Between Integrals and Optima: New Methods for Scalable Machine Learning

Chris J. Maddison

Exeter College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

To Maia.

Acknowledgements

The work in this thesis reflects the support of an entire community.

Foremost in this group are my supervisors, Arnaud Doucet and Yee Whye Teh, and my mentors, Danny Tarlow and Ilya Sutskever. They provided me with opportunities, shared with me their ideas, and inspired me by their relentless vision and drive. I would also like to thank Geoffrey Hinton, David Silver, and Tom Minka for their leadership and their support.

Many parts of this thesis reflect collaborative work. I would like to thank Andriy Mnih, for defending the baselines; to thank Dieterich Lawson and George Tucker, for their excitement and many long hours; and to thank Daniel Paulin, for our year-long adventure into the depths of optimization. This work is made stronger by the participation of many voices.

Many people provided thoughts, comments, and critique throughout the research process. I would like to thank Jimmy Ba, David Balduzzi, Gabriel Barth-Maron, Lars Buesing, Stefano Favarro, Patrick Rebeschini, Daniel J. Rezende, Sushant Sachdeva, and Theophane Weber.

My work was supported by a DeepMind Graduate Scholarship, a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada, and the Open Phil AI Fellowship. In particular, I would like to thank DeepMind for inviting me to participate in their vibrant research community, and Daniel Dewey at Open Philanthropy for his work in building a space for fellowship.

Finally, for friendship in research and in life, I would like to thank Ryan Adams, Noam Brown, George Dahl, Emily Denton, Jessica Forde, Marta Garnelo, Erin Grant, Roger Grosse, Frauke Harms, Tamir Hazan, James Martens, Emile Mathieu, Robert Nishihara, Aditi Raghunathan, Jasper Snoek, Ioan Stefanovici, Kevin Swersky, and Wojciech Zaremba.

Abstract

The success of machine learning is due in part to the effectiveness of scalable computational methods, like stochastic gradient descent or Monte Carlo methods, that undergird learning algorithms. This thesis contributes four new scalable methods for distinct problems that arise in machine learning. It introduces a new method for gradient estimation in discrete variable models, a new objective for maximum likelihood learning in the presence of latent variables, and two new gradient-based differentiable optimization methods. Although quite different, these contributions address distinct, critical parts of a typical machine learning workflow. Furthermore, each contribution is inspired by an interplay between the numerical problems of optimization and integration, an interplay that forms the central theme of this thesis.

Contents

1 Introduction **1**

1.1 Integrals and Optima 1

1.2 Gradient Estimation and the Gumbel max trick 2

1.3 Maximum Likliehood via Variational Objectives 5

1.4 Gradient-based Optimization and Momentum 6

1.5 Overview 9

2 The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables **10**

3 Filtering Variational Objectives **24**

4 Hamiltonian Descent **38**

4.1 Abstract 38

4.2 Introduction 38

4.2.1 Notation and Convex Analysis Review 42

4.2.2 Related Literature 43

4.3 Continuous Dynamics 44

4.3.1 Hamiltonian Systems 45

4.3.2 Continuously Descending the Hamiltonian 45

4.3.3 Continuous Hamiltonian Descent on Convex Functions 47

4.3.4 Partial Lower Bounds 53

4.4 Optimization Algorithms 56

4.4.1 Implicit Method 56

4.4.2 First Explicit Method, with Analysis via the Hessian of f . . . 58

4.4.3 Second Explicit Method, with Analysis via the Hessian of k . 59

4.4.4 First Explicit Method on Non-Convex f 62

4.5 Kinetic Maps for Functions with Power Behavior 63

4.5.1 Power Kinetic Energies 63

4.5.2 Matching power kinetic ∇k with assumptions on f 67
4.5.3 Matching relativistic kinetic ∇k with assumptions on f 72
4.6 Conclusion 78

5 Dual Space Preconditioning for Gradient Descent 81
5.1 Abstract 81
5.2 Introduction 81
5.3 Related literature 84
5.4 Convex analysis background 85
5.4.1 Convex conjugate and Legendre functions 85
5.4.2 Relative smoothness and relative strong convexity 87
5.5 Analysis of the dual preconditioned scheme 90
5.5.1 Motivation 90
5.5.2 Relative conditions in the dual space 91
5.5.3 Convergence rates under dual relative smoothness 95
5.6 Applications 99
5.6.1 Exponential Penalty Functions 99
5.6.2 p-norm Regression 103
5.7 Discussion 108
5.7.1 Special cases and related methods 108
5.7.2 Conclusions 110

6 Conclusions 113

A Appendix to Hamiltonian Descent 115
Ap.1 Proofs for convergence of continuous systems 115
Ap.2 Proofs for partial lower bounds 117
Ap.3 Proofs of convergence for discrete systems 125
Ap.3.1 Implicit Method 125
Ap.3.2 First Explicit Method 131
Ap.3.3 Second Explicit Method 135
Ap.3.4 Explicit Method on Non-Convex f 140
Ap.4 Proofs for power kinetic energies 141

Bibliography 157

Chapter 1

Introduction

Machine learning enjoyed a series of high-profile achievements in, e.g., image classification [140], speech recognition [118], automatic machine translation [269], speech synthesis [253], and games playing [235]. These successes are enabled partly by software packages that automate learning [245, 1, 200, 43]; partly by the use of expressive black-box models and predictive accuracy as the metric for validating them [45, 147]; and partly by the unreasonable effectiveness of computational methods, like stochastic gradient descent [37] or Monte Carlo methods [220], that undergird learning algorithms.

This thesis contributes to the computational methodology of machine learning through new methods for gradient estimation in discrete variable models (Chapter 2), new objectives for maximum likelihood learning in the presence of latent variables (Chapter 3), and gradient-based differentiable optimization (Chapters 4 & 5). Although quite distinct as methods, the problems of computing derivatives, defining objective functions, and optimizing are all critical parts of a typical machine learning workflow. Furthermore, each new method is thematically unified by an interplay between the numerical problems of optimization and integration.

1.1 Integrals and Optima

Optimization and integration hide at the core of many machine learning methods. Consider for example the maximum likelihood objective [79], which is one most widely-used for fitting machine learning models [178]. In all but the simplest models, finding the maximum likelihood parameters requires an algorithm that searches over the likelihood surface for an extreme point [86]. This is the problem of optimization, and is usually accomplished using some kind of method for numerical optimization [38]. Maximum likelihood in many graphical models requires computing marginal

distributions, which are essentially counts or volumes of high-dimensional spaces. This is the problem of integration, and it is a well-studied topic for graphical models [260].

At first, these two problems do not seem related. Integrals measure the volume under a function's surface taken over its domain. Integration is fundamentally an accumulative process; small changes anywhere in a function's surface will generally lead to different values for its integral. Optima are the extreme points of a function's surface. Many changes to the surface would not change the location of its optima, and for differentiable functions, these optima are completely characterized by local properties. Thus, optima appear to be mostly *local* phenomena, while integrals appear to be mostly *global* ones.

Yet, this apparent distinction is only superficial. On the one hand, in some settings integration can be reduced to optimization through a variational approach [128]. On the other hand, some optimization algorithms may be interpreted as numerical integrators of ordinary differential equations [206, 238]. Consider a third specific example in the problem of sampling from probability distributions, which is most often used as a subroutine in the estimation of integrals via Monte Carlo estimation. Markov chain Monte Carlo methods [46] are frequently used to solve this problem, and they can be shown to reduce a real-valued objective function over probability distributions [129, 221], i.e., to be performing optimization in the space of measures. This view of sampling as optimization can be used to analyze rates [169] and design methods [26, 262]. Each of the three contributions of this thesis are seeded by an observation of this type. Chapter 2 is based on the Gumbel max trick, which reduces random variate simulation to optimization. Chapter 3 is based on variational inference, which reduces posterior inference to optimization. Chapters 4 and 5 are inspired by continuous-time perspectives on optimization, which cast optimization methods as numerical integrators of differential equations. We introduce each of these contexts in turn.

1.2 Gradient Estimation and the Gumbel max trick

Software packages for automatic differentiation [245, 1, 200, 43] have been a driving force in the deep learning revolution [147], because they liberate practitioners to explore quickly in the space of models and objectives, without requiring separate derivative derivations. Therefore, expanding the class of objectives or models that are amenable to automatic differentiation can have an outsized impact on practice.

An important class of objectives in machine learning are integrals whose density is the parameter we wish to optimize over. A typical example are the objectives of reinforcement learning, in which the density in the integrand represents a distribution over actions and the aim is to maximize the expected reward over the action distribution [242]. Another example, which we will cover in the next section, are variational objectives over approximate posteriors in models with latent variables [137, 218]. For objectives over densities in the integrand, gradient estimators [177, 10] may be used in lieu of exact gradients under the Robbins-Monro conditions [219]. In this section we introduce two popular generic strategies for designing gradient estimators in machine learning. We return to the interplay between optimization and integration at the end of the section with the Gumbel max trick.

REINFORCE [264] and reparameterization tricks [137, 218] are two of the most widely-used gradient estimation techniques in machine learning. Let $(\pi_\phi)_{\phi\in\mathbb{R}^n}$ be a family of probability distribution on $\mathbb{R}^d$ (equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$) parameterized by $\phi \in \mathbb{R}^n$. Assume that each π_ϕ admits a density p_ϕ with respect to the Lebesgue measure dx. Suppose we are given a differentiable objective function $f : \mathbb{R}^d \to \mathbb{R}$, then our goal is to approximate the following vector of first partial derivatives for $X \sim \pi_\phi$:

$$\nabla_\phi \, \mathbb{E}[f(X)] = \nabla_\phi \left[\int f(x) p_\phi(x) \, dx \right] = \left(\frac{\partial}{\partial \phi_i} \left[\int f(x) p_\phi(x) \, dx \right] \right)_{i=1}^{n}. \tag{1.1}$$

The REINFORCE estimator [264], also known as the score function estimator [98], is based on the following observation that (assuming that we may exchange derivatives and integrals):

$$\nabla_\phi \int f(x) p_\phi(x) \, dx = \int f(x) \nabla_\phi p_\phi(x) \, dx = \int f(x) \nabla_\phi \log p_\phi(x) p_\phi(x) \, dx, \tag{1.2}$$

for p_ϕ differentiable and positive a.e. This suggests the REINFORCE estimator, $f(X)\nabla_\phi \log p_\phi(X)$ for $X \sim \pi_\phi$, which may be used a Monte Carlo estimator of the true gradient. A sufficient condition guaranteeing that one can exchange the derivative and integral in REINFORCE is that p_ϕ be continuously differentiable in ϕ (a.e. in x), $f \in L^2$, and $|\partial p_\phi(x)/\partial \phi_i| \leq M_i(x)$ (a.e. in x) for some $M_i \in L^2$ and all $\phi \in \mathbb{R}^n$ [10, Prop. 3.5 of Chap. 7].

Reparameterization tricks [137, 218, 225], also known as infinitesimal perturbation analysis [96, 52], are based on a change of variables. We consider a simple example to introduce the idea. Let $X \sim \mathcal{N}(\phi, I) = \pi_\phi$ be a Gaussian random variable with

mean $\phi \in \mathbb{R}^d$. The observation driving the reparameterization tricks is the following (assuming again that we may exchange derivatives and integrals):

$$\nabla_\phi \mathbb{E}[f(X)] = \nabla_\phi \mathbb{E}[f(\phi + \epsilon)] = \mathbb{E}[\nabla_\phi [f(\phi + \epsilon)]], \tag{1.3}$$

where $\epsilon \sim \mathcal{N}(0, I)$. This suggests the reparameterization gradient estimator, $\nabla_\phi [f(\phi + \epsilon)]$. More generally, these estimators can be derived when there exists a differentiable change of variables $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^d$ such that $X \overset{d}{=} g(\phi, \epsilon)$ for some random variable ϵ with range $\mathbb{R}^m$. The specific requirements on g are that $f(g(\phi, \epsilon))$ is differentiable in ϕ (a.e. in ϵ) and that $f(g(\phi, \epsilon))$ satisfies a local Lipschitz condition for some Lipschitz constant in L^2 [10, Prop. 2.3, Chap. 7]. The rediscovery of reparameterization gradient estimators in machine learning [137, 218] revolutionized probabilistic methods for deep learning, because these estimators are simple to implement and generally lower variance than the REINFORCE estimator [88].

Reparameterization tricks are not always readily available. Consider when the base measure of (1.1) is a counting measure and not Lebesgue. The REINFORCE estimators are typically applicable, but these tend to have prohibitively high variance. Control variates can be used to reduce the variance of REINFORCE [e.g., 172], but one may prefer a reparameterization estimator. Discrete distributions admit a generic change of variables (reparameterization), known in machine learning as the Gumbel max trick [164], with a long history in random choice theory as the Plackett-Luce model [205, 273, 158]. Let $X \sim \pi_\phi$ be a discrete random variable whose range $\mathcal{X} = \{x \in \{0, 1\}^n : \sum_i x_i = 1\}$ are the vertices of the $(n-1)$-dimensional simplex. If p_ϕ is its mass function and $G_i \sim$ Gumbel [110] i.i.d., then

$$X \overset{d}{=} \arg\max_{x \in \mathcal{X}} \{x^T (G + \log p_\phi(x))\}. \tag{1.4}$$

Unfortunately, this reparameterization of X does not obviously get us closer to our goal of deriving a reparameterization gradient estimator for $\mathbb{E}[f(X)]$. The partial derivatives in ϕ_i of the right hand side of (1.4) are 0 almost surely, and the exchange of derivative and integral cannot be justified.

Nevertheless, (1.4) is the starting point of the first contribution of this thesis: improved gradient estimators for discrete variable models (Chapter 2). The Gumbel max trick is also a return to our central theme. (1.4) shows that simulating a discrete random variable (an integration-type procedure) can be be accomplished by optimizing a random function (an optimization-type procedure). Indeed, [164, 160] show that this is not unique to counting measures. One can define a certain random function whose optimum is distributed according to any given probability distribution, and generalize the Gumbel max trick to continuous distributions.

1.3 Maximum Likliehood via Variational Objectives

In section 1.2, we considered the problem of computing derivatives. A major application of the gradient estimators of the previous section occurs in the development of large, non-linear, black-box statistical models for high dimensional natural data. For example, [253] designed an autoregressive model of audio waveform data, called WaveNet. When fit to human speech, WaveNet is capable of simulating a novel, realistic-sounding utterance in the voice of a given speaker, even famous celebrities. Collectively called "generative modeling", these machine learning techniques seek to fit some parametric family of distribution to the empirical distribution of a given dataset [137, 218, 67, 100, 196, 248, 176]. Many of these models incorporate unobserved latent variables, which poses a considerable challenge for methods like maximum likelihood estimation. A set of scalable approaches [137, 218, 213], collectively called variational autoencoders, based on variational inference (VI) [128] and amortized inference have recently made these models trainable. Although VI has a long history [32], our introduction will emphasize VI as seen from recent developments in statistical deep learning. Furthermore, we will see that VI is another example of integration reducing to optimization.

First, we define a typical maximum likelihood objective for latent variable models. Denote the observations of a dataset by Y, a realization of an $\mathbb{R}^{d_1}$-valued random variable. Let us assume that Y is jointly distributed with a random variable $X \in \mathbb{R}^{d_2}$ with joint density $p_\theta(x, y)$ (with respect to the Lebesgue measure) for some parameter $\theta \in \mathbb{R}^n$. It is typical that p_θ is specified via a parametric "prior" density $p_\theta(x)$ and a conditional "likelihood" density $p_\theta(y|x)$ parameterized by some parametric function of x. The goal of maximum likelihood estimation (MLE) [79] is to recover $\theta \in \mathbb{R}^n$ that maximizes the marginal log-likelihood,

$$\log p_\theta(Y) = \log \left(\int p_\theta(x, Y) \, dx \right). \tag{1.5}$$

In such generality, this model is not particularly new, e.g. mixture models are captured in this family. It is the expressivity of the likelihood function (generally computed by some neural network) that make such models suitable for complex natural data and the likelihood is a point of major innovation, e.g., [55]. The assumption that the observed data is actually distributed according to this distribution for some $\theta \in \mathbb{R}^n$ is essentially never verified, but approaches that approximately optimize the likelihood (1.5) of these large, non-linear, black-box models tend to work well in practice, e.g. [142].

Ideally, one would compute derivatives of the objective (1.5) and run an optimization routine to approximate the MLE. Unfortunately, even computing the derivative is intractable, owing to the integral inside the logarithm. VI is a strategy for overcoming this intractability [128, 17]. Let $q_\phi(x|y)$ be a conditional density in some parametric family, defined for all y in the support of $p_\theta(x, \cdot)$. The variational lower bound is defined as the following objective,

$$\mathcal{L}(Y, \theta, \phi) = \int \left[\log \frac{p_\theta(x, Y)}{q_\phi(x|Y)}\right] q_\theta(x|Y)\, dx \leq \log p_\theta(Y) \tag{1.6}$$

The bound is tight when $q_\phi(x|y)$ is the true conditional $p_\theta(x|y)$ under the model p_θ. Assuming the parametric family of q contains the true posteriors of p, the joint optimum of (1.6) in θ and ϕ is the MLE. The EM algorithm [66] is a classical algorithm used to optimize (1.6), but the updates of the EM algorithm are generally too expensive for the models considered in statistical deep learning.

There are two insights behind the variational autoencoder approach to using (1.6) to perform approximate MLE in deep, non-linear models. The first is that (1.6) is in the same form as (1.1) and its gradients in ϕ, θ may be estimated with the approaches of section 1.2. The second, is that a single, highly-expressive q_ϕ can approximate the true conditional $p_\theta(x|y)$ over an entire dataset. The second insight is typically called amortized inference, because the cost of inference is "amortized" in the learning of the parameters ϕ. In practice, in deep learning, this objective is then optimized with some gradient-based optimization routine jointly over θ and ϕ [137, 218].

Returning now to the theme of the thesis, note that, when the variational bound (1.6) is tight, the problem of integrating out the latent variables (computing the marginal log-likelihood) can be cast as an optimization problem (maximizing over q_ϕ). In Chapter 3, we design generalizations of (1.6) specialized for sequential models.

1.4 Gradient-based Optimization and Momentum

In section 1.3 we introduced variational objectives for maximum likelihood estimation, and in section 1.2 we introduced gradient estimators for such objectives. In this section, we complete the picture by introducing the basics of gradient-based optimization. We will consider a simplified scenario. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function that is bounded below. We are interested in finding a minimizing argument,

$$x_{\min} = \arg\min_{x \in \mathbb{R}^d} f(x). \tag{1.7}$$

The gradient descent algorithm is one of the simplest algorithms for this problem. It begins with the observation that the negative gradient (the vector of first partial derivatives, $\nabla f(x) = (\partial f(x)/\partial x^{(n)})$) points in the direction of greatest instantaneous decrease on the function's surface. Given some initial point $x_0 \in \mathbb{R}^d$ and a step size $1/L > 0$, the iterates of gradient descent are given by,

$$x_{i+1} = x_i - \nabla f(x_i)/L. \tag{1.8}$$

If the step size is small enough and the function smooth enough, the iterates of this scheme will iteratively descend the function's surface. Methods like gradient descent, which only use the first-order derivatives [190, 207, 193], are among the most popular in machine learning. This is partly due to their ease of implementation and scalability [130]. One of the first questions to ask about any optimization scheme is, when can one guarantee its convergence and at what rate? This is the topic of the final two chapters of this thesis (Chapters 4 & 5).

Smoothness and strong convexity are two typical conditions used to guarantee the convergence of gradient descent. For smooth convex f, the iterates $f(x_i) - f(x_{\min})$ of gradient descent converge at a rate of $\mathcal{O}(1/i)$. For smooth, strongly convex f, the iterates $f(x_i) - f(x_{\min})$ of gradient descent converge geometrically at a rate of $\mathcal{O}(\lambda^i)$ for some $0 < \lambda < 1$. Smoothness is a condition that guarantees that f grows no faster than a quadratic, and strong convexity is a condition that guarantees that f grows no slower than a quadratic. Intuitively, smoothness guarantees a descent in f and strong convexity guarantees a sufficient descent for fast convergence. See [190, 207, 193] for these classical results.

Gradient-based optimizers are generally improved by incorporating momentum into the dynamics of the iterates. Although there is flexibility in how it is implemented, momentum adds a persistence of "motion" that improves convergence in directions of persistent gradient signal. The first introduction of momentum is due to Boris Polyak [206], and his iteration is given by,

$$x_{i+1} = x_i - \alpha \nabla f(x_i) + \beta(x_i - x_{i-1}), \tag{1.9}$$

where $\alpha, \beta > 0$. Note that the only distinction between (1.9) and the gradient method (1.8) is the last term, which is exactly the term that produces a persistence of motion. Polyak showed that (1.9) is capable of achieving optimal local convergence rates (an improvement over the gradient method) on twice-differentiable, smooth, and strongly convex f. Nesterov's celebrated accelerated gradient algorithms are another form of

momentum, and these achieve optimal global rates for smooth convex f and smooth, strongly convex f [195].

In the final two chapters of this thesis we ask whether the momentum method (Chapters 4) and the gradient method (Chapters 5) can be generalized to converge under conditions that generalize smoothness or strong convexity. As we show, this generalization is possible, and the inspiration comes from the Monte Carlo literature (another return to the central theme of the interplay between optiization and integration).

First, consider what happens to the sequence of iterates defined by (1.9) as $\alpha \to 0$ and $\beta = \sqrt{\alpha}(1-\gamma)$ for $0 < \gamma < 1$. The iterates (under reasonable smoothness conditions) approach solutions of the following ordinary differential equation [233].

$$\begin{aligned} x_t' &= p_t \\ p_t' &= -\nabla f(x_t) - \gamma p_t \end{aligned} \tag{1.10}$$

Very similar differential equations describe dynamics used in the Monte Carlo literature. In particular, the Hamiltonian Monte Carlo algorithm (HMC) is one of the most successful Monte Carlo methods for continuous distributions and it uses discretizations of the Hamiltonian differential equations (below) to propose moves in a Metropolis-Hastings scheme [185],

$$\begin{aligned} x_t' &= \nabla k(p_t) \\ p_t' &= -\nabla f(x_t) \end{aligned} \tag{1.11}$$

where $k : \mathbb{R}^d \to \mathbb{R}$ is generally a convex differentiable function and f is the log-density of the measure whose integrals we wish to approximate. The dynamics described by (1.11) are used in physics as a model of frictionless dynamics that preserve energy. (1.10) describes dynamics exposed to a constant source of friction, which dissipates energy [165]. This dissipation is important for optimization, because otherwise the system would fail to converge. Indeed, (1.10) and (1.11) describe the same system when $\gamma = 0$ and $\nabla k(p) = p$. Therefore, momentum optimizers may be seen as the dissipative cousin of the dynamics used in HMC.

Drawing parallels between optimization and Monte Carlo methods can inspire methodological progress. This is because some questions, more naturally posed in one literature (Monte Carlo or optimization), may also be relevant for the other. For example, the function k in (1.11) is a free parameter that a user may design, and recent work in the HMC literature [149] argued that a sensible choice of k achieves $\nabla k \approx (\nabla f)^{-1}$. This suggests considering following differential equation for optimization,

$$\begin{aligned} x_t' &= \nabla k(p_t) \\ p_t' &= -\nabla f(x_t) - \gamma p_t \end{aligned} \tag{1.12}$$

where $k : \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function designed by the user. The natural question to ask of (1.12) is, how should k and f relate for the convergence of discretizations of (1.12) to the minimum of f? This is the starting point of Chapter 4, and we find a rather satisfying answer; ∇k should approximate $(\nabla f)^{-1} - x_{\min}$ in a sense made precise by conditions that are similar to smoothness and strong convexity. In Chapter 5 we simplify these conditions to ones that exactly generalize smoothness and strong convexity, and we present a version of non-linear preconditioning of the gradient descent algorithm that exploits them. The conditions that we are arrive at are closely related to, but distinct from, conditions recently introduced to study mirror descent [16, 156].

1.5 Overview

This thesis presents four methodological contributions to distinct, complementary parts of a standard machine learning workflow. It is an integrated thesis formed of four chapters, each corresponding to a separate paper and presented in the form that they were published or submitted. Two papers were published at machine learning conferences, one is in review, and one is in preprint. * indicates joint first authorship.

1. Chris J. Maddison*, Dieterich Lawson*, George Tucker*, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, Yee Whye Teh. Filtering Variational Objectives. In *Advances in Neural Information Processing Systems*, 2017.

2. Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, 2017.

3. Chris J. Maddison*, Daniel Paulin*, Yee Whye Teh, Arnaud Doucet. Hamiltonian Descent Methods. In *Preprint*, 2018.

4. Chris J. Maddison*, Daniel Paulin*, Yee Whye Teh, Arnaud Doucet. Dual Space Preconditioning for Gradient Descent. In *Review*, 2019.

THE CONCRETE DISTRIBUTION: A CONTINUOUS RELAXATION OF DISCRETE RANDOM VARIABLES

Chris J. Maddison[1,2], Andriy Mnih[1], & Yee Whye Teh[1]
[1]DeepMind, London, United Kingdom
[2]University of Oxford, Oxford, United Kingdom
cmaddis@stats.ox.ac.uk

ABSTRACT

The reparameterization trick enables optimizing large scale stochastic computation graphs via gradient descent. The essence of the trick is to refactor each stochastic node into a differentiable function of its parameters and a random variable with fixed distribution. After refactoring, the gradients of the loss propagated by the chain rule through the graph are low variance unbiased estimators of the gradients of the expected loss. While many continuous random variables have such reparameterizations, discrete random variables lack useful reparameterizations due to the discontinuous nature of discrete states. In this work we introduce CONCRETE random variables—CONtinuous relaxations of disCRETE random variables. The Concrete distribution is a new family of distributions with closed form densities and a simple reparameterization. Whenever a discrete stochastic node of a computation graph can be refactored into a one-hot bit representation that is treated continuously, Concrete stochastic nodes can be used with automatic differentiation to produce low-variance biased gradients of objectives (including objectives that depend on the log-probability of latent stochastic nodes) on the corresponding discrete graph. We demonstrate the effectiveness of Concrete relaxations on density estimation and structured prediction tasks using neural networks.

1 INTRODUCTION

Software libraries for automatic differentiation (AD) [1, 245] are enjoying broad use, spurred on by the success of neural networks on some of the most challenging problems of machine learning. The dominant mode of development in these libraries is to define a forward parametric computation, in the form of a directed acyclic graph, that computes the desired objective. If the components of the graph are differentiable, then a backwards computation for the gradient of the objective can be derived automatically with the chain rule. The ease of use and unreasonable effectiveness of gradient descent has led to an explosion in the diversity of architectures and objective functions. Thus, expanding the range of useful continuous operations can have an outsized impact on the development of new models. For example, a topic of recent attention has been the optimization of stochastic computation graphs from samples of their states. Here, the observation that AD "just works" when stochastic nodes[1] can be reparameterized into deterministic functions of their parameters and a fixed noise distribution [137, 218], has liberated researchers in the development of large complex stochastic architectures [106].

Computing with discrete stochastic nodes still poses a significant challenge for AD libraries. Deterministic discreteness can be relaxed and approximated reasonably well with sigmoidal functions or the softmax [104, 102], but, if a distribution over discrete states is needed, there is no clear solution. There are well known unbiased estimators for the gradients of the parameters of a discrete stochastic node from samples. While these can be made to work with AD, they involve special casing

[1]For our purposes a stochastic node of a computation graph is just a random variable whose distribution depends in some deterministic way on the values of the parent nodes.

and defining surrogate objectives [231], and even then they can have high variance. Still, reasoning about discrete computation comes naturally to humans, and so, despite the difficulty associated, many modern architectures incorporate discrete stochasticity [271, 138].

This work is inspired by the observation that many architectures treat discrete nodes continuously, and gradients rich with counterfactual information are available for each of their possible states. We introduce a CONtinuous relaxation of disCRETE random variables, CONCRETE for short, which allow gradients to flow through their states. The *Concrete distribution* is a new parametric family of continuous distributions on the simplex with closed form densities. Sampling from the Concrete distribution is as simple as taking the softmax of logits perturbed by fixed additive noise. This reparameterization means that Concrete stochastic nodes are quick to implement in a way that "just works" with AD. Crucially, every discrete random variable corresponds to the zero temperature limit of a Concrete one. In this view optimizing an objective over an architecture with discrete stochastic nodes can be accomplished by gradient descent on the samples of the corresponding Concrete relaxation. When the objective depends, as in variational inference, on the log-probability of discrete nodes, the Concrete density is used during training in place of the discrete mass. At test time, the graph with discrete nodes is evaluated.

The paper is organized as follows. We provide a background on stochastic computation graphs and their optimization in Section 2. Section 3 reviews a reparameterization for discrete random variables, introduces the Concrete distribution, and discusses its application as a relaxation. Section 4 reviews related work. In Section 5 we present results on a density estimation task and a structured prediction task on the MNIST and Omniglot datasets. When comparing the effectiveness of gradients obtained via Concrete relaxations to a state-of-the-art-method (VIMCO) [175], we find that they are competitive—occasionally outperforming and occasionally underperforming—all the while being implemented in an AD library without special casing.

2 BACKGROUND

2.1 OPTIMIZING STOCHASTIC COMPUTATION GRAPHS

Stochastic computation graphs (SCGs) provide a formalism for specifying input-output mappings, potentially stochastic, with learnable parameters using directed acyclic graphs (see [231] for a review). The state of each non-input node in such a graph is obtained from the states of its parent nodes by either evaluating a deterministic function or sampling from a conditional distribution. Many training objectives in supervised, unsupervised, and reinforcement learning can be expressed in terms of SCGs.

To optimize an objective represented as a SCG, we need estimates of its parameter gradients. We will concentrate on graphs with some stochastic nodes (backpropagation covers the rest). For simplicity, we restrict our attention to graphs with a single stochastic node X. We can interpret the forward pass in the graph as first sampling X from the conditional distribution $p_\phi(x)$ of the stochastic node given its parents, then evaluating a deterministic function $f_\theta(x)$ at X. We can think of $f_\theta(X)$ as a noisy objective, and we are interested in optimizing its expected value $L(\theta, \phi) = \mathbb{E}_{X\sim p_\phi(x)}[f_\theta(X)]$ w.r.t. parameters θ, ϕ.

In general, both the objective and its gradients are intractable. We will side-step this issue by estimating them with samples from $p_\phi(x)$. The gradient w.r.t. to the parameters θ has the form

$$\nabla_\theta L(\theta, \phi) = \nabla_\theta \mathbb{E}_{X\sim p_\phi(x)}[f_\theta(X)] = \mathbb{E}_{X\sim p_\phi(x)}[\nabla_\theta f_\theta(X)] \tag{1}$$

and can be easily estimated using Monte Carlo sampling:

$$\nabla_\theta L(\theta, \phi) \simeq \frac{1}{S}\sum_{s=1}^{S} \nabla_\theta f_\theta(X^s), \tag{2}$$

where $X^s \sim p_\phi(x)$ i.i.d. The more challenging task is to compute the gradient w.r.t. the parameters ϕ of $p_\phi(x)$. The expression obtained by differentiating the expected objective,

$$\nabla_\phi L(\theta, \phi) = \nabla_\phi \int p_\phi(x) f_\theta(x)\,\mathrm{d}x = \int f_\theta(x) \nabla_\phi p_\phi(x)\,\mathrm{d}x, \tag{3}$$

does not have the form of an expectation w.r.t. x and thus does not directly lead to a Monte Carlo gradient estimator. However, there are two ways of getting around this difficulty which lead to the two classes of estimators we will now discuss.

2.2 SCORE FUNCTION ESTIMATORS

The *score function estimator* (SFE) [87], also known as the REINFORCE [264] or likelihood-ratio estimator [98], is based on the identity $\nabla_\phi p_\phi(x) = p_\phi(x)\nabla_\phi \log p_\phi(x)$, which allows the gradient in Eq. 3 to be written as an expectation:

$$\nabla_\phi L(\theta, \phi) = \mathbb{E}_{X\sim p_\phi(x)}\left[f_\theta(X)\nabla_\phi \log p_\phi(X)\right]. \tag{4}$$

Estimating this expectation using naive Monte Carlo gives the estimator

$$\nabla_\phi L(\theta, \phi) \simeq \frac{1}{S}\sum\nolimits_{s=1}^{S} f_\theta(X^s)\nabla_\phi \log p_\phi(X^s), \tag{5}$$

where $X^s \sim p_\phi(x)$ i.i.d. This is a very general estimator that is applicable whenever $\log p_\phi(x)$ is differentiable w.r.t. ϕ. As it does not require $f_\theta(x)$ to be differentiable or even continuous as a function of x, the SFE can be used with both discrete and continuous random variables.

Though the basic version of the estimator can suffer from high variance, various variance reduction techniques can be used to make the estimator much more effective [103]. Baselines are the most important and widely used of these techniques [264]. A number of score function estimators have been developed in machine learning [198, 213, 172, 247, 109], which differ primarily in the variance reduction techniques used.

2.3 REPARAMETERIZATION TRICK

In many cases we can sample from $p_\phi(x)$ by first sampling Z from some fixed distribution $q(z)$ and then transforming the sample using some function $g_\phi(z)$. For example, a sample from $\mathrm{Normal}(\mu, \sigma^2)$ can be obtained by sampling Z from the standard form of the distribution $\mathrm{Normal}(0, 1)$ and then transforming it using $g_{\mu,\sigma}(Z) = \mu + \sigma Z$. This two-stage reformulation of the sampling process, called the *reparameterization trick*, allows us to transfer the dependence on ϕ from p into f by writing $f_\theta(x) = f_\theta(g_\phi(z))$ for $x = g_\phi(z)$, making it possible to reduce the problem of estimating the gradient w.r.t. parameters of a distribution to the simpler problem of estimating the gradient w.r.t. parameters of a deterministic function.

Having reparameterized $p_\phi(x)$, we can now express the objective as an expectation w.r.t. $q(z)$:

$$L(\theta, \phi) = \mathbb{E}_{X\sim p_\phi(x)}[f_\theta(X)] = \mathbb{E}_{Z\sim q(z)}[f_\theta(g_\phi(Z))]. \tag{6}$$

As $q(z)$ does not depend on ϕ, we can estimate the gradient w.r.t. ϕ in exactly the same way we estimated the gradient w.r.t. θ in Eq. 1. Assuming differentiability of $f_\theta(x)$ w.r.t. x and of $g_\phi(z)$ w.r.t. ϕ and using the chain rule gives

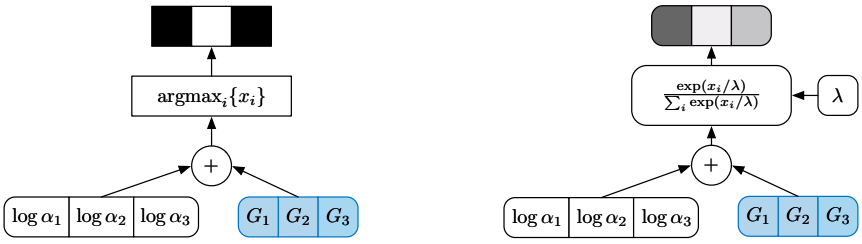
$$\nabla_\phi L(\theta, \phi) = \mathbb{E}_{Z\sim q(z)}[\nabla_\phi f_\theta(g_\phi(Z))] = \mathbb{E}_{Z\sim q(z)}\left[f'_\theta(g_\phi(Z))\nabla_\phi g_\phi(Z)\right]. \tag{7}$$

The reparameterization trick, introduced in the context of variational inference independently by [137], [218], and [246], is usually the estimator of choice when it is applicable. For continuous latent variables which are not directly reparameterizable, new hybrid estimators have also been developed, by combining partial reparameterizations with score function estimators [225, 179].

2.4 APPLICATION: VARIATIONAL TRAINING OF LATENT VARIABLE MODELS

We will now see how the task of training latent variable models can be formulated in the SCG framework. Such models assume that each observation x is obtained by first sampling a vector of latent variables Z from the prior $p_\theta(z)$ before sampling the observation itself from $p_\theta(x \mid z)$. Thus the probability of observation x is $p_\theta(x) = \sum_z p_\theta(z)p_\theta(x \mid z)$. Maximum likelihood training of such models is infeasible, because the log-likelihood (LL) objective $L(\theta) = \log p_\theta(x) = \log \mathbb{E}_{Z\sim p_\theta(z)}[p_\theta(x \mid Z)]$ is typically intractable and does not fit into the above framework due to the expectation being inside the $\log$. The multi-sample variational objective [50],

$$\mathcal{L}_m(\theta, \phi) = \mathop{\mathbb{E}}_{Z^i\sim q_\phi(z|x)}\left[\log\left(\frac{1}{m}\sum_{i=1}^{m}\frac{p_\theta(Z^i,\, x)}{q_\phi(Z^i \mid x)}\right)\right]. \tag{8}$$



(a) Discrete(α) (b) Concrete(α, λ)

Figure 1: Visualization of sampling graphs for 3-ary discrete $D \sim \text{Discrete}(\alpha)$ and 3-ary Concrete $X \sim \text{Concrete}(\alpha, \lambda)$. White operations are deterministic, blue are stochastic, rounded are continuous, square discrete. The top node is an example state; brightness indicates a value in [0,1].

provides a convenient alternative which has precisely the form we considered in Section 2.1. This approach relies on introducing an auxiliary distribution $q_\phi(z \mid x)$ with its own parameters, which serves as approximation to the intractable posterior $p_\theta(z \mid x)$. The model is trained by jointly maximizing the objective w.r.t. to the parameters of p and q. The number of samples used inside the objective m allows trading off the computational cost against the tightness of the bound. For $m = 1$, $\mathcal{L}_m(\theta, \phi)$ becomes is the widely used evidence lower bound (ELBO) [120] on $\log p_\theta(x)$, while for $m > 1$, it is known as the importance weighted bound [50].

3 The Concrete Distribution

3.1 Discrete Random Variables and the Gumbel-Max Trick

To motivate the construction of Concrete random variables, we review a method for sampling from discrete distributions called the Gumbel-Max trick [273, 116, 164]. We restrict ourselves to a representation of discrete states as vectors $d \in \{0, 1\}^n$ of bits that are one-hot, or $\sum_{k=1}^n d_k = 1$. This is a flexible representation in a computation graph; to achieve an integral representation take the inner product of d with $(1, \ldots, n)$, and to achieve a point mass representation in $\mathbb{R}^m$ take Wd where $W \in \mathbb{R}^{m \times n}$.

Consider an unnormalized parameterization $(\alpha_1, \ldots, \alpha_n)$ where $\alpha_k \in (0, \infty)$ of a discrete distribution $D \sim \text{Discrete}(\alpha)$—we can assume that states with 0 probability are excluded. The Gumbel-Max trick proceeds as follows: sample $U_k \sim \text{Uniform}(0, 1)$ i.i.d. for each k, find k that maximizes $\{\log \alpha_k - \log(-\log U_k)\}$, set $D_k = 1$ and the remaining $D_i = 0$ for $i \neq k$. Then

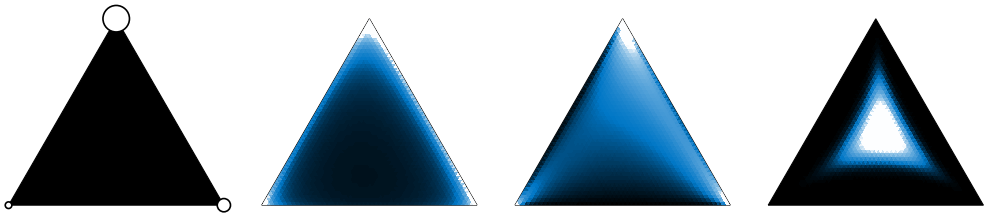
$$\mathbb{P}(D_k = 1) = \frac{\alpha_k}{\sum_{i=1}^n \alpha_i}. \tag{9}$$

In other words, the sampling of a discrete random variable can be refactored into a deterministic function—componentwise addition followed by argmax—of the parameters $\log \alpha_k$ and fixed distribution $-\log(-\log U_k)$. See Figure 1a for a visualization.

The apparently arbitrary choice of noise gives the trick its name, as $-\log(-\log U)$ has a Gumbel distribution. This distribution features in extreme value theory [110] where it plays a central role similar to the Normal distribution: the Gumbel distribution is stable under $\max$ operations, and for some distributions, the order statistics (suitably normalized) of i.i.d. draws approach the Gumbel in distribution. The Gumbel can also be recognized as a $-\log$-transformed exponential random variable. So, the correctness of (9) also reduces to a well known result regarding the argmin of exponential random variables.

3.2 Concrete Random Variables

The derivative of the argmax is 0 everywhere except at the boundary of state changes, where it is undefined. For this reason the Gumbel-Max trick is not a suitable reparameterization for use in SCGs with AD. Here we introduce the Concrete distribution motivated by considering a graph, which is the same as Figure 1a up to a continuous relaxation of the argmax computation, see Figure 1b. This will ultimately allow the optimization of parameters α_k via gradients.



(a) $\lambda = 0$ (b) $\lambda = 1/2$ (c) $\lambda = 1$ (d) $\lambda = 2$

Figure 2: A discrete distribution with unnormalized probabilities $(\alpha_1, \alpha_2, \alpha_3) = (2, 0.5, 1)$ and three corresponding Concrete densities at increasing temperatures λ. Each triangle represents the set of points (y_1, y_2, y_3) in the simplex $\Delta^2 = \{(y_1, y_2, y_3) \mid y_k \in (0, 1), y_1 + y_2 + y_3 = 1\}$. For $\lambda = 0$ the size of white circles represents the mass assigned to each vertex of the simplex under the discrete distribution. For $\lambda \in \{2, 1, 0.5\}$ the intensity of the shading represents the value of $p_{\alpha,\lambda}(y)$.

The argmax computation returns states on the vertices of the simplex $\Delta^{n-1} = \{x \in \mathbb{R}^n \mid x_k \in [0, 1], \sum_{k=1}^n x_k = 1\}$. The idea behind Concrete random variables is to relax the state of a discrete variable from the vertices into the interior where it is a random probability vector—a vector of numbers between 0 and 1 that sum to 1. To sample a Concrete random variable $X \in \Delta^{n-1}$ at temperature $\lambda \in (0, \infty)$ with parameters $\alpha_k \in (0, \infty)$, sample $G_k \sim \text{Gumbel}$ i.i.d. and set

$$X_k = \frac{\exp((\log \alpha_k + G_k)/\lambda)}{\sum_{i=1}^n \exp((\log \alpha_i + G_i)/\lambda)}. \tag{10}$$

The softmax computation of (10) smoothly approaches the discrete argmax computation as $\lambda \to 0$ while preserving the relative order of the Gumbels $\log \alpha_k + G_k$. So, imagine making a series of forward passes on the graphs of Figure 1. Both graphs return a stochastic value for each forward pass, but for smaller temperatures the outputs of Figure 1b become more discrete and eventually indistinguishable from a typical forward pass of Figure 1a.

The distribution of X sampled via (10) has a closed form density on the simplex. Because there may be other ways to sample a Concrete random variable, we take the density to be its definition.

Definition 1 (Concrete Random Variables)**.** Let $\alpha \in (0, \infty)^n$ and $\lambda \in (0, \infty)$. $X \in \Delta^{n-1}$ has a Concrete distribution $X \sim \text{Concrete}(\alpha, \lambda)$ with location α and temperature λ, if its density is:

$$p_{\alpha,\lambda}(x) = (n-1)!\, \lambda^{n-1} \prod_{k=1}^n \left(\frac{\alpha_k x_k^{-\lambda-1}}{\sum_{i=1}^n \alpha_i x_i^{-\lambda}} \right). \tag{11}$$

Proposition 1 lists a few properties of the Concrete distribution. (a) is confirmation that our definition corresponds to the sampling routine (10). (b) confirms that rounding a Concrete random variable results in the discrete random variable whose distribution is described by the logits $\log \alpha_k$, (c) confirms that taking the zero temperature limit of a Concrete random variable is the same as rounding. Finally, (d) is a convexity result on the density. We prove these results in Appendix A.

Proposition 1 (Some Properties of Concrete Random Variables)**.** *Let $X \sim \text{Concrete}(\alpha, \lambda)$ with location parameters $\alpha \in (0, \infty)^n$ and temperature $\lambda \in (0, \infty)$, then*

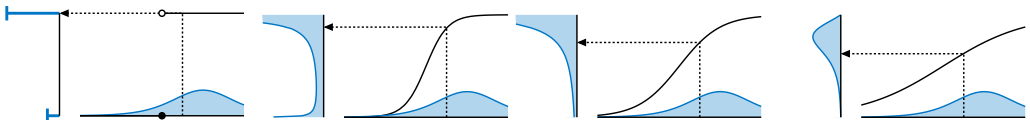
(a) (Reparameterization) If $G_k \sim \text{Gumbel}$ i.i.d., then $X_k \overset{d}{=} \frac{\exp((\log \alpha_k + G_k)/\lambda)}{\sum_{i=1}^n \exp((\log \alpha_i + G_i)/\lambda)}$,

(b) (Rounding) $\mathbb{P}\left(X_k > X_i \text{ for } i \neq k\right) = \alpha_k / (\sum_{i=1}^n \alpha_i)$,

(c) (Zero temperature) $\mathbb{P}\left(\lim_{\lambda \to 0} X_k = 1\right) = \alpha_k / (\sum_{i=1}^n \alpha_i)$,

(d) (Convex eventually) If $\lambda \leq (n-1)^{-1}$, then $p_{\alpha,\lambda}(x)$ is log-convex in x.

The binary case of the Gumbel-Max trick simplifies to passing additive noise through a step function. The corresponding Concrete relaxation is implemented by passing additive noise through a sigmoid—see Figure 3. We cover this more thoroughly in Appendix B.



(a) $\lambda = 0$ (b) $\lambda = 1/2$ (c) $\lambda = 1$ (d) $\lambda = 2$

Figure 3: A visualization of the binary special case. (a) shows the discrete trick, which works by passing a noisy logit through the unit step function. (b), (c), (d) show Concrete relaxations; the horizontal blue densities show the density of the input distribution and the vertical densities show the corresponding Binary Concrete density on $(0, 1)$ for varying λ.

3.3 Concrete Relaxations

Concrete random variables may have some intrinsic value, but we investigate them simply as surrogates for optimizing a SCG with discrete nodes. When it is computationally feasible to integrate over the discreteness, that will always be a better choice. Thus, we consider the use case of optimizing a large graph with discrete stochastic nodes from samples.

First, we outline our proposal for how to use Concrete relaxations by considering a variational autoencoder with a single discrete latent variable. Let $P_a(d)$ be the mass function of some n-dimensional one-hot discrete random variable with unnormalized probabilities $a \in (0, \infty)^n$ and $p_\theta(x|d)$ some distribution over a data point x given $d \in (0, 1)^n$ one-hot. The generative model is then $p_{\theta,a}(x, d) = p_\theta(x|d)P_a(d)$. Let $Q_\alpha(d|x)$ be an approximating posterior over $d \in (0, 1)^n$ one-hot whose unnormalized probabilities $\alpha(x) \in (0, \infty)^n$ depend on x. All together the variational lowerbound we care about stochastically optimizing is

$$\mathcal{L}_1(\theta, a, \alpha) = \mathop{\mathbb{E}}_{D \sim Q_\alpha(d|x)} \left[\log \frac{p_\theta(x|D)P_a(D)}{Q_\alpha(D|x)} \right], \tag{12}$$

with respect to θ, a, and any parameters of α. First, we relax the stochastic computation $D \sim \text{Discrete}(\alpha(x))$ by replacing D with a Concrete random variable $Z \sim \text{Concrete}(\alpha(x), \lambda_1)$ with density $q_{\alpha,\lambda_1}(z|x)$. Simply replacing every instance of D with Z in Eq. 12 will result in a non-interpretable objective, which does not necessarily lowerbound $\log p(x)$, because $\mathbb{E}_{Z \sim q_{\alpha,\lambda_1}(a|x)}[-\log Q_\alpha(Z|x)/P_a(Z)]$ is not a KL divergence. Thus we propose "relaxing" the terms $P_a(d)$ and $Q_\alpha(d|x)$ to reflect the true sampling distribution. Thus, the relaxed objective is:

$$\mathcal{L}_1(\theta, a, \alpha) \overset{\text{relax}}{\rightsquigarrow} \mathop{\mathbb{E}}_{Z \sim q_{\alpha,\lambda_1}(z|x)} \left[\log \frac{p_\theta(x|Z)p_{a,\lambda_2}(Z)}{q_{\alpha,\lambda_1}(Z|x)} \right] \tag{13}$$

where $p_{a,\lambda_2}(z)$ is a Concrete density with location a and temperature λ_2. At test time we evaluate the discrete lowerbound $\mathcal{L}_1(\theta, a, \alpha)$.

Thus, the basic paradigm we propose is the following: during training replace every discrete node with a Concrete node at some fixed temperature (or with an annealing schedule). The graphs are identical up to the softmax / argmax computations, so the parameters of the relaxed graph and discrete graph are the same. When an objective depends on the log-probability of discrete variables in the SCG, as the variational lowerbound does, we propose that the log-probability terms are also "relaxed" to represent the true distribution of the relaxed node. At test time the original discrete loss is evaluated. This is possible, because the discretization of any Concrete distribution has a closed form mass function, and the relaxation of any discrete distribution into a Concrete distribution has a closed form density. This is not always possible. For example, the multinomial probit model—the Gumbel-Max trick with Gaussians replacing Gumbels—does not have a closed form mass.

The success of Concrete relaxations will depend on the choice of temperature during training. It is important that the relaxed nodes are not able to represent a precise real valued mode in the interior of the simplex as in Figure 2d. If this is the case, it is possible for the relaxed random variable to communicate much more than $\log_2(n)$ bits of information about its α parameters. This might lead the relaxation to prefer the interior of the simplex to the vertices, and as a result there will be a large integrality gap in the overall performance of the discrete graph. Therefore Proposition 1 (d) is a conservative guideline for generic n-ary Concrete relaxations; at temperatures lower than $(n-1)^{-1}$ we are guaranteed not to have any modes in the interior for any $\alpha \in (0, \infty)^n$. Ultimately the best choice of λ and the performance of the relaxation for any specific n will be an empirical question.

4 RELATED WORK

Perhaps the most common distribution over the simplex is the Dirichlet with density $p_\alpha(x) \propto \prod_{k=1}^{n} x_k^{\alpha_k - 1}$ on $x \in \Delta^{n-1}$. The Dirichlet can be characterized by strong independence properties, and a great deal of work has been done to generalize it [61, 6, 215, 77]. Of note is the Logistic Normal distribution [11], which can be simulated by taking the softmax of $n-1$ normal random variables and an nth logit that is deterministically zero. The Logistic Normal is an important distribution, because it can effectively model correlations within the simplex. To our knowledge the Concrete distribution does not fall completely into any family of distributions previously described. For $\lambda \leq 1$ the Concrete is in a class of normalized infinitely divisible distributions (S. Favaro, personal communication), and the results of [77] apply.

The idea of using a softmax of Gumbels as a relaxation for a discrete random variable was concurrently considered by [124], where it was called the Gumbel-Softmax. They do not use the density in the relaxed objective, opting instead to compute all aspects of the graph, including discrete log-probability computations, with the relaxed stochastic state of the graph. In the case of variational inference, this relaxed objective is not a lower bound on the marginal likelihood of the observations, and care needs to be taken when optimizing it. The idea of using sigmoidal functions with additive input noise to approximate discreteness is also not a new idea. [85] introduced nonlinear Gaussian units which computed their activation by passing Gaussian noise with the mean and variance specified by the input to the unit through a nonlinearity, such as the logistic function. [227] binarized real-valued codes of an autoencoder by adding (Gaussian) noise to the logits before passing them through the logistic function. Most recently, to avoid the difficulty associated with likelihood-ratio methods [138] relaxed the discrete sampling operation by sampling a vector of Gaussians instead and passing those through a softmax.

There is another family of gradient estimators that have been studied in the context of training neural networks with discrete units. These are usually collected under the umbrella of straight-through estimators [212]. The basic idea they use is passing forward discrete values, but taking gradients through the expected value. They have good empirical performance, but have not been shown to be the estimators of any loss function. This is in contrast to gradients from Concrete relaxations, which are biased with respect to the discrete graph, but unbiased with respect to the continuous one.

5 EXPERIMENTS

5.1 PROTOCOL

The aim of our experiments was to evaluate the effectiveness of the gradients of Concrete relaxations for optimizing SCGs with discrete nodes. We considered the tasks in [175]: structured output prediction and density estimation. Both tasks are difficult optimization problems involving fitting probability distributions with hundreds of latent discrete nodes. We compared the performance of Concrete reparameterizations to two state-of-the-art score function estimators: VIMCO [175] for optimizing the multisample variational objective ($m > 1$) and NVIL [172] for optimizing the single-sample one ($m = 1$). We performed the experiments using the MNIST and Omniglot datasets. These are datasets of 28×28 images of handwritten digits (MNIST) or letters (Omniglot). For MNIST we used the fixed binarization of [228] and the standard 50,000/10,000/10,000 split into training/validation/testing sets. For Omniglot we sampled a fixed binarization and used the standard 24,345/8,070 split into training/testing sets. We report the negative log-likelihood (NLL) of the discrete graph on the test data as the performance metric.

All of our models were neural networks with layers of n-ary discrete stochastic nodes with values on the corners of the hypercube $\{-1, 1\}^{\log_2(n)}$. The distributions were parameterized by n real values $\log \alpha_k \in \mathbb{R}$, which we took to be the logits of a discrete random variable $D \sim \text{Discrete}(\alpha)$ with n states. Model descriptions are of the form "(200V–200H∼784V)", read from left to right. This describes the order of conditional sampling, again from left to right, with each integer representing the number of stochastic units in a layer. The letters V and H represent observed and latent variables, respectively. If the leftmost layer is H, then it was sampled unconditionally from some parameters. Conditioning functions are described by {–, ∼}, where "–" means a linear function of the previous layer and "∼" means a non-linear function. A "layer" of these units is simply the concatenation

binary model	m	MNIST NLL				Omniglot NLL			
		Test		Train		Test		Train	
		Concrete	VIMCO	Concrete	VIMCO	Concrete	VIMCO	Concrete	VIMCO
(200H – 784V)	1	107.3	**104.4**	107.5	**104.2**	118.7	**115.7**	117.0	**112.2**
	5	104.9	**101.9**	104.9	**101.5**	118.0	**113.5**	115.8	**110.8**
	50	104.3	**98.8**	104.2	**98.3**	118.9	**113.0**	115.8	**110.0**
(200H – 200H – 784V)	1	102.1	**92.9**	102.3	**91.7**	116.3	**109.2**	114.4	**104.8**
	5	99.9	**91.7**	100.0	**90.8**	116.0	**107.5**	113.5	**103.6**
	50	99.5	**90.7**	99.4	**89.7**	117.0	**108.1**	113.9	**103.6**
(200H ∼784V)	1	**92.1**	93.8	**91.2**	91.5	**108.4**	116.4	**103.6**	110.3
	5	**89.5**	91.4	**88.1**	88.6	**107.5**	118.2	**101.4**	102.3
	50	**88.5**	89.3	**86.4**	86.5	**108.1**	116.0	**100.5**	100.8
(200H ∼200H ∼784V)	1	**87.9**	88.4	86.5	**85.8**	**105.9**	111.7	**100.2**	105.7
	5	**86.3**	86.4	84.1	**82.5**	**105.8**	108.2	**98.6**	101.1
	50	85.7	**85.5**	83.1	**81.8**	**106.8**	113.2	97.5	**95.2**

Table 1: Density estimation with binary latent variables. When $m = 1$, VIMCO stands for NVIL.

of some number of independent nodes whose parameters are determined as a function the previous layer. For example a 240 binary layer is a factored distribution over the $\{-1, 1\}^{240}$ hypercube. Whereas a 240 8-ary layer can be seen as a distribution over the same hypercube where each of the 80 triples of units are sampled independently from an 8 way discrete distribution over $\{-1, 1\}^3$. All models were initialized with the heuristic of [97] and optimized using Adam [135]. All temperatures were fixed throughout training. Appendix C for hyperparameter details.

5.2 Density Estimation

Density estimation, or generative modelling, is the problem of fitting the distribution of data. We took the latent variable approach described in Section 2.4 and trained the models by optimizing the variational objective $\mathcal{L}_m(\theta, \phi)$ given by Eq. 8 averaged uniformly over minibatches of data points x. Both our generative models $p_\theta(z,\ x)$ and variational distributions $q_\phi(z \mid x)$ were parameterized with neural networks as described above. We trained models with $\mathcal{L}_m(\theta, \phi)$ for $m \in \{1, 5, 50\}$ and approximated the NLL with $\mathcal{L}_{50,000}(\theta, \phi)$ averaged uniformly over the whole dataset.

The results are shown in Table 1. In general, VIMCO outperformed Concrete relaxations for linear models and Concrete relaxations outperformed VIMCO for non-linear models. We also tested the effectiveness of Concrete relaxations on generative models with n-ary layers on the $\mathcal{L}_5(\theta, \phi)$ objective. The best 4-ary model achieved test/train NLL 86.7/83.3, the best 8-ary achieved 87.4/84.6 with Concrete relaxations. The relatively poor performance of the 8-ary model may be because moving from 4 to 8 results in a more difficult objective without much added capacity. As a control we trained n-ary models using logistic normals as relaxations of discrete distributions (with retuned temperature hyperparameters). Because the discrete zero temperature limit of logistic Normals is a multinomial probit whose mass function is not known, we evaluated the discrete model by sampling from the discrete distribution parameterized by the logits learned during training. The best 4-ary model achieved test/train NLL of 88.7/85.0, the best 8-ary model achieved 89.1/85.1.

5.3 Structured Output Prediction

Structured output prediction is concerned with modelling the high-dimensional distribution of the observation given a context and can be seen as conditional density estimation. We considered the task of predicting the bottom half x_1 of an image of an MNIST digit given its top half x_2, as introduced by [212]. We followed [212] in using a model with layers of discrete stochastic units between the context and the observation. Conditioned on the top half x_2 the network samples from a distribution $p_\phi(z \mid x_2)$ over layers of stochastic units z then predicts x_1 by sampling from a distribution $p_\theta(x_1 \mid z)$. The training objective for a single pair (x_1, x_2) is

$$\mathcal{L}_m^{SP}(\theta, \phi) = \mathop{\mathbb{E}}_{Z_i \sim p_\phi(z|x_2)} \left[\log \left(\frac{1}{m} \sum_{i=1}^{m} p_\theta(x_1 \mid Z_i) \right) \right].$$

binary model	m	Test NLL		Train NLL	
		Concrete	VIMCO	Concrete	VIMCO
(392V–240H –240H–392V)	1	**58.5**	61.4	**54.2**	59.3
	5	**54.3**	54.5	**49.2**	52.7
	50	53.4	**51.8**	**48.2**	49.6
(392V–240H –240H–240H –392V)	1	**56.3**	59.7	**51.6**	58.4
	5	**52.7**	53.5	**46.9**	51.6
	50	52.0	**50.2**	**45.9**	47.9

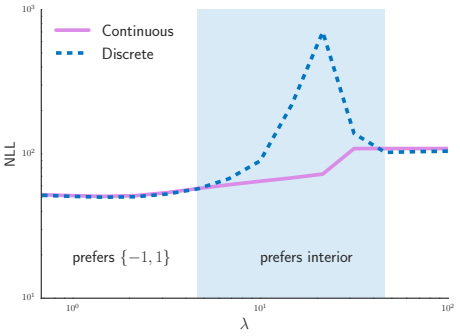


Figure 4: Results for structured prediction on MNIST comparing Concrete relaxations to VIMCO. When $m = 1$ VIMCO stands for NVIL. The plot on the right shows the objective (lower is better) for the continuous and discrete graph trained at temperatures λ. In the shaded region, units prefer to communicate real values in the interior of $(-1, 1)$ and the discretization suffers an integrality gap.

This objective is a special case of $\mathcal{L}_m(\theta, \phi)$ (Eq. 8) where we use the prior $p_\phi(z|x_2)$ as the variational distribution. Thus, the objective is a lower bound on $\log p_{\theta,\phi}(x_1 \mid x_2)$.

We trained the models by optimizing $\mathcal{L}_m^{SP}(\theta, \phi)$ for $m \in \{1, 5, 50\}$ averaged uniformly over mini-batches and evaluated them by computing $\mathcal{L}_{100}^{SP}(\theta, \phi)$ averaged uniformly over the entire dataset. The results are shown in Figure 4. Concrete relaxations more uniformly outperformed VIMCO in this instance. We also trained n-ary (392V–240H–240H–240H–392V) models on the $\mathcal{L}_1^{SP}(\theta, \phi)$ objective using the best temperature hyperparameters from density estimation. 4-ary achieved a test/train NLL of 55.4/46.0 and 8-ary achieved 54.7/44.8. As opposed to density estimation, increasing arity uniformly improved the models. We also investigated the hypothesis that for higher temperatures Concrete relaxations might prefer the interior of the interval to the boundary points $\{-1, 1\}$. Figure 4 was generated with binary (392V–240H–240H–240H–392V) model trained on $\mathcal{L}_1^{SP}(\theta, \phi)$.

6 Conclusion

We introduced the Concrete distribution, a continuous relaxation of discrete random variables. The Concrete distribution is a new distribution on the simplex with a closed form density parameterized by a vector of positive location parameters and a positive temperature. Crucially, the zero temperature limit of every Concrete distribution corresponds to a discrete distribution, and any discrete distribution can be seen as the discretization of a Concrete one. The application we considered was training stochastic computation graphs with discrete stochastic nodes. The gradients of Concrete relaxations are biased with respect to the original discrete objective, but they are low variance unbiased estimators of a continuous surrogate objective. We showed in a series of experiments that stochastic nodes with Concrete distributions can be used effectively to optimize the parameters of a stochastic computation graph with discrete stochastic nodes. We did not find that annealing or automatically tuning the temperature was important for these experiments, but it remains interesting and possibly valuable future work.

Acknowledgments

We thank Jimmy Ba for the excitement and ideas in the early days, Stefano Favarro for some analysis of the distribution. We also thank Gabriel Barth-Maron and Roger Grosse.

A Proof of Proposition 1

Let $X \sim \text{Concrete}(\alpha, \lambda)$ with location parameters $\alpha \in (0, \infty)^n$ and temperature $\lambda \in (0, \infty)$.

1. Let $G_k \sim \text{Gumbel}$ i.i.d., consider

$$Y_k = \frac{\exp((\log \alpha_k + G_k)/\lambda)}{\sum_{i=1}^n \exp((\log \alpha_i + G_i)/\lambda)}$$

Let $Z_k = \log \alpha_k + G_k$, which has density

$$\alpha_k \exp(-z_k) \exp(-\alpha_k \exp(-z_k))$$

We will consider the invertible transformation

$$F(z_1, \ldots, z_n) = (y_1, \ldots, y_{n-1}, c)$$

where

$$y_k = \exp(z_k/\lambda)c^{-1}$$
$$c = \sum_{i=1}^{n} \exp(z_i/\lambda)$$

then

$$F^{-1}(y_1, \ldots, y_{n-1}, c) = (\lambda(\log y_1 + \log c), \ldots, \lambda(\log y_{n-1} + \log c), \lambda(\log y_n + \log c))$$

where $y_n = 1 - \sum_{i=1}^{n-1} y_i$. This has Jacobian

$$\begin{bmatrix} \lambda y_1^{-1} & 0 & 0 & 0 & \ldots & 0 & \lambda c^{-1} \\ 0 & \lambda y_2^{-1} & 0 & 0 & \ldots & 0 & \lambda c^{-1} \\ 0 & 0 & \lambda y_3^{-1} & 0 & \ldots & 0 & \lambda c^{-1} \\ & & \vdots & & & & \\ -\lambda y_n^{-1} & -\lambda y_n^{-1} & -\lambda y_n^{-1} & -\lambda y_n^{-1} & \ldots & -\lambda y_n^{-1} & \lambda c^{-1} \end{bmatrix}$$

by adding y_i/y_n times each of the top $n-1$ rows to the bottom row we see that this Jacobian has the same determinant as

$$\begin{bmatrix} \lambda y_1^{-1} & 0 & 0 & 0 & \ldots & 0 & \lambda c^{-1} \\ 0 & \lambda y_2^{-1} & 0 & 0 & \ldots & 0 & \lambda c^{-1} \\ 0 & 0 & \lambda y_3^{-1} & 0 & \ldots & 0 & \lambda c^{-1} \\ & & \vdots & & & & \\ 0 & 0 & 0 & 0 & \ldots & 0 & \lambda (c y_n)^{-1} \end{bmatrix}$$

and thus the determinant is equal to

$$\frac{\lambda^n}{c \prod_{i=1}^{k} y_i}$$

all together we have the density

$$\frac{\lambda^n \prod_{k=1}^{n} \alpha_k \exp(-\lambda \log y_k - \lambda \log c) \exp(-\alpha_k \exp(-\lambda \log y_k - \lambda \log c))}{c \prod_{i=1}^{n} y_i}$$

with $r = \log c$ change of variables we have density

$$\frac{\lambda^n \prod_{k=1}^{n} \alpha_k \exp(-\lambda r) \exp(-\alpha_k \exp(-\lambda \log y_k - \lambda r))}{\prod_{i=1}^{n} y_i^{\lambda+1}} =$$
$$\frac{\lambda^n \prod_{k=1}^{n} \alpha_k}{\prod_{i=1}^{n} y_i^{\lambda+1}} \exp(-n\lambda r) \exp(-\sum_{i=1}^{n} \alpha_i \exp(-\lambda \log y_i - \lambda r)) =$$

letting $\gamma = \log(\sum_{n=1}^{n} \alpha_k y_k^{-\lambda})$

$$\frac{\lambda^n \prod_{k=1}^{n} \alpha_k}{\prod_{i=1}^{n} y_i^{\lambda+1} \exp(\gamma)} \exp(-n\lambda r + \gamma) \exp(-\exp(-\lambda r + \gamma)) =$$

integrating out r

$$\frac{\lambda^n \prod_{k=1}^{n} \alpha_k}{\prod_{i=1}^{n} y_i^{\lambda+1} \exp(\gamma)} \left(\frac{\exp(-\gamma n + \gamma)\Gamma(n)}{\lambda} \right) =$$
$$\frac{\lambda^{n-1} \prod_{k=1}^{n} \alpha_k}{\prod_{i=1}^{n} y_i^{\lambda+1}} (\exp(-\gamma n)\Gamma(n)) =$$
$$(n-1)! \lambda^{n-1} \frac{\prod_{k=1}^{n} \alpha_k y_k^{-\lambda-1}}{(\sum_{n=1}^{n} \alpha_k y_k^{-\lambda})^n}$$

Thus $Y \stackrel{d}{=} X$.

2. Follows directly from (a) and the Gumbel-Max trick [160].
3. Follows directly from (a) and the Gumbel-Max trick [160].
4. Let $\lambda \leq (n-1)^{-1}$. The density of X can be rewritten as

$$\begin{aligned} p_{\alpha,\lambda}(x) &\propto \prod_{k=1}^{n} \frac{\alpha_k y^{-\lambda-1}}{\sum_{i=1}^{n} \alpha_i y_i^{-\lambda}} \\ &= \prod_{k=1}^{n} \frac{\alpha_k y_k^{\lambda(n-1)-1}}{\sum_{i=1}^{n} \alpha_i \prod_{j \neq i} y_j^{\lambda}} \end{aligned}$$

Thus, the log density is up to an additive constant C

$$\log p_{\alpha,\lambda}(x) = \sum_{k=1}^{n} (\lambda(n-1)-1) \log y_k - n \log \left(\sum_{k=1}^{n} \alpha_k \prod_{j \neq k} y_j^{\lambda} \right) + C$$

If $\lambda \leq (n-1)^{-1}$, then the first n terms are convex, because $-\log$ is convex. For the last term, $-\log$ is convex and non-increasing and $\prod_{j \neq k} y_j^{\lambda}$ is concave for $\lambda \leq (n-1)^{-1}$. Thus, their composition is convex. The sum of convex terms is convex, finishing the proof.

B THE BINARY SPECIAL CASE

Bernoulli random variables are an important special case of discrete distributions taking states in $\{0, 1\}$. Here we consider the binary special case of the Gumbel-Max trick from Figure 1a along with the corresponding Concrete relaxation.

Let $D \sim \text{Discrete}(\alpha)$ for $\alpha \in (0, \infty)^2$ be a two state discrete random variable on $\{0, 1\}^2$ such that $D_1 + D_2 = 1$, parameterized as in Figure 1a by $\alpha_1, \alpha_2 > 0$:

$$\mathbb{P}(D_1 = 1) = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{14}$$

The distribution is degenerate, because $D_1 = 1 - D_2$. Therefore we consider just D_1. Under the Gumbel-Max reparameterization, the event that $D_1 = 1$ is the event that $\{G_1 + \log \alpha_1 > G_2 + \log \alpha_2\}$ where $G_k \sim \text{Gumbel}$ i.i.d. The difference of two Gumbels is a Logistic distribution $G_1 - G_2 \sim \text{Logistic}$, which can be sampled in the following way, $G_1 - G_2 \stackrel{d}{=} \log U - \log(1 - U)$ where $U \sim \text{Uniform}(0, 1)$. So, if $\alpha = \alpha_1/\alpha_2$, then we have

$$\mathbb{P}(D_1 = 1) = \mathbb{P}(G_1 + \log \alpha_1 > G_2 + \log \alpha_2) = \mathbb{P}(\log U - \log(1 - U) + \log \alpha > 0) \tag{15}$$

Thus, $D_1 \stackrel{d}{=} H(\log \alpha + \log U - \log(1 - U))$, where H is the unit step function.

Correspondingly, we can consider the Binary Concrete relaxation that results from this process. As in the n-ary case, we consider the sampling routine for a Binary Concrete random variable $X \in (0, 1)$ first. To sample X, sample $L \sim \text{Logistic}$ and set

$$X = \frac{1}{1 + \exp(-(\log \alpha + L)/\lambda)} \tag{16}$$

We define the Binary Concrete random variable X by its density on the unit interval.

Definition 2 (Binary Concrete Random Variables)**.** Let $\alpha \in (0, \infty)$ and $\lambda \in (0, \infty)$. $X \in (0, 1)$ has a Binary Concrete distribution $X \sim \text{BinConcrete}(\alpha, \lambda)$ with location α and temperature λ, if its density is:

$$p_{\alpha,\lambda}(x) = \frac{\lambda \alpha x^{-\lambda-1}(1-x)^{-\lambda-1}}{(\alpha x^{-\lambda} + (1-x)^{-\lambda})^2}. \tag{17}$$

We state without proof the special case of Proposition 1 for Binary Concrete distributions

Proposition 2 (Some Properties of Binary Concrete Random Variables)**.** *Let* $X \sim \text{BinConcrete}(\alpha, \lambda)$ *with location parameter* $\alpha \in (0, \infty)$ *and temperature* $\lambda \in (0, \infty)$*, then*

(a) (Reparameterization) If $L \sim \text{Logistic}$*, then* $X \stackrel{d}{=} \frac{1}{1+\exp(-(\log\alpha+L)/\lambda)}$*,*

(b) (Rounding) $\mathbb{P}(X > 0.5) = \alpha/(1+\alpha)$*,*

(c) (Zero temperature) $\mathbb{P}(\lim_{\lambda\to 0} X = 1) = \alpha/(1+\alpha)$*,*

(d) (Convex eventually) If $\lambda \leq 1$*, then* $p_{\alpha,\lambda}(x)$ *is log-convex in* x*.*

We can generalize the binary circuit beyond Logistic random variables. Consider an arbitrary random variable X with infinite support on $\mathbb{R}$. If $\Phi : \mathbb{R} \to [0,1]$ is the CDF of X, then

$$\mathbb{P}(H(X) = 1) = 1 - \Phi(0)$$

If we want this to have a Bernoulli distribution with probability $\alpha/(1+\alpha)$, then we should solve the equation

$$1 - \Phi(0) = \frac{\alpha}{1+\alpha}.$$

This gives $\Phi(0) = 1/(1+\alpha)$, which can be accomplished by relocating the random variable Y with CDF Φ to be $X = Y - \Phi^{-1}(1/(1+\alpha))$.

C Experimental Details

The basic model architectures we considered are exactly analogous to those in [50] with Concrete/discrete random variables replacing Gaussians.

C.1 — vs ∼

The conditioning functions we used were either linear or non-linear. Non-linear consisted of two $\tanh$ layers of the same size as the preceding stochastic layer in the computation graph.

C.2 n-ary layers

All our models are neural networks with layers of n-ary discrete stochastic nodes with $\log_2(n)$-dimensional states on the corners of the hypercube $\{-1,1\}^{\log_2(n)}$. For a generic n-ary node sampling proceeds as follows. Sample a n-ary discrete random variable $D \sim \text{Discrete}(\alpha)$ for $\alpha \in (0,\infty)^n$. If C is the $\log_2(n) \times n$ matrix, which lists the corners of the hypercube $\{-1,1\}^{\log_2(n)}$ as columns, then we took $Y = CD$ as downstream computation on D. The corresponding Concrete relaxation is to take $X \sim \text{Concrete}(\alpha,\lambda)$ for some fixed temperature $\lambda \in (0,\infty)$ and set $\tilde{Y} = CX$. For the binary case, this amounts to simply sampling $U \sim \text{Uniform}(0,1)$ and taking $Y = 2H(\log U - \log(1-U) + \log\alpha) - 1$. The corresponding Binary Concrete relaxation is $\tilde{Y} = 2\sigma((\log U - \log(1-U) + \log\alpha)/\lambda) - 1$.

C.3 Bias Initialization

All biases were initialized to 0 with the exception of the biases in the prior decoder distribution over the 784 or 392 observed units. These were initialized to the logit of the base rate averaged over the respective dataset (MNIST or Omniglot).

C.4 Centering

We also found it beneficial to center the layers of the inference network during training. The activity in $(-1,1)^d$ of each stochastic layer was centered during training by maintaining a exponentially decaying average with rate 0.9 over minibatches. This running average was subtracted from the activity of the layer *before* it was updated. Gradients did not flow throw this computation, so it simply amounted to a dynamic offset. The averages were *not* updated during the evaluation.

C.5 HYPERPARAMETER SELECTION

All models were initialized with the heuristic of [97] and optimized using Adam [135] with parameters $\beta_1 = 0.9, \beta_2 = 0.999$ for 10^7 steps on minibatches of size 64. Hyperparameters were selected on the MNIST dataset by grid search taking the values that performed best on the validation set. Learning rates were chosen from $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\}$ and weight decay from $\{0, 10^{-2}, 10^{-1}, 1\}$. Two sets of hyperparameters were selected, one for linear models and one for non-linear models. The linear models' hyperparameters were selected with the 200H–200H–784V density model on the $\mathcal{L}_5(\theta, \phi)$ objective. The non-linear models' hyperparameters were selected with the 200H$\sim$200H$\sim$784V density model on the $\mathcal{L}_5(\theta, \phi)$ objective. For density estimation, the Concrete relaxation hyperparameters were (weight decay = 0, learning rate = $3 \cdot 10^{-4}$) for linear and (weight decay = 0, learning rate = 10^{-4}) for non-linear. For structured prediction Concrete relaxations used (weight decay = 10^{-3}, learning rate = $3 \cdot 10^{-4}$).

In addition to tuning learning rate and weight decay, we tuned temperatures for the Concrete relaxations on the density estimation task. We found it valuable to have different values for the prior and posterior distributions. In particular, for binary we found that (prior $\lambda_2 = 1/2$, posterior $\lambda_1 = 2/3$) was best, for 4-ary we found (prior $\lambda_2 = 2/3$, posterior $\lambda_1 = 1$) was best, and (prior $\lambda_2 = 2/5$, posterior $\lambda_1 = 2/3$) for 8-ary. No temperature annealing was used. For structured prediction we used just the corresponding posterior λ_1 as the temperature for the whole graph, as there was no variational posterior.

We performed early stopping when training with the score function estimators (VIMCO/NVIL) as they were much more prone to overfitting.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

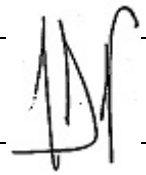
Title of Paper	The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables
Publication Status	☒Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, 2017.

Student Confirmation

Student Name:	Chris J. Maddison		
Contribution to the Paper	• I proposed the idea, and did all of the analysis in the paper. • I wrote all of the experimental code with the exception of the VIMCO and NVIL baselines, which were written by Andriy Mnih. • I wrote the majority of the paper, with the exception of the background section, which was written by Andriy Mnih. • All authors contributed to the development of the paper through discussions and ideas, and all authors reviewed the final draft.		
Signature		Date	05 May 2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Arnaud Doucet			
Supervisor comments I agree that the candidate has made a substantial contribution to the publication.			
Signature		Date	05 May 2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Filtering Variational Objectives

Chris J. Maddison[1,3,*], Dieterich Lawson,[2,*] George Tucker[2,*]
Nicolas Heess[1], Mohammad Norouzi[2], Andriy Mnih[1], Arnaud Doucet[3], Yee Whye Teh[1]

[1]DeepMind, [2]Google Brain, [3]University of Oxford
{cmaddis, dieterichl, gjt}@google.com

Abstract

When used as a surrogate objective for maximum likelihood estimation in latent variable models, the evidence lower bound (ELBO) produces state-of-the-art results. Inspired by this, we consider the extension of the ELBO to a family of lower bounds defined by a particle filter's estimator of the marginal likelihood, the *filtering variational objectives* (FIVOs). FIVOs take the same arguments as the ELBO, but can exploit a model's sequential structure to form tighter bounds. We present results that relate the tightness of FIVO's bound to the variance of the particle filter's estimator by considering the generic case of bounds defined as log-transformed likelihood estimators. Experimentally, we show that training with FIVO results in substantial improvements over training the same model architecture with the ELBO on sequential data.

1 Introduction

Learning in statistical models via gradient descent is straightforward when the objective function and its gradients are tractable. In the presence of latent variables, however, many objectives become intractable. For neural generative models with latent variables, there are currently a few dominant approaches: optimizing lower bounds on the marginal log-likelihood [137, 218], restricting to a class of invertible models [67], or using likelihood-free methods [100, 196, 248, 176]. In this work, we focus on the first approach and introduce *filtering variational objectives* (FIVOs), a tractable family of objectives for maximum likelihood estimation (MLE) in latent variable models with sequential structure.

Specifically, let x denote an observation of an $\mathcal{X}$-valued random variable. We assume that the process generating x involves an unobserved $\mathcal{Z}$-valued random variable z with joint density $p(x, z)$ in some family $\mathcal{P}$. The goal of MLE is to recover $p \in \mathcal{P}$ that maximizes the marginal log-likelihood, $\log p(x) = \log \left(\int p(x, z)\, dz \right)$[1]. The difficulty in carrying out this optimization is that the log-likelihood function is defined via a generally intractable integral. To circumvent marginalization, a common approach [137, 218] is to optimize a variational lower bound on the marginal log-likelihood [128, 17]. The evidence lower bound $\mathcal{L}(x, p, q)$ (ELBO) is the most common such bound and is defined by a variational posterior distribution $q(z|x)$ whose support includes p's,

$$\mathcal{L}(x, p, q) = \mathop{\mathbb{E}}_{q(z|x)} \left[\log \frac{p(x, z)}{q(z|x)} \right] = \log p(x) - \mathrm{KL}(q(z|x) \parallel p(z|x)) \leq \log p(x)\ . \tag{1}$$

$\mathcal{L}(x, p, q)$ lower-bounds the marginal log-likelihood for any choice of q, and the bound is tight when q is the true posterior $p(z|x)$. Thus, the joint optimum of $\mathcal{L}(x, p, q)$ in p and q is the MLE. In practice,

[*]Equal contribution.
[1]We reuse p to denote the conditionals and marginals of the joint density.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

it is common to restrict q to a tractable family of distributions (e.g., a factored distribution) and to jointly optimize the ELBO over p and q with stochastic gradient ascent [137, 218, 213, 141]. Because of the KL penalty from q to p, optimizing (1) under these assumptions tends to force p's posterior to satisfy the factorizing assumptions of the variational family which reduces the capacity of the model p. One strategy for addressing this is to decouple the tightness of the bound from the quality of q. For example, [50] observed that Eq. (1) can be interpreted as the log of an unnormalized importance weight with the proposal given by q, and that using N samples from the same proposal produces a tighter bound, known as the importance weighted auto-encoder bound, or IWAE.

Indeed, it follows from Jensen's inequality that the log of *any* unbiased positive Monte Carlo estimator of the marginal likelihood results in a lower bound that can be optimized for MLE. The filtering variational objectives (FIVOs) build on this idea by treating the log of a particle filter's likelihood estimator as an objective function. Following [174], we call objectives defined as log-transformed likelihood estimators Monte Carlo objectives (MCOs). In this work, we show that the tightness of an MCO scales like the relative variance of the estimator from which it is constructed. It is well-known that the variance of a particle filter's likelihood estimator scales more favourably than simple importance sampling for models with sequential structure [53, 25]. Thus, FIVO can potentially form a much tighter bound on the marginal log-likelihood than IWAE.

The main contributions of this work are introducing filtering variational objectives and a more careful study of Monte Carlo objectives. In Section 2, we review maximum likelihood estimation via maximizing the ELBO. In Section 3, we study Monte Carlo objectives and provide some of their basic properties. We define filtering variational objectives in Section 4, discuss details of their optimization, and present a sharpness result. Finally, we cover related work and present experiments showing that sequential models trained with FIVO outperform models trained with ELBO or IWAE in practice.

2 Background

We briefly review techniques for optimizing the ELBO as a surrogate MLE objective. We restrict our focus to latent variable models in which the model $p_\theta(x, z)$ factors into tractable conditionals $p_\theta(z)$ and $p_\theta(x|z)$ that are parameterized differentiably by parameters θ. MLE in these models is then the problem of optimizing $\log p_\theta(x)$ in θ. The expectation-maximization (EM) algorithm is an approach to this problem which can be seen as coordinate ascent, fully maximizing $\mathcal{L}(x, p_\theta, q)$ alternately in q and θ at each iteration [66, 268, 186]. Yet, EM rarely applies in general, because maximizing over q for a fixed θ corresponds to a generally intractable inference problem.

Instead, an approach with mild assumptions on the model is to perform gradient ascent following a Monte Carlo estimator of the ELBO's gradient [120, 213]. We assume that q is taken from a family of distributions parameterized differentiably by parameters ϕ. We can follow an unbiased estimator of the ELBO's gradient by sampling $z \sim q_\phi(z|x)$ and updating the parameters by $\theta' = \theta + \eta \nabla_\theta \log p_\theta(x, z)$ and $\phi' = \phi + \eta(\log p_\theta(x, z) - \log q_\phi(z|x)) \nabla_\phi \log q_\phi(z|x)$, where the gradients are computed conditional on the sample z and η is a learning rate. Such estimators follow the ELBO's gradient in expectation, but variance reduction techniques are usually necessary [213, 173, 174].

A lower variance gradient estimator can be derived if q_ϕ is a reparameterizable distribution [137, 218, 88]. Reparameterizable distributions are those that can be simulated by sampling from a distribution $\epsilon \sim d(\epsilon)$, which does not depend on ϕ, and then applying a deterministic transformation $z = f_\phi(x, \epsilon)$. When p_θ, q_ϕ, and f_ϕ are differentiable, an unbiased estimator of the ELBO gradient consists of sampling ϵ and updating the parameter by $(\theta', \phi') = (\theta, \phi) + \eta \nabla_{(\theta,\phi)} (\log p_\theta(x, f_\phi(x, \epsilon)) - \log q_\phi(f_\phi(x, \epsilon)|x))$. Given ϵ, the gradients of the sampling process can flow through $z = f_\phi(x, \epsilon)$.

Unfortunately, when the variational family of q_ϕ is restricted, following gradients of $-\mathrm{KL}(q_\phi(z|x) \parallel p_\theta(z|x))$ tends to reduce the capacity of the model p_θ to match the assumptions of the variational family. This KL penalty can be "removed" by considering generalizations of the ELBO whose tightness can be controlled by means other than the closenesss of p and q, e.g., [50]. We consider this in the next section.

3 Monte Carlo Objectives (MCOs)

Monte Carlo objectives (MCOs) [174] generalize the ELBO to objectives defined by taking the $\log$ of a positive, unbiased estimator of the marginal likelihood. The key property of MCOs is that they are lower bounds on the marginal log-likelihood, and thus can be used for MLE. Motivated by the previous section, we present results on the convergence of generic MCOs to the marginal log-likelihood and show that the tightness of an MCO is closely related to the variance of the estimator that defines it.

One can verify that the ELBO is a lower bound by using the concavity of log and Jensen's inequality,

$$\mathop{\mathbb{E}}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] \;\leq\; \log \int \frac{p(x,z)}{q(z|x)} q(z|x)\, dz \;=\; \log p(x). \tag{2}$$

This argument only relies only on unbiasedness of $p(x,z)/q(z|x)$ when $z \sim q(z|x)$. Thus, we can generalize this by considering any unbiased marginal likelihood estimator $\hat{p}_N(x)$ and treating $\mathbb{E}[\log \hat{p}_N(x)]$ as an objective function over models p. Here $N \in \mathbb{N}$ indexes the amount of computation needed to simulate $\hat{p}_N(x)$, e.g., the number of samples or particles.

Definition 1. Monte Carlo Objectives. Let $\hat{p}_N(x)$ be an unbiased positive estimator of $p(x)$, $\mathbb{E}[\hat{p}_N(x)] = p(x)$, then the Monte Carlo objective $\mathcal{L}_{\mathrm{N}}(x,p)$ over $p \in \mathcal{P}$ defined by $\hat{p}_N(x)$ is

$$\mathcal{L}_{\mathrm{N}}(x,p) \;=\; \mathbb{E}[\log \hat{p}_N(x)] \tag{3}$$

For example, the ELBO is constructed from a single unnormalized importance weight $\hat{p}(x) = p(x,z)/q(z|x)$. The IWAE bound [50] takes $\hat{p}_N(x)$ to be N averaged i.i.d. importance weights,

$$\mathcal{L}^{\mathrm{IWAE}}_{\mathrm{N}}(x,p,q) \;=\; \mathop{\mathbb{E}}_{q(z^i|x)}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N}\frac{p(x,z^i)}{q(z^i|x)}\right)\right] \tag{4}$$

We consider additional examples in the Appendix. To avoid notational clutter, we omit the arguments to an MCO, e.g., the observations x or model p, when the default arguments are clear from context. Whether we can compute stochastic gradients of $\mathcal{L}_{\mathrm{N}}$ efficiently depends on the specific form of the estimator and the underlying random variables that define it.

Many likelihood estimators $\hat{p}_N(x)$ converge to $p(x)$ almost surely as $N \to \infty$ (known as strong consistency). The advantage of a consistent estimator is that its MCO can be driven towards $\log p(x)$ by increasing N. We present sufficient conditions for this convergence and a description of the rate:

Proposition 1. Properties of Monte Carlo Objectives. *Let $\mathcal{L}_{\mathrm{N}}(x,p)$ be a Monte Carlo objective defined by an unbiased positive estimator $\hat{p}_N(x)$ of $p(x)$. Then,*

- *(a) (Bound) $\mathcal{L}_{\mathrm{N}}(x,p) \leq \log p(x)$.*
- *(b) (Consistency) If $\log \hat{p}_N(x)$ is uniformly integrable (see Appendix for definition) and $\hat{p}_N(x)$ is strongly consistent, then $\mathcal{L}_{\mathrm{N}}(x,p) \to \log p(x)$ as $N \to \infty$.*
- *(c) (Asymptotic Bias) Let $g(N) = \mathbb{E}[(\hat{p}_N(x) - p(x))^6]$ be the 6th central moment. If the 1st inverse moment is bounded, $\limsup_{N\to\infty} \mathbb{E}[\hat{p}_N(x)^{-1}] < \infty$, then*

$$\log p(x) - \mathcal{L}_{\mathrm{N}}(x,p) = \frac{1}{2}\operatorname{var}\left(\frac{\hat{p}_N(x)}{p(x)}\right) + \mathcal{O}(\sqrt{g(N)}). \tag{5}$$

Proof. See the Appendix for the proof and a sufficient condition for controlling the first inverse moment when $\hat{p}_N(x)$ is the average of i.i.d. random variables. □

In some cases, convergence of the bound to $\log p(x)$ is monotonic, e.g., IWAE [50], but this is not true in general. The relative variance of estimators, $\operatorname{var}(\hat{p}_N(x)/p(x))$, tends to be well studied, so property (c) gives us a tool for comparing the convergence rate of distinct MCOs. For example, [53, 25] study marginal likelihood estimators defined by particle filters and find that the relative variance of these estimators scales favorably in comparison to naive importance sampling. This suggests that a particle filter's MCO, introduced in the next section, will generally be a tighter bound than IWAE.

Algorithm 1 Simulating $\mathcal{L}^{\text{FIVO}}_{\text{N}}(x_{1:T}, p, q)$

FIVO$(x_{1:T}, p, q, N)$:
$\{w_0^i\}_{i=1}^N = \{1/N\}_{i=1}^N$
for $t \in \{1, \ldots, T\}$ **do**
  for $i \in \{1, \ldots, N\}$ **do**
    $z_t^i \sim q_t(z_t | x_{1:t}, z_{1:t-1}^i)$
    $z_{1:t}^i = \textbf{CONCAT}(z_{1:t-1}^i, z_t^i)$
  $\hat{p}_t = \left(\sum_{i=1}^N w_{t-1}^i \alpha_t(z_{1:t}^i)\right)$
  $\hat{p}_N(x_{1:t}) = \hat{p}_N(x_{1:t-1})\hat{p}_t$
  $\{w_t^i\}_{i=1}^N = \{w_{t-1}^i \alpha_t(z_{1:t}^i)/\hat{p}_t\}_{i=1}^N$
  if resampling criteria satisfied by $\{w_t^i\}_{i=1}^N$ **then**
    $\{w_t^i, z_{1:t}^i\}_{i=1}^N = \textbf{RSAMP}(\{w_t^i, z_{1:t}^i\}_{i=1}^N)$
return $\log \hat{p}_N(x_{1:T})$

RSAMP$(\{w^i, z^i\}_{i=1}^N)$:
for $i \in \{1, \ldots, N\}$ **do**
  $a \sim \text{Categorical}(\{w^i\}_{i=1}^N)$
  $y^i = z^a$
return $\{\frac{1}{N}, y^i\}_{i=1}^N$

4 Filtering Variational Objectives (FIVOs)

The filtering variational objectives (FIVOs) are a family of MCOs defined by the marginal likelihood estimator of a particle filter. For models with sequential structure, e.g., latent variable models of audio and text, the relative variance of a naive importance sampling estimator tends to scale exponentially in the number of steps. In contrast, the relative variance of particle filter estimators can scale more favorably with the number of steps—linearly in some cases [53, 25]. Thus, the results of Section 3 suggest that FIVOs can serve as tighter objectives than IWAE for MLE in sequential models.

Let our observations be sequences of T $\mathcal{X}$-valued random variables denoted $x_{1:T}$, where $x_{i:j} \equiv (x_i, \ldots, x_j)$. We also assume that the data generation process relies on a sequence of T unobserved $\mathcal{Z}$-valued latent variables denoted $z_{1:T}$. We focus on sequential latent variable models that factor as a series of tractable conditionals, $p(x_{1:T}, z_{1:T}) = p_1(x_1, z_1) \prod_{t=2}^T p_t(x_t, z_t | x_{1:t-1}, z_{1:t-1})$.

A particle filter is a sequential Monte Carlo algorithm, which propagates a population of N weighted particles for T steps using a combination of importance sampling and resampling steps, see Alg. 1. In detail, the particle filter takes as arguments an observation $x_{1:T}$, the number of particles N, the model distribution p, and a variational posterior $q(z_{1:T}|x_{1:T})$ factored over t,

$$q(z_{1:T}|x_{1:T}) = \prod_{t=1}^T q_t(z_t | x_{1:t}, z_{1:t-1}) \,. \tag{6}$$

The particle filter maintains a population $\{w_{t-1}^i, z_{1:t-1}^i\}_{i=1}^N$ of particles $z_{1:t-1}^i$ with weights w_{t-1}^i. At step t, the filter independently proposes an extension $z_t^i \sim q_t(z_t | x_{1:t}, z_{1:t-1}^i)$ to each particle's trajectory $z_{1:t-1}^i$. The weights w_{t-1}^i are multiplied by the incremental importance weights,

$$\alpha_t(z_{1:t}^i) = \frac{p_t(x_t, z_t^i | x_{1:t-1}, z_{1:t-1}^i)}{q_t(z_t^i | x_{1:t}, z_{1:t-1}^i)}, \tag{7}$$

and renormalized. If the current weights w_t^i satisfy a resampling criteria, then a resampling step is performed and N particles $z_{1:t}^i$ are sampled in proportion to their weights from the current population with replacement. Common resampling schemes include resampling at every step and resampling if the effective sample size (ESS) of the population $(\sum_{i=1}^N (w_t^i)^2)^{-1}$ drops below $N/2$ [68]. After resampling the weights are reset to 1. Otherwise, the particles $z_{1:t}^i$ are copied to the next step along with the accumulated weights. See Fig. 1 for a visualization.

Instead of viewing Alg. 1 as an inference algorithm, we treat the quantity $\mathbb{E}[\log \hat{p}_N(x_{1:T})]$ as an objective function over p. Because $\hat{p}_N(x_{1:T})$ is an unbiased estimator of $p(x_{1:T})$, proven in the Appendix and in [63, 64, 8, 204], it defines an MCO, which we call FIVO:

Definition 2. Filtering Variational Objectives. Let $\log \hat{p}_N(x_{1:T})$ be the output of Alg. 1 with inputs $(x_{1:T}, p, q, N)$, then $\mathcal{L}^{\text{FIVO}}_{\text{N}}(x_{1:T}, p, q) = \mathbb{E}[\log \hat{p}_N(x_{1:T})]$ is a filtering variational objective.

$\hat{p}_N(x_{1:T})$ is a strongly consistent estimator [63, 64]. So if $\log \hat{p}_N(x_{1:T})$ is uniformly integrable, then $\mathcal{L}^{\text{FIVO}}_{\text{N}}(x_{1:T}, p, q) \to \log p(x_{1:T})$ as $N \to \infty$. Resampling is the distinguishing feature of $\mathcal{L}^{\text{FIVO}}_{\text{N}}$; if resampling is removed, then FIVO reduces to IWAE. Resampling does add an amount of immediate variance, but it allows the filter to discard low weight particles with high probability. This has the

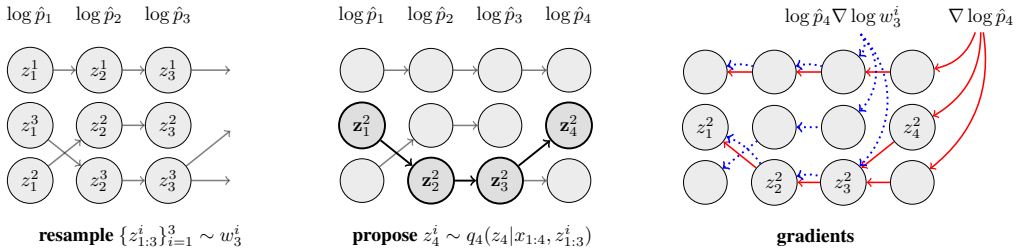


Figure 1: Visualizing FIVO; (Left) Resample from particle trajectories to determine inheritance in next step, (middle) propose with q_t and accumulate loss $\log \hat{p}_t$, (right) gradients (in the reparameterized case) flow through the lattice, objective gradients in solid red and resampling gradients in dotted blue.

effect of refocusing the distribution of particles to regions of higher mass under the posterior, and in some sequential models can reduce the variance from exponential to linear in the number of time steps [53, 25]. Resampling is a greedy process, and it is possible that a particle discarded at step t, could have attained a high mass at step T. In practice, the best trade-off is to use adaptive resampling schemes [68]. If for a given $x_{1:T}, p, q$ a particle filter's likelihood estimator improves over simple importance sampling in terms of variance, we expect $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}$ to be a tighter bound than $\mathcal{L}$ or $\mathcal{L}_{\mathrm{N}}^{\mathrm{IWAE}}$.

4.1 Optimization

The FIVO bound can be optimized with the same stochastic gradient ascent framework used for the ELBO. We found in practice it was effective simply to follow a Monte Carlo estimator of the biased gradient $\mathbb{E}[\nabla_{(\theta,\phi)} \log \hat{p}_N(x_{1:T})]$ with reparameterized z_t^i. This gradient estimator is biased, as the full FIVO gradient has three kinds of terms: it has the term $\mathbb{E}[\nabla_{\theta,\phi} \log \hat{p}_N(x_{1:T})]$, where $\nabla_{\theta,\phi} \log \hat{p}_N(x_{1:T})$ is defined conditional on the random variables of Alg. 1; it has gradient terms for every distribution of Alg. 1 that depends on the parameters; and, if adaptive resampling is used, then it has additional terms that account for the change in FIVO with respect to the decision to resample. In this section, we derive the FIVO gradient when z_t^i are reparameterized and a fixed resampling schedule is followed. We derive the full gradient in the Appendix.

In more detail, we assume that p and q are parameterized in a differentiable way by θ and ϕ. Assume that q is from a reparameterizable family and that z_t^i of Alg. 1 are reparameterized. Assume that we use a fixed resampling schedule, and let $\mathbb{I}(\text{resampling at step } t)$ be an indicator function indicating whether a resampling occured at step t. Now, $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}$ depends on the parameters via $\log \hat{p}_N(x_{1:T})$ and the resampling probabilities w_t^i in the density. Thus, $\nabla_{(\theta,\phi)} \mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}} =$

$$\mathbb{E}\left[\nabla_{(\theta,\phi)} \log \hat{p}_N(x_{1:T}) + \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{I}(\text{resampling at step } t) \log \frac{\hat{p}_N(x_{1:T})}{\hat{p}_N(x_{1:t})} \nabla_{(\theta,\phi)} \log w_t^i\right] \quad (8)$$

Given a single forward pass of Alg. 1 with reparameterized z_t^i, the terms inside the expectation form a Monte Carlo estimator of Eq. (8). However, the terms from resampling events contribute to the majority of the variance of the estimator. Thus, the gradient estimator that we found most effective in practice consists only of the gradient $\nabla_{(\theta,\phi)} \log \hat{p}_N(x_{1:T})$, the solid red arrows of Figure 1. We explore this experimentally in Section 6.3.

4.2 Sharpness

As with the ELBO, FIVO is a variational objective taking a variational posterior q as an argument. An important question is whether FIVO achieves the marginal log-likelihood at its optimal q. We can only guarantee this for models in which $z_{1:t-1}$ and x_t are independent given $x_{1:t-1}$.

Proposition 2. Sharpness of Filtering Variational Objectives. *Let* $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q)$ *be a FIVO, and* $q^*(x_{1:T}, p) = \mathrm{argmax}_q \mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q)$. *If* p *has independence structure such that* $p(z_{1:t-1}|x_{1:t}) = p(z_{1:t-1}|x_{1:t-1})$ *for* $t \in \{2, \ldots, T\}$, *then*

$$q^*(x_{1:T}, p)(z_{1:T}) = p(z_{1:T}|x_{1:T}) \quad \textit{and} \quad \mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q^*(x_{1:T}, p)) = \log p(x_{1:T}) \,.$$

Proof. See Appendix. □

Most models do not satisfy this assumption, and deriving the optimal q in general is complicated by the resampling dynamics. For the restricted the model class in Proposition 2, the optimal q_t does not condition on future observations $x_{t+1:T}$. We explored this experimentally with richer models in Section 6.4, and found that allowing q_t to condition on $x_{t+1:T}$ does not reliably improve FIVO. This is consistent with the view of resampling as a greedy process that responds to each intermediate distribution as if it were the final. Still, we found that the impact of this effect was outweighed by the advantage of optimizing a tighter bound.

5 Related Work

The marginal log-likelihood is a central quantity in statistics and probability, and there has long been an interest in bounding it [260]. The literature relating to the bounds we call Monte Carlo objectives has typically focused on the problem of estimating the marginal likelihood itself. [107, 49] use Jensen's inequality in a forward and reverse estimator to detect the failure of inference methods. IWAE [50] is a clear influence on this work, and FIVO can be seen as an extension of this bound. The ELBO enjoys a long history [128] and there have been efforts to improve the ELBO itself. [214] generalize the ELBO by considering arbitrary operators of the model and variational posterior. More closely related to this work is a body of work improving the ELBO by increasing the expressiveness of the variational posterior. For example, [217, 136] augment the variational posterior with deterministic transformations with fixed Jacobians, and [229] extend the variational posterior to admit a Markov chain.

Other approaches to learning in neural latent variable models include [35], who use importance sampling to approximate gradients under the posterior, and [108], who use sequential Monte Carlo to approximate gradients under the posterior. These are distinct from our contribution in the sense that for them inference for the sake of estimation is the ultimate goal. To our knowledge the idea of treating the output of inference as an objective in and of itself, while not completely novel, has not been fully appreciated in the literature. Although, this idea shares inspiration with methods that optimize the convergence of Markov chains [23].

We note that the idea to optimize the log estimator of a particle filter was independently and concurrently considered in [180, 146]. In [180] the bound we call FIVO is cast as a tractable lower bound on the ELBO defined by the particle filter's non-parameteric approximation to the posterior. [146] additionally derive an expression for FIVO's bias as the KL between the filter's distribution and a certain target process. Our work is distinguished by our study of the convergence of MCOs in N, which includes FIVO, our investigation of FIVO sharpness, and our experimental results on stochastic RNNs.

6 Experiments

In our experiments, we sought to: (a) compare models trained with ELBO, IWAE, and FIVO bounds in terms of final test log-likelihoods, (b) explore the effect of the resampling gradient terms on FIVO, (c) investigate how the lack of sharpness affects FIVO, and (d) consider how models trained with FIVO use the stochastic state. To explore these questions, we trained variational recurrent neural networks (VRNN) [58] with the ELBO, IWAE, and FIVO bounds using TensorFlow [1] on two benchmark sequential modeling tasks: natural speech waveforms and polyphonic music. These datasets are known to be difficult to model without stochastic latent states [82].

The VRNN is a sequential latent variable model that combines a deterministic recurrent neural network (RNN) with stochastic latent states z_t at each step. The observation distribution over x_t is conditioned directly on z_t and indirectly on $z_{1:t-1}$ via the RNN's state $h_t(z_{t-1}, x_{t-1}, h_{t-1})$. For a length T sequence, the model's posterior factors into the conditionals $\prod_{t=1}^{T} p_t(z_t|h_t(z_{t-1}, x_{t-1}, h_{t-1}))g_t(x_t|z_t, h_t(z_{t-1}, x_{t-1}, h_{t-1}))$, and the variational posterior factors as $\prod_{t=1}^{T} q_t(z_t|h_t(z_{t-1}, x_{t-1}, h_{t-1}), x_t)$. All distributions over latent variables are factorized Gaussians, and the output distributions g_t depend on the dataset. The RNN is a single-layer LSTM and the conditionals are parameterized by fully connected neural networks with one hidden layer of the same size as the LSTM hidden layer. We used the residual parameterization [82] for the variational posterior.

N	Bound	Nottingham	JSB	MuseData	Piano-midi.de
4	ELBO	-3.00	-8.60	-7.15	-7.81
	IWAE	-2.75	-7.86	-7.20	-7.86
	FIVO	**-2.68**	**-6.90**	**-6.20**	**-7.76**
8	ELBO	-3.01	-8.61	-7.19	-7.83
	IWAE	-2.90	-7.40	-7.15	-7.84
	FIVO	**-2.77**	**-6.79**	**-6.12**	**-7.45**
16	ELBO	-3.02	-8.63	-7.18	-7.85
	IWAE	-2.85	-7.41	-7.13	-7.79
	FIVO	**-2.58**	**-6.72**	**-5.89**	**-7.43**

		TIMIT	
N	Bound	64 units	256 units
4	ELBO	0	10,438
	IWAE	-160	11,054
	FIVO	**5,691**	**17,822**
8	ELBO	2,771	9,819
	IWAE	3,977	11,623
	FIVO	**6,023**	**21,449**
16	ELBO	1,676	9,918
	IWAE	3,236	13,069
	FIVO	**8,630**	**21,536**

Table 1: Test set marginal log-likelihood bounds for models trained with ELBO, IWAE, and FIVO. For ELBO and IWAE models, we report $\max\{\mathcal{L}, \mathcal{L}_{128}^{\text{IWAE}}, \mathcal{L}_{128}^{\text{FIVO}}\}$. For FIVO models, we report $\mathcal{L}_{128}^{\text{FIVO}}$. Pianoroll results are in nats per timestep, TIMIT results are in nats per sequence relative to ELBO with $N = 4$. For details on our evaluation methodology and absolute numbers see the Appendix.

For FIVO we resampled when the ESS of the particles dropped below $N/2$. For FIVO and IWAE we used a batch size of 4, and for the ELBO, we used batch sizes of $4N$ to match computational budgets (resampling is $\mathcal{O}(N)$ with the alias method). For all models we report bounds using the variational posterior trained jointly with the model. For models trained with FIVO we report $\mathcal{L}_{128}^{\text{FIVO}}$. To provide strong baselines, we report the maximum across bounds, $\max\{\mathcal{L}, \mathcal{L}_{128}^{\text{IWAE}}, \mathcal{L}_{128}^{\text{FIVO}}\}$, for models trained with ELBO and IWAE. Additional details in the Appendix.

6.1 Polyphonic Music

We evaluated VRNNs trained with the ELBO, IWAE, and FIVO bounds on 4 polyphonic music datasets: the Nottingham folk tunes, the JSB chorales, the MuseData library of classical piano and orchestral music, and the Piano-midi.de MIDI archive [40]. Each dataset is split into standard train, valid, and test sets and is represented as a sequence of 88-dimensional binary vectors denoting the notes active at the current timestep. We mean-centered the input data and modeled the output as a set of 88 factorized Bernoulli variables. We used 64 units for the RNN hidden state and latent state size for all polyphonic music models except for JSB chorales models, which used 32 units. We report bounds on average log-likelihood per timestep in Table 1. Models trained with the FIVO bound significantly outperformed models trained with either the ELBO or the IWAE bounds on all four datasets. In some cases, the improvements exceeded 1 nat *per timestep*, and in all cases optimizing FIVO with $N = 4$ outperformed optimizing IWAE or ELBO for $N = \{4, 8, 16\}$.

6.2 Speech

The TIMIT dataset is a standard benchmark for sequential models that contains 6300 utterances with an average duration of 3.1 seconds spoken by 630 different speakers. The 6300 utterances are divided into a training set of size 4620 and a test set of size 1680. We further divided the training set into a validation set of size 231 and a training set of size 4389, with the splits exactly as in [82]. Each TIMIT utterance is represented as a sequence of real-valued amplitudes which we split into a sequence of 200-dimensional frames, as in [58, 82]. Data preprocessing was limited to mean centering and variance normalization as in [82]. For TIMIT, the output distribution was a factorized Gaussian, and we report the average log-likelihood bound per sequence relative to models trained with ELBO. Again, models trained with FIVO significantly outperformed models trained with IWAE or ELBO, see Table 1.

6.3 Resampling Gradients

All models in this work (except those in this section) were trained with gradients that did not include the term in Eq. (8) that comes from resampling steps. We omitted this term because it has an outsized effect on gradient variance, often increasing it by 6 orders of magnitude. To explore the effects of this term experimentally, we trained VRNNs with and without the resampling gradient term on the TIMIT and polyphonic music datasets. When using the resampling term, we attempted to control its variance

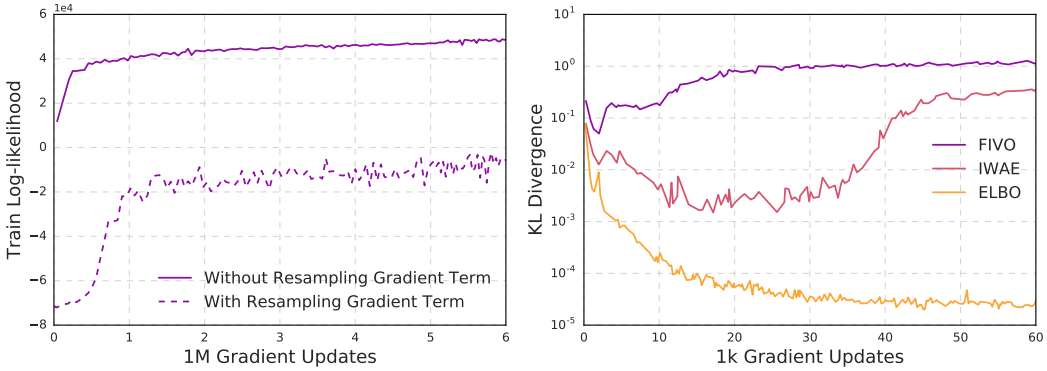


Figure 2: (Left) Graph of $\mathcal{L}_{128}^{\text{FIVO}}$ over training comparing models trained with and without the resampling gradient terms on TIMIT with $N = 4$. (Right) KL divergence from $q(z_{1:T}|x_{1:T})$ to $p(z_{1:T})$ for models trained on the JSB chorales with $N = 16$.

Bound	Nottingham	JSB	MuseData	Piano-midi.de	TIMIT
ELBO	**-2.40**	**-5.48**	-6.54	**-6.68**	**0**
ELBO+s	-2.59	-5.53	**-6.48**	-6.77	-925
IWAE	-2.52	-5.77	-6.54	**-6.74**	1,469
IWAE+s	**-2.37**	**-4.63**	**-6.47**	**-6.74**	**2,630**
FIVO	**-2.29**	-4.08	**-5.80**	-6.41	6,991
FIVO+s	-2.34	**-3.83**	-5.87	**-6.34**	**9,773**

Table 2: Train set marginal log-likelihood bounds for models comparing smoothing (+s) and non-smoothing variational posteriors. We report $\max\{\mathcal{L}, \mathcal{L}_{128}^{\text{IWAE}}, \mathcal{L}_{128}^{\text{FIVO}}\}$ for ELBO and IWAE models and $\mathcal{L}_{128}^{\text{FIVO}}$ for FIVO models. All models were trained with $N = 4$. Pianoroll results are in nats per timestep, TIMIT results are in nats per sequence relative to non-smoothing ELBO. For details on our evaluation methodology and absolute numbers see the Appendix.

using a moving-average baseline linear in the number of timesteps. For all datasets, models trained without the resampling gradient term outperformed models trained with the term by a large margin on both the training set and held-out data. Many runs with resampling gradients failed to improve beyond random initialization. A representative pair of train log-likelihood curves is shown in Figure 2 — gradients without the resampling term led to earlier convergence and a better solution. We stress that this is an empirical result — in principle biased gradients can lead to divergent behaviour. We leave exploring strategies to reduce the variance of the unbiased estimator to future work.

6.4 Sharpness

FIVO does not achieve the marginal log-likelihood at its optimal variational posterior q^*, because the optimal q^* does not condition on future observations (see Section 4.2). In contrast, ELBO and IWAE are sharp, and their q^*s depend on future observations. To investigate the effects of this, we defined a smoothing variant of the VRNN in which q takes as additional input the hidden state of a deterministic RNN run backwards over the observations, allowing q to condition on future observations. We trained smoothing VRNNs using ELBO, IWAE, and FIVO, and report evaluation on the training set (to isolate the effect on optimization performance) in Table 2 . Smoothing helped models trained with IWAE, but not enough to outperform models trained with FIVO. As expected, smoothing did not reliably improve models trained with FIVO. Test set performance was similar, see the Appendix for details.

6.5 Use of Stochastic State

A known pathology when training stochastic latent variable models with the ELBO is that stochastic states can go unused. Empirically, this is associated with the collapse of variational posterior $q(z|x)$ network to the model prior $p(z)$ [41]. To investigate this, we plot the KL divergence from $q(z_{1:T}|x_{1:T})$ to $p(z_{1:T})$ averaged over the dataset (Figure 2). Indeed, the KL of models trained with

ELBO collapsed during training, whereas the KL of models trained with FIVO remained high, even while achieving a higher log-likelihood bound.

7 Conclusions

We introduced the family of filtering variational objectives, a class of lower bounds on the log marginal likelihood that extend the evidence lower bound. FIVOs are suited for MLE in neural latent variable models. We trained models with the ELBO, IWAE, and FIVO bounds and found that the models trained with FIVO significantly outperformed other models across four polyphonic music modeling tasks and a speech waveform modeling task. Future work will include exploring control variates for the resampling gradients, FIVOs defined by more sophisticated filtering algorithms, and new MCOs based on differentiable operators like leapfrog operators with deterministically annealed temperatures. In general, we hope that this paper inspires the machine learning community to take a fresh look at the literature of marginal likelihood estimators—seeing them as objectives instead of algorithms for inference.

Acknowledgments

We thank Matt Hoffman, Matt Johnson, Danilo J. Rezende, Jascha Sohl-Dickstein, and Theophane Weber for helpful discussions and support in this project. A. Doucet was partially supported by the EPSRC grant EP/K000276/1. Y. W. Teh's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

Appendix to Filtering Variational Objectives

Proof of Proposition 1.

Let $\mathbb{E}[\hat{p}_N(x)] = p(x)$ and define $\mathcal{L}_N(x, p) = \mathbb{E}[\log \hat{p}_N(x)]$ as the Monte Carlo objective defined by $\hat{p}_N(x)$.

(a) By the concavity of $\log$ and Jensen's inequality,

$$\mathcal{L}_N(x, p) = \mathbb{E}[\log \hat{p}_N(x)] \leq \log \mathbb{E}[\hat{p}_N(x)] = \log p(x)$$

(b) Assume

- $\hat{p}_N(x)$ is strongly consistent, i.e. $\hat{p}_N(x) \xrightarrow{a.s.} p(x)$ as $N \to \infty$.
- $\log \hat{p}_N(x)$ is uniformly integrable. That is, let $(\Omega, \mathcal{F}, \mu)$ be the probability space on which $\log \hat{p}_N(x)$ is defined. The random variables $\{\log \hat{p}_N(x)\}_{N=1}^{\infty}$ are uniformly integrable if $\mathbb{E}[|\log \hat{p}_N(x)|] < \infty$ and if for any $\epsilon > 0$, there exists $\delta > 0$, such that for all N and $E \in \mathcal{F}$, $\mu(E) < \delta$ implies $\mathbb{E}[|\log \hat{p}_N(x)|\mathbb{I}(E)] < \epsilon$, where $\mathbb{I}(E)$ is an indicator function of the set E.

Then by continuity of $\log$, $\log \hat{p}_N(x)$ converges almost surely to $\log p(x)$. By Vitali's convergence theorem (using the uniform integrability assumption), we get $\mathcal{L}_N(x, p) = \mathbb{E}[\log \hat{p}_N(x)] \to \log p(x)$ as $N \to \infty$.

(c) Let $g(N) = \mathbb{E}[(\hat{p}_N(x) - p(x))^6]$, and assume $\limsup_{N\to\infty} \mathbb{E}[(\hat{p}_N(x))^{-1}] < \infty$. Define the relative error

$$\Delta = \frac{\hat{p}_N(x) - p(x)}{p(x)} \tag{9}$$

Then the bias $\log p(x) - \mathcal{L}_N(x, p) = -\mathbb{E}[\log(1 + \Delta)]$. Now, Taylor expand $\log(1 + \Delta)$ about 0,

$$\log(1 + \Delta) = \Delta - \frac{1}{2}\Delta^2 + \int_0^{\Delta} \left(\frac{1}{1+x} - 1 + x\right) dx \tag{10}$$

$$= \Delta - \frac{1}{2}\Delta^2 + \int_0^{\Delta} \left(\frac{x^2}{1+x}\right) dx \tag{11}$$

and in expectation

$$-\mathbb{E}[\log(1+\Delta)] = \frac{1}{2}\Delta^2 - \mathbb{E}\left[\int_0^\Delta \left(\frac{x^2}{1+x}\right)\, dx\right] \tag{12}$$

Our aim is to show

$$\left|\mathbb{E}\left[\int_0^\Delta \frac{x^2}{1+x}\, dx\right]\right| \in \mathcal{O}(g(N)^{1/2}) \tag{13}$$

In particular, by Cauchy-Schwarz

$$\left|\mathbb{E}\left[\int_0^\Delta \left(\frac{x^2}{1+x}\right)\, dx\right]\right| \leq \mathbb{E}\left[\left|\int_0^\Delta \frac{1}{(1+x)^2}\, dx\right|^{1/2} \left|\int_0^\Delta x^4\, dx\right|^{1/2}\right] \tag{14}$$

$$= \mathbb{E}\left[\left|\frac{\Delta}{1+\Delta}\right|^{1/2} \left|\frac{\Delta^5}{5}\right|^{1/2}\right] \tag{15}$$

$$= \mathbb{E}\left[\left|\frac{1}{1+\Delta}\right|^{1/2} \left|\frac{\Delta^6}{5}\right|^{1/2}\right] \tag{16}$$

and again by Cauchy-Schwarz

$$\leq \left(\mathbb{E}\left[\left|\frac{1}{1+\Delta}\right|\right]\right)^{1/2} \left(\mathbb{E}\left[\frac{\Delta^6}{5}\right]\right)^{1/2}. \tag{17}$$

This concludes the proof.

Controlling the first inverse moment.

We provide a sufficient condition that guarantees that the inverse moment of the average of i.i.d. random variables is bounded, a condition used in Proposition 1 (c). Intuitively, this is a fairly weak condition, because it only requires that the mass in an arbitrarily small neighbourhood of zero is bounded.

Lemma 3. *Let w_i be i.i.d. positive random variables and $\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^N w_i$. If there exist $M, C, \epsilon > 0$ such that $\mathbb{P}(w_i < w) \leq Cw^{1+\epsilon}$ for $w \in [0, M)$, then $\mathbb{E}[\hat{p}_N(x)^{-1}] \leq C\frac{M^\epsilon}{\epsilon} + \frac{1}{M}$.*

Proof. Let $M, C, \epsilon > 0$ be such that $\mathbb{P}(w_i < w) \leq Cw^{1+\epsilon}$ for $w \in [0, M)$. We proceed in two cases. If $N = 1$, then

$$\begin{aligned}
\mathbb{E}[\hat{p}_N(x)^{-1}] &= \int_0^\infty \mathbb{P}(w_1^{-1} > u)\, du \\
&= \int_0^\infty \mathbb{P}(w_1 < 1/u)\, du \\
&= \int_0^M \frac{\mathbb{P}(w_1 < w)}{w^2}\, dw + \int_M^\infty \frac{\mathbb{P}(w_1 < w)}{w^2}\, dw \\
&\leq \int_0^M \frac{Cw^{1+\epsilon}}{w^2}\, dw + \int_M^\infty \frac{1}{w^2}\, dw \\
&= C\frac{M^\epsilon}{\epsilon} + \frac{1}{M}
\end{aligned}$$

For $N > 1$, we show that $\mathbb{E}[\hat{p}_N(x)^{-1}] \leq \mathbb{E}[\hat{p}_1(x)^{-1}]$, so the same condition is sufficient for any N. The AM-GM inequality tells us that

$$\sum_{i=1}^N \frac{w_i}{N} \geq \left(\prod_{i=1}^N w_i\right)^{1/N}$$

so

$$\begin{aligned}\mathbb{E}[\hat{p}_N(x)^{-1}] &\leq \mathbb{E}\left[\left(\prod_{i=1}^{N} w_i\right)^{-1/N}\right] \\ &= \prod_{i=1}^{N} \mathbb{E}\left[w_i^{-1/N}\right] \\ &= \mathbb{E}\left[w_1^{-1/N}\right]^N\end{aligned}$$

and by Lyapunov's inequality, we have

$$\leq \mathbb{E}\left[\left(w_1^{-1/N}\right)^N\right] = \mathbb{E}[\hat{p}_1(x)^{-1}]$$

This concludes the proof. □

Gradients of $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q)$.

We formulate unbiased gradients of $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q)$ by considering Algorithm 1 as a method for simulating FIVO. We consider the cases when the sampling of z_t^i is and is not reparameterized. We also consider the case where we make adaptive resampling decisions.

First, we assume that the decision to resample is not adaptive (i.e., depends in some way on the random variables already produced until that point in Algorithm 1), and are fixed ahead of time. When the sampling z_t^i is not reparameterized there are three terms to the gradient: (1) the gradients of $\log \hat{p}_N(x_{1:T})$ with respect to the parameters conditional on the latent states, (2) gradients of the densities q_t with respect to their parameters, and (3) gradients of the resampling probabilities with respect to the parameters. All together, the following is a gradient of FIVO,

$$\begin{aligned}\mathbb{E}\Big[\nabla_{\theta,\phi} \log \hat{p}_N(x_{1:T}) + \sum_{t=1}^{T}\sum_{i=1}^{N}\Big(\log \frac{\hat{p}_N(x_{1:T})}{\hat{p}_N(x_{1:t-1})} \nabla_\phi \log q_{t,\phi}(z_t^i | x_{1:t}, z_{1:t-1}^i) + \\ \mathbb{I}(\text{resampling at step } t) \log \frac{\hat{p}_N(x_{1:T})}{\hat{p}_N(x_{1:t})} \nabla_{\theta,\phi} \log w_t^i\Big)\Big]\end{aligned} \tag{18}$$

where $\mathbb{I}(A)$ is an indicator function. If z_t^i is reparameterized, then the first and third terms suffice for an unbiased gradient,

$$\mathbb{E}\left[\nabla_{\theta,\phi} \log \hat{p}_N(x_{1:T}) + \sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{I}(\text{resampling at step } t) \log \frac{\hat{p}_N(x_{1:T})}{\hat{p}_N(x_{1:t})} \nabla_{\theta,\phi} \log w_t^i\right] \tag{19}$$

In this work we only considered reparameterized q_ts, and we dropped the terms of the gradient that arise from resampling.

Second, when the decision to resample is adaptive, the domain of the random variables involved in simulating $\log \hat{p}_N(x_{1:T})$ can be partitioned into 2^T regions, over each of which the density is differentiable. Between those regions, the density experiences a jump discontinuity. Thus, there are additional terms to the gradient of $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q)$ that correspond to the change in the regions of continuity as the parameters change. These terms can be written as surface integrals over the boundaries of the regions. We drop these terms in practice.

Proof of Proposition 2.

Assume $p(z_{1:t-1}|x_{1:t}) = p(z_{1:t-1}|x_{1:t-1})$ for all $t \in \{2, \ldots, T\}$. We will show $\mathcal{L}_{\mathrm{N}}^{\mathrm{FIVO}}(x_{1:T}, p, q) = \log p(x_{1:T})$ at $q(z_t|z_{1:t-1}, x_{1:t}) = p(z_t|z_{1:t-1}, x_{1:t})$. We will do this by induction, showing that every particle has a constant weight and that $\hat{p}_N(x_{1:T}) = p(x_{1:T})$ is a constant. For $t = 1$ we have

$$\alpha_1^i(z_1) = \frac{p_1(x_1, z_1)}{p(z_1|x_1)} = p_1(x_1) \tag{20}$$

Thus, all particles have the same weight and $\hat{p}_1 = p_1(x_1)$. Now for any t we have that the weights must be $1/N$ since the particles all have the same weight and

$$\alpha_t^i(z_{1:t}) = \frac{p_t(x_t, z_t|z_{1:t-1}, x_{1:t-1})}{p(z_t|z_{1:t-1}, x_{1:t})} \tag{21}$$

$$= \frac{p(z_{1:t}, x_{1:t})}{p(z_{1:t-1}, x_{1:t-1})p(z_t|z_{1:t-1}, x_{1:t})} \tag{22}$$

$$= \frac{p(x_{1:t})}{p(x_{1:t-1})} \frac{p(z_{1:t}|x_{1:t})}{p(z_{1:t-1}|x_{1:t-1})p(z_t|z_{1:t-1}, x_{1:t})} \tag{23}$$

$$= \frac{p(x_{1:t})}{p(x_{1:t-1})} \frac{p(z_{1:t}|x_{1:t})}{p(z_{1:t-1}|x_{1:t})p(z_t|z_{1:t-1}, x_{1:t})} \tag{24}$$

$$= \frac{p(x_{1:t})}{p(x_{1:t-1})} \tag{25}$$

and thus,

$$\hat{p}_N(x_{1:T}) = p_1(x_1) \prod_{t=2}^{T} \frac{p(x_{1:t})}{p(x_{1:t-1})} = p(x_{1:T}) \tag{26}$$

Implementation details

We initialized weights using the Xavier initialization [97] and used the Adam optimizer [135] with a batch size of 4. During training, we did not truncate sequences and performed full backpropagation through time for all datasets. For the results presented in Sections 6.1 and 6.2 we performed a grid search over learning rates $\{3 \times 10^{-4}, 1 \times 10^{-4}, 3 \times 10^{-5}, 1 \times 10^{-5}\}$ and picked the run and early stopping step by the validation performance.

Evaluation and Comparison of Bounds

Comparing models trained with different log-likelihood lower bounds is challenging because calculating the actual log-likelihood is intractable. Burda *et al.* [50] showed that the IWAE bound is at least as tight as the ELBO and monotonically increases with N. This suggests comparing models based on the IWAE bound evaluated with a large N. However, we found that IWAE and ELBO bounds tended to diverge for models trained with FIVO.

Although FIVO is not provably a tighter bound than the ELBO or IWAE, our experiments suggest that this tends to be the case in practice. In Figure 3, we plotted all three bounds over training for a representative experiment. All plots use the same model architecture, but the training objective changes in each panel. For the model trained with IWAE, the FIVO and IWAE bounds are tighter than their counterparts on the model trained with ELBO, suggesting that the model trained with IWAE is superior. The ELBO bound evaluated on the model trained with IWAE, however, is lower than its counterpart on the model trained with the ELBO. For the model trained with FIVO, both IWAE and ELBO bounds seem to diverge, but the FIVO bound outperforms the FIVO bounds on both of the other models. As in the figure, we generally found that the same model evaluated with FIVO, IWAE, and ELBO produced values descending in that order.

We suspect that q distributions trained under the FIVO bound are more entropic than those trained under ELBO or IWAE because of the resampling operation. During training under FIVO, q is able to propose state transitions that could poorly explain the observations because the bad states will be resampled away without harming the final bound value. Then, when a FIVO-trained q is evaluated with ELBO or IWAE it proposes poor states that are not resampled away, leading to a poor final bound value. Conversely, qs trained with ELBO and IWAE are not able to fully leverage the resampling operation when evaluated with the FIVO bound.

Because of this behavior, we chose to optimistically evaluate models trained with IWAE and ELBO by reporting the maximum across all the bounds. For models trained with FIVO, we reported only the FIVO bound. We felt this evaluation scheme provided the strongest comparison to existing bounds.

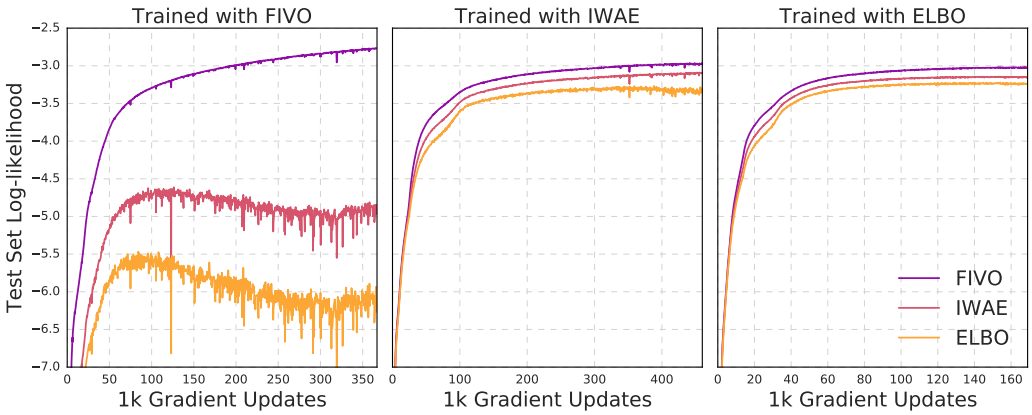


Figure 3: Comparison of ELBO, IWAE, and FIVO bounds. We plot the ELBO ($\mathcal{L}$), IWAE ($\mathcal{L}_{128}^{\text{IWAE}}$), and FIVO ($\mathcal{L}_{128}^{\text{FIVO}}$) test log-likelihood lower bounds for a fixed model architecture trained with FIVO (left), IWAE (middle), and ELBO (right). The models are VRNNs trained on the Nottingham dataset with 64 units, $N = 16$, and learning rate 3×10^{-5}.

Evaluating TIMIT Log-Likelihoods

We reported log-likelihood scores for TIMIT relative to an ELBO baseline instead of raw log-likelihoods. Previous papers (e.g., [58, 82]) report the log-likelihood of data that have been mean centered and variance normalized, but it would be more proper to report the results on the un-standardized data. Specifically, if the training set has mean μ and variance σ^2 and the model outputs $\hat{\mu}$ and $\hat{\sigma}^2$, then the un-standardized test data would be evaluated under a $\mathcal{N}(\hat{\mu}\sigma + \mu, \hat{\sigma}^2\sigma^2)$ distribution.

Log-likelihoods produced by these approaches differ by a constant offset that depends on σ. Because the offset is a function of only training set statistics, it does not affect relative comparison between methods. Because of this we chose to report log-likelihoods relative to a baseline instead of absolute numbers. Absolute numbers calculated on standardized data are reported in Tables 3, 4, and 5 to allow for comparisons with other papers.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

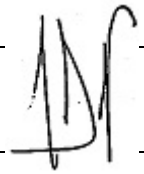
Title of Paper	Filtering Variational Objectives
Publication Status	☒ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Chris J. Maddison*, Dieterich Lawson*, George Tucker*, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, Yee Whye Teh. Filtering Variational Objectives. In *Advances in Neural Information Processing Systems*, 2017.

Student Confirmation

Student Name:	Chris J. Maddison		
Contribution to the Paper	• I proposed the idea and worked out the correctness of the approach. • The proofs were joint work with Arnaud Doucet. Arnaud helped me complete the proof of the rate of convergence and he provided the proof of conditions that control the first inverse moment. • I wrote a full set of experimental code for an RL application, but the experiments that I performed were not included in the final version. A separate codebase was used for the generative modeling experiments in the paper, and the reported experiments were all run by Dieterich Lawson. • I wrote the majority of the paper, with the exception of the experimental section, which was written by Dieterich Lawson and George Tucker. • All authors contributed to the development of the paper through discussions and ideas, and all authors reviewed the final draft.		
Signature		Date	05 May 2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Arnaud Doucet			
Supervisor comments I agree that the candidate has made a substantial contribution to the publication.			
Signature		Date	05 May 2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Hamiltonian Descent

4.1 Abstract

We propose a family of optimization methods that achieve linear convergence using first-order gradient information and constant step sizes on a class of convex functions much larger than the smooth and strongly convex ones. This larger class includes functions whose second derivatives may be singular or unbounded at their minima or near infinity. Our methods are discretizations of conformal Hamiltonian dynamics, which generalizes the classical momentum method to model the motion of a particle with non-standard kinetic energy exposed to a dissipative force and the gradient field of the function of interest. They are first-order in the sense that they require only gradient computation. Yet, crucially the kinetic gradient map can be designed to incorporate global information about the convex conjugate in a fashion that allows for linear convergence on convex functions that may be non-smooth or non-strongly convex. We study in detail one implicit and two explicit methods. For one explicit method, we provide conditions under which it converges to stationary points of non-convex functions. For all, we provide conditions on the convex function and kinetic energy pair that guarantee linear convergence, and show that these conditions can be satisfied by functions with power growth. In sum, these methods expand the class of convex functions on which linear convergence is possible with first-order computation.

4.2 Introduction

We consider the problem of unconstrained minimization of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\min_{x \in \mathbb{R}^d} f(x), \tag{4.1}$$

by iterative methods that require only the partial derivatives $\nabla f(x) = (\partial f(x)/\partial x^{(n)}) \in \mathbb{R}^d$ of f, known also as first-order methods [190, 207, 193]. These methods produce a sequence of iterates $x_i \in \mathbb{R}^d$, and our emphasis is on those that achieve linear convergence, i.e., as a function of the iteration i they satisfy $f(x_i) - f(x_{\min}) = \mathcal{O}(\lambda^{-i})$ for some rate $\lambda > 1$ and $x_{\min} \in \mathbb{R}^d$ a global minimizer. We briefly consider non-convex differentiable f, but the bulk of our analysis focuses on the case of convex differentiable f. Our results will also occasionally require twice differentiability of f.

The convergence rates of first-order methods on convex functions can be broadly separated by the properties of strong convexity and Lipschitz smoothness. Taken together these properties for convex f are equivalent to the conditions that the following left hand bound (strong convexity) and right hand bound (smoothness) hold for some $\mu, L \in (0, \infty)$ and all $x, y \in \mathbb{R}^d$,

$$\frac{\mu}{2} \|x - y\|_2^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2} \|x - y\|_2^2 , \tag{4.2}$$

where $\langle x, y \rangle = \sum_{n=1}^{d} x^{(n)} y^{(n)}$ is the standard inner product and $\|x\|_2 = \sqrt{\langle x, x \rangle}$ is the Euclidean norm. For twice differentiable f, these properties are equivalent to the conditions that eigenvalues of the matrix of second-order partial derivatives $\nabla^2 f(x) = (\partial^2 f(x)/\partial x^{(n)} \partial x^{(m)}) \in \mathbb{R}^{d \times d}$ are everywhere lower bounded by μ and upper bounded by L, respectively. Thus, functions whose second derivatives are continuously unbounded or approaching 0, cannot be both strongly convex and smooth. Both bounds play an important role in the performance of first-order methods. On the one hand, for smooth and strongly convex f, the iterates of many first-order methods converge linearly. On the other hand, for any first-order method, there exist smooth convex functions and non-smooth strongly convex functions on which its convergence is sub-linear, i.e., $f(x_i) - f(x_{\min}) \geq \mathcal{O}(i^{-2})$ for any first-order method on smooth convex functions. See [190, 207, 193] for these classical results and [131] for other more exotic scenarios. Moreover, for a given method it can sometimes be very easy to find examples on which its convergence is slow; see Figure 4.1, in which gradient descent with a fixed step size converges slowly on $f(x) = [x^{(1)} + x^{(2)}]^4 + [x^{(1)}/2 - x^{(2)}/2]^4$, which is not strongly convex as its Hessian is singular at $(0, 0)$.

The central assumption in the worst case analyses of first-order methods is that information about f is restricted to black box evaluations of f and ∇f locally at points $x \in \mathbb{R}^d$, see [190, 193]. In this paper we assume additional access to first-order information of a second differentiable function $k : \mathbb{R}^d \to \mathbb{R}$ and show how ∇k can be designed to incorporate information about f to yield practical methods that converge linearly on convex functions. These methods are derived by discretizing the conformal

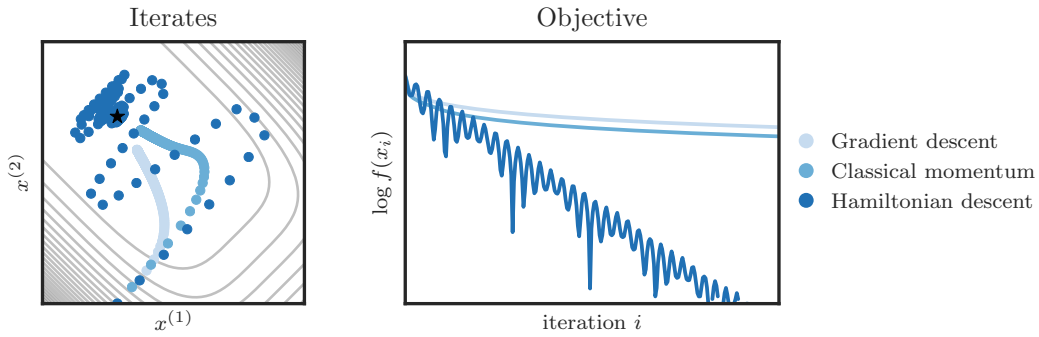


Figure 4.1: Optimizing $f(x) = [x^{(1)} + x^{(2)}]^4 + [x^{(1)}/2 - x^{(2)}/2]^4$ with three methods: gradient descent with fixed step size equal to $1/L_0$ where $L_0 = \lambda_{\max}(\nabla^2 f(x_0))$ is the maximum eigenvalue of the Hessian $\nabla^2 f$ at x_0; classical momentum, which is a particular case of our first explicit method with $k(p) = [(p^{(1)})^2 + (p^{(2)})^2]/2$ and fixed step size equal to $1/L_0$; and Hamiltonian descent, which is our first explicit method with $k(p) = (3/4)[(p^{(1)})^{4/3} + (p^{(2)})^{4/3}]$ and a fixed step size.

Hamiltonian system [165]. These systems are parameterized by $f, k : \mathbb{R}^d \to \mathbb{R}$ and $\gamma \in (0, \infty)$ with solutions $(x_t, p_t) \in \mathbb{R}^{2d}$,

$$\begin{aligned} x_t' &= \nabla k(p_t) \\ p_t' &= -\nabla f(x_t) - \gamma p_t. \end{aligned} \tag{4.3}$$

From a physical perspective, these systems model the dynamics of a single particle located at x_t with momentum p_t and kinetic energy $k(p_t)$ being exposed to a force field ∇f and a dissipative force. For this reason we refer to k as, the *kinetic energy*, and ∇k, the *kinetic map*. When the kinetic map ∇k is the identity, $\nabla k(p) = p$, these dynamics are the continuous time analog of Polyak's heavy ball method [206]. Let $f_c(x) = f(x + x_{\min}) - f(x_{\min})$ denote the centered version of f, which takes its minimum at 0, with minimum value 0. Our key observation in this regard is that when f is convex, and k is chosen as $k(p) = (f_c^*(p) + f_c^*(-p))/2$ (where $f_c^*(p) = \sup\{\langle x, p\rangle - f_c(x) : x \in \mathbb{R}^d\}$ is the convex conjugate of f_c), these dynamics have linear convergence with rate independent of f. In other words, this choice of k acts as a preconditioner, a generalization of using $k(p) = \langle p, A^{-1}p\rangle/2$ for $f(x) = \langle x, Ax\rangle/2$. Thus ∇k can exploit global information provided by the conjugate f_c^* to condition convergence for generic convex functions.

To preview the flavor of our results in detail, consider the special case of optimizing the power function $f(x) = |x|^b/b$ for $x \in \mathbb{R}$ and $b \in (1, \infty)$ initialized at $x_0 > 0$ using system (4.3) (or discretizations of it) with $k(p) = |p|^a/a$ for $p \in \mathbb{R}$ and $a \in (1, \infty)$.

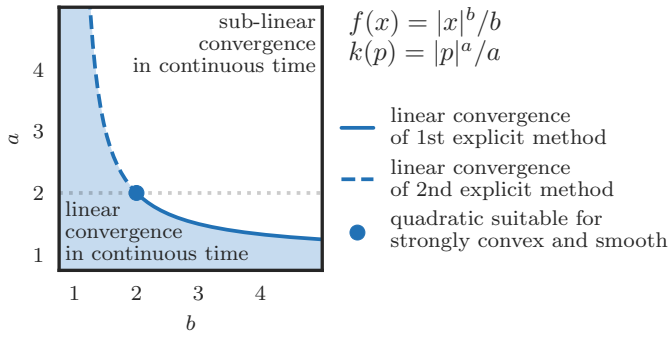


Figure 4.2: Convergence Regions for Power Functions. Shown are regions of distinct convergence types for Hamiltonian descent systems with $f(x) = |x|^b/b, k(p) = |p|^a/a$ for $x, p \in \mathbb{R}$ and $a, b \in (1, \infty)$. We show in Section 4.3 convergence is linear in continuous time iff $1/a + 1/b \geq 1$. In Section 4.5 we show that the assumptions of the explicit discretizations can be satisfied if $1/a + 1/b = 1$, leaving this as the only suitable pairing for linear convergence. Light dotted line is the line occupied by classical momentum with $k(p) = p^2/2$.

FOr this choice of f, it can be shown that $f_c^*(p) = f_c^*(-p) = k(p)$ when $a = b/(b-1)$. In line with this, in Section 4.3 we show that (4.3) exhibits linear convergence in continuous time if and only if $1/a + 1/b \geq 1$. In Section 4.4 we propose two explicit discretizations with fixed step sizes; in Section 4.5 we show that the first explicit discretization converges if $1/a + 1/b = 1$ and $b \geq 2$, and the second converges if $1/a + 1/b = 1$ and $1 < b \leq 2$. This means that the only suitable pairing corresponds in this case to the choice $k(p) \propto f_c^*(p) + f_c^*(-p)$. Figure 4.2 summarizes this discussion. Returning to Figure 4.1, we can compare the use of the kinetic energy of Polyak's heavy ball with a kinetic energy that relates appropriately to the convex conjugate of $f(x) = [x^{(1)} + x^{(2)}]^4 + [x^{(1)}/2 - x^{(2)}/2]^4$.

Most convex functions are not simple power functions, and computing $f_c^*(p) + f_c^*(-p)$ exactly is rarely feasible. To make our observations useful for numerical optimization, we show that linear convergence is still achievable in continuous time even if $k(p) \geq \alpha \max\{f_c^*(p), f_c^*(-p)\}$ for some $0 < \alpha \leq 1$ within a region defined by x_0. We study three discretizations of (4.3), one implicit method and two explicit ones (which are suitable for functions that grow asymptotically fast or slow, respectively). We prove linear convergence rates for these under appropriate additional assumptions. We introduce a family of kinetic energies that generalize the power functions to capture distinct power growth near zero and asymptotically far from zero. We show that the additional assumptions of discretization can be satisfied for this family of

k. We derive conditions on f that guarantee the linear convergence of our methods when paired with a specific choice of k from this family. These conditions generalize the quadratic growth implied by smoothness and strong convexity, extending it to general power growth that may be distinct near the minimum and asymptotically far from the minimum, which we refer to as tail and body behavior, respectively. Step sizes can be fixed independently of the initial position (and often dimension), and do not require adaptation, which often leads to convergence problems, see [267]. Indeed, we analyze a kinetic map ∇k that resembles the iterate updates of some popular adaptive gradient methods [75, 274, 90, 135], and show that it conditions the optimization of strongly convex functions with very fast growing tails (non-smooth). Thus, our methods provide a framework optimizing potentially non-smooth or non-strongly convex functions with linear rates using first-order computation.

The organization of the paper is as follows. In the rest of this section, we cover notation, review a few results from convex analysis, and give an overview of the related literature. In Section 4.3, we show the linear convergence of (4.3) under conditions on the relation between the kinetic energy k and f. We show a partial converse that in some settings our conditions are necessary. In Section 4.4, we present the three discretizations of the continuous dynamics and study the assumptions under which linear rates can be guaranteed for convex functions. For one of the discretizations, we also provide conditions under which it converges to stationary points of non-convex functions. In Section 4.5, we study a family of kinetic energies suitable for functions with power growth. We describe the class of functions for which the assumptions of the discretizations can be satisfied when using these kinetic energies.

4.2.1 Notation and Convex Analysis Review

We let $\langle x, y\rangle = \sum_{n=1}^{d} x^{(n)} y^{(n)}$ denote the standard inner product for $x, y \in \mathbb{R}^d$ and $\|x\|_2 = \sqrt{\langle x, x\rangle}$ the Euclidean norm. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, the gradient $\nabla f(x) = (\partial f(x)/\partial x^{(n)}) \in \mathbb{R}^d$ is the vector of partial derivatives at x. For twice-differentiable f, the Hessian $\nabla^2 h(x) = (\partial^2 f(x)/\partial x^{(n)} \partial x^{(m)}) \in \mathbb{R}^{d\times d}$ is the matrix of second-order partial derivatives at x. The notation x_t denotes the solution $x_t : [0, \infty) \to \mathbb{R}^d$ to a differential equation with derivative in t denoted x_t'. x_i denotes the iterates $x_i : \{0, 1, \ldots\} \to \mathbb{R}^d$ of a discrete system.

Consider a convex function $h : C \to \mathbb{R}$ that is defined on a convex domain $C \subseteq \mathbb{R}^d$ and differentiable on the interior $\operatorname{int}(C)$. The convex conjugate $h^* : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$h^*(p) = \sup\{\langle x, p\rangle - h(x) : x \in C\} \tag{4.4}$$

and it is itself convex. It is easy to show from the definition that if $g : C \to \mathbb{R}$ is another convex function such that $g(x) \leq h(x)$ for all $x \in C$, then $h^*(p) \leq g^*(p)$ for all $p \in \mathbb{R}^d$. Because we make such extensive use of it, we remind readers of the Fenchel-Young inequality: for $x \in C$ and $p \in \mathbb{R}^d$,

$$\langle x, p \rangle \leq h(x) + h^*(p), \tag{4.5}$$

which is easily derived from the definition of h^*, or see Section 12 of [222]. For $x \in \text{int}(C)$ by Theorem 26.4 of [222],

$$\langle x, \nabla h(x) \rangle = h(x) + h^*(\nabla h(x)). \tag{4.6}$$

Let $y \in \mathbb{R}^d$, $c \in \mathbb{R} \setminus \{0\}$. If $g(x) = h(x + y) - c$, then $g^*(p) = h^*(p) - \langle p, y \rangle + c$ (Theorem 12.3 [222]). If $h(x) = |x|^b/b$ for $x \in \mathbb{R}$ and $b \in (1, \infty)$, then $h^*(p) = |p|^a/a$ where $a = b/(b-1)$ (page 106 of [222]). If $g(x) = ch(x)$, then $g^*(p) = ch^*(p/c)$ (Table 3.2 [36]). For these and more on h^*, we refer readers to [222, 42, 36].

4.2.2 Related Literature

Standard references on convex optimization and the convergence analysis of first-order methods include [190, 207, 28, 42, 193, 47].

The heavy ball method was introduced by Polyak in his seminal paper [206]. In this paper, local convergence with linear rate was shown (i.e., when the initial position is sufficiently close to the local minimum). For quadratic functions, it can be shown that the convergence rate for optimally chosen step sizes is proportional to the square root of the conditional number of the Hessian, similarly to conjugate gradient descent (see e.g., [216]). As far as we know, global convergence of the heavy ball method for non-quadratic functions was only recently established in [91] and [148], see [111] for an extension to stochastic average gradients. The heavy ball method forms the basis of the some of the most successful optimization methods for deep learning, see e.g., [240, 135], and the recent review [38]. Hereafter, classical momentum refers to any first-order discretization of the continuous analog of Polyak's heavy ball (with possibly suboptimal step sizes).

Nesterov obtained upper and lower bounds of matching order for first-order methods for smooth convex functions and smooth strongly convex functions, see [193]. In Necoara *et al.* [187], the assumption of strong convexity was relaxed, and under a weaker quadratic growth condition, linear rates were obtained by several well known optimization methods. Several other authors obtained linear rates for various classes

of non-strongly convex or non-uniformly smooth functions, see e.g., [188, 133, 73, 272, 78, 223].

In recent years, there has been interest in the optimization community in looking at the continuous time ODE limit of optimization methods, when the step size tends to zero. Su *et al.* [238, 239] have found the continuous time limit of Nesterov's accelerated gradient descent. This result improves the intuition about Nesterov's method, as the proofs of convergence rates in continuous time are rather elegant and clear, while the previous proofs in discrete time are not as transparent. Follow-ups have studied the continuous time counterparts to accelerated mirror descent [139] as well as higher order discretizations of such systems [263, 266]. Studying continuous time systems for optimization can separate the concerns of designing an optimizer from the difficulties of discretization. This perspective has resulted in numerous other recent works that propose new optimization methods, and study existing ones via their continuous time limit, see e.g., [29, 7, 80, 126, 72, 83, 84].

Conformal Hamiltonian systems (4.3) are studied in geometry [165, 30], because their solutions preserve symplectic area up to a constant; when $\gamma = 0$ symplectic area is exactly preserved, when $\gamma > 0$ symplectic area dissipates uniformly at an exponential rate [165]. In classical mechanics, Hamiltonian dynamics (system (4.3) with $\gamma = 0$) are used to describe the motion of a particle exposed to the force field ∇f. Here, the most common form for k is $k(p) = \langle p, p \rangle / 2m$, where m is the mass, or in relativistic mechanics, $k(p) = c\sqrt{\langle p, p \rangle + m^2 c^2}$ where c is the speed of light, see [99]. In the Markov Chain Monte Carlo literature, where (discretized) Hamiltonian dynamics (again $\gamma = 0$) are used to propose moves in a Metropolis–Hastings algorithm [170, 115, 74, 185], k is viewed as a degree of freedom that can be used to improve the mixing properties of the Markov chain [93, 149]. Stochastic differential equations similar to (4.3) with $\gamma > 0$ have been studied from the perspective of designing k [157, 237].

4.3 Continuous Dynamics

In this section, we motivate the discrete optimization algorithms by introducing their continuous time counterparts. These systems are differential equations described by a Hamiltonian vector field plus a dissipation field. Thus, we briefly review Hamiltonian dynamics, the continuous dynamics of Hamiltonian descent, and derive convergence rates for convex f in continuous time.

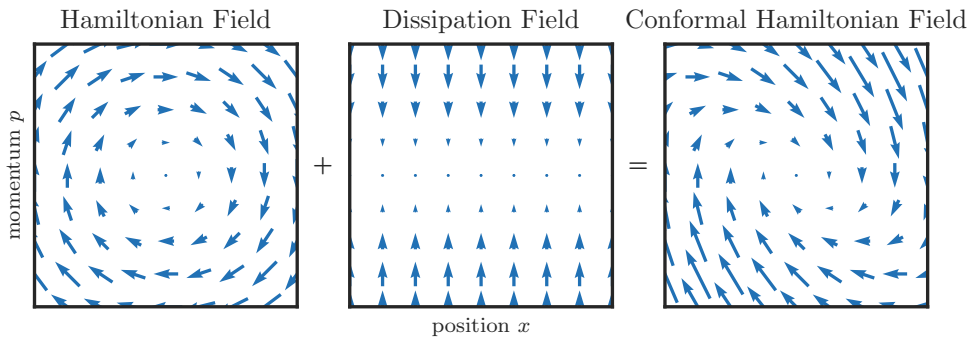


Figure 4.3: A visualization of a conformal Hamiltonian system.

4.3.1 Hamiltonian Systems

In the Hamiltonian formulation of mechanics, the evolution of a particle exposed to a force field ∇f is described by its location $x_t : [0, \infty) \to \mathbb{R}^d$ and momentum $p_t : [0, \infty) \to \mathbb{R}^d$ as functions of time. The system is characterized by the total energy, or Hamiltonian,

$$\mathcal{H}(x, p) = k(p) + f(x) - f(x_{\min}), \tag{4.7}$$

where $x_{\min}$ is one of the global minimizers of f and $k : \mathbb{R}^d \to \mathbb{R}$ is called the kinetic energy. Throughout, we consider kinetic energies k that are a strictly convex functions with minimum at $k(0) = 0$. The Hamiltonian $\mathcal{H}$ defines the trajectory of a particle x_t and its momentum p_t via the ordinary differential equation,

$$\begin{aligned} x_t' &= \nabla_p \mathcal{H}(x_t, p_t) = \nabla k(p_t) \\ p_t' &= -\nabla_x \mathcal{H}(x_t, p_t) = -\nabla f(x_t). \end{aligned} \tag{4.8}$$

For any solution of this system, the value of the total energy over time $\mathcal{H}_t = \mathcal{H}(x_t, p_t)$ is conserved as $\mathcal{H}_t' = \langle \nabla k(p_t), p_t' \rangle + \langle \nabla f(x_t), x_t' \rangle = 0$. Thus, the solutions of the Hamiltonian field oscillate, exchanging energy from x to p and back again.

4.3.2 Continuously Descending the Hamiltonian

The solutions of a Hamiltonian system remain in the level set $\{(x_t, p_t) : \mathcal{H}_t = H_0\}$. To drive such a system towards stationary points, the total energy must reduce over time. Consider as a motivating example the continuous system $x_t'' = -\nabla f(x_t) - \gamma x_t'$, which describes Polyak's heavy ball algorithm in continuous time [206]. Letting $x_t' = p_t$, the

heavy ball system can be rewritten as

$$\begin{aligned} x_t' &= p_t \\ p_t' &= -\nabla f(x_t) - \gamma p_t. \end{aligned} \tag{4.9}$$

Note that this system can be viewed as a combination of a Hamiltonian field with $k(p) = \langle p, p \rangle / 2$ and a dissipation field, i.e., $(x_t', p_t') = F(x_t, p_t) + G(x_t, p_t)$ where $F(x_t, p_t) = (p_t, -\nabla f(x_t))$ and $G(x_t, p_t) = (0, -\gamma p_t)$, see Figure 4.3 for a visualization. This is naturally extended to define the more general conformal Hamiltonian system [165],

$$\begin{aligned} x_t' &= \nabla k(p_t) \\ p_t' &= -\nabla f(x_t) - \gamma p_t. \end{aligned} \tag{4.3 revisited}$$

with $\gamma \in (0, \infty)$. When k is convex with a minimum $k(0) = 0$, these systems descend the level sets of the Hamiltonian. We can see this by showing that the total energy $\mathcal{H}_t$ is reduced along the trajectory (x_t, p_t),

$$\mathcal{H}_t' = \langle \nabla k(p_t), p_t' \rangle + \langle \nabla f(x_t), x_t' \rangle = -\gamma \langle \nabla k(p_t), p_t \rangle \leq -\gamma k(p_t) \leq 0, \tag{4.10}$$

where we have used the convexity of k, and the fact that it is minimised at $k(0) = 0$.

The following proposition shows some existence and uniqueness results for the dynamics (4.3). We say that $\mathcal{H}$ is radially unbounded if $\mathcal{H}(x, p) \to \infty$ when $\|(x, p)\|_2 \to \infty$, e.g., this would be implied if f and k were strictly convex with unique minima.

Proposition 4.3.1 (Existence and uniqueness)**.** *If ∇f and ∇k are continuous, k is convex with a minimum $k(0) = 0$, and $\mathcal{H}$ is radially unbounded, then for every $x, p \in \mathbb{R}^d$, there exists a solution (x_t, p_t) of (4.3) defined for every $t \geq 0$ with $(x_0, p_0) = (x, p)$. If in addition, ∇f and ∇k are continuously differentiable, then this solution is unique.*

Proof. First, only assuming continuity, it follows from Peano's existence theorem [201] that there exists a local solution on an interval $t \in [-a, a]$ for some $a > 0$. Let $[0, A)$ denote the right maximal interval where a solution of (4.3) satisfying that $x_0 = x$ and $p_0 = p$ exist. From (4.10), it follows that $\mathcal{H}_t' \leq 0$, and hence $\mathcal{H}_t \leq \mathcal{H}_0$ for every $t \in [0, A)$. Now by the radial unboundedness of $\mathcal{H}$, and the fact that $\mathcal{H}_t \leq \mathcal{H}_0$, it follows that the compact set $\{(x, p) : \mathcal{H}(x, p) \leq \mathcal{H}_0\}$ is never left by the dynamics, and hence by Theorem 10.1 of [114] (page 140), we must have $A = \infty$. The uniqueness under continuous differentiability follows from the Fundamental Existence–Uniqueness Theorem on page 74 of [202]. □

As shown in the next proposition, (4.10) implies that conformal Hamiltonian systems approach stationary points of f.

Proposition 4.3.2 (Convergence to a stationary point)**.** *Let (x_t, p_t) be a solution to the system* (4.3) *with initial conditions $(x_0, p_0) = (x, p) \in \mathbb{R}^{2d}$, f continuously differentiable, and k continuously differentiable, strictly convex with minimum at* 0 *and $k(0) = 0$. If f is bounded below and $\mathcal{H}$ is radially unbounded, then $\|\nabla f(x_t)\|_2 \to 0$.*

Proof. Since f is bounded below, $\mathcal{H}_t \geq 0$. Since $\mathcal{H}$ is radially unbounded, the set $B := \{(x, p) \in \mathbb{R}^{2d} : \mathcal{H}(x, p) \leq \mathcal{H}(x_0, p_0) + 1\}$ is a compact set that contains (x_0, p_0) in its interior. Moreover, by (4.10), we also have $(x_t, p_t) \in B$ for all $t > 0$. Consider the set $M = \{(x_t, p_t) : \mathcal{H}'_t = 0\} \cap B$. Since k is strictly convex, this set is equivalent to $\{(x_t, p_t) : \|p_t\|_2 = 0\} \cap B$. The largest invariant set of the dynamics (4.3) inside M is $I = \{(x, p) \in \mathbb{R}^{2d} : \|p\|_2 = 0, \|\nabla f(x)\|_2 = 0\} \cap B$. By LaSalle's principle [145], all trajectories started from B must approach I. Since f is a continuous bounded function on the compact set B, there is a point $x_* \in B$ such that $f(x_*) \leq f(x)$ for every $x \in B$ (i.e. the minimum is attained in B) by the extreme value theorem (see [224]). Moreover, due to the definition of B, x_* is in its interior, hence $\|\nabla f(x_*)\|_2 = 0$ and therefore $(x_*, 0) \in I$. Thus the set I is non-empty (note that I might contain other local minima as well). □

Remark 4.3.3. This construction can be generalized by modifying the $-\gamma p_t$ component of (4.3) to a more general dissipation field $-\gamma D(p_t)$. If the dissipation field is everywhere aligned with the kinetic map, $\langle \nabla k(p), D(p) \rangle \geq 0$, then these systems dissipate energy. We have not found alternatives to $D(p) = \gamma p$ that result in linear convergence in general.

4.3.3 Continuous Hamiltonian Descent on Convex Functions

In this section we study how k can be designed to condition the system (4.3) for linear convergence in $\log(f(x_t) - f(x_{\min}))$. Although the solutions x_t, p_t of (4.3) approach stationary points under weak conditions, to derive rates we consider the case when f is convex. To motivate our choice of k, consider the quadratic function $f(x) = \langle x, Ax \rangle / 2$ with $k(p) = \langle p, A^{-1}p \rangle / 2$ for positive definite symmetric $A \in \mathbb{R}^{d \times d}$. Now (4.3) becomes,

$$\begin{aligned} x'_t &= A^{-1} p_t \\ p'_t &= -A x_t - \gamma p_t. \end{aligned} \tag{4.11}$$

By the change of variables $v_t = A^{-1} p_t$, this is equivalent to

$$\begin{aligned} x'_t &= v_t \\ v'_t &= -x_t - \gamma v_t, \end{aligned} \tag{4.12}$$

which is a universal equation and hence the convergence rate of (4.11) is independent of A. Although this kinetic energy implements a constant preconditioner for any f, for this specific f k is its convex conjugate f^*. This suggests the core idea of this paper: taking k related in some sense to f^* for more general convex functions may condition the convergence of (4.3). Indeed, we show in this section that, if the kinetic energy $k(p)$ upper bounds a centered version of $f^*(p)$, then the convergence of (4.3) is linear.

More precisely, define the following centered function $f_c : \mathbb{R}^d \to \mathbb{R}$,

$$f_c(x) = f(x + x_{\min}) - f(x_{\min}). \tag{4.13}$$

The convex conjugate of f_c is given by $f_c^*(p) = f^*(p) - \langle x_{\min}, p\rangle + f(x_{\min})$ and is minimized at $f_c^*(0) = 0$. Importantly, as we will show in the final lemma of this section, taking a kinetic energy such that $k(p) \geq \alpha \max(f_c^*(p), f_c^*(-p))$ for some $\alpha \in (0, 1]$ suffices to achieve linear rates on any differentiable convex f in continuous time. The constant α is included to capture the fact that k may under estimate f_c^* by some constant factor, so long as it is positive. If α does not depend in any fashion on f, then the convergence rate of (4.3) is independent of f. In Section 4.3.4 we also show a partial converse — for some simple problems taking a k not satisfying those assumptions results in sub-linear convergence for almost every path (except for one unique curve and its mirror).

Remark 4.3.4. There is an interesting connection to duality theory for a specific choice of k. In a slight abuse of representation, consider rewriting the original problem as

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{2}(f(x) + f(x)).$$

The Fenchel dual of this problem is equivalent to the following problem after a small reparameterization of p (see Chapter 31 of [222]),

$$\max_{p\in\mathbb{R}^d} \frac{1}{2}(-f^*(p) - f^*(-p)).$$

The Fenchel duality theorem guarantees that for a given pair of primal-dual variables $(x, p) \in \mathbb{R}^d$, the duality gap between the primal objective $f(x)$ and the dual objective $(-f^*(p) - f^*(-p))/2$ is positive. Thus,

$$\begin{aligned} f(x) - (-f^*(p) - f^*(-p))/2 &= f(x) - f(x_{\min}) + (f^*(p) + f^*(-p))/2 + f(x_{\min}) \\ &= f(x) - f(x_{\min}) + (f_c^*(p) + f_c^*(-p))/2 \geq 0. \end{aligned}$$

Thus, for the choice $k(p) = (f_c^*(p) + f_c^*(-p))/2$, which as we will show implies linear convergence of (4.3), the Hamiltonian $\mathcal{H}(x,p)$ is exactly the duality gap between the primal and dual objectives.

Linear rates in continuous time can be derived by a Lyapunov function $\mathcal{V} : \mathbb{R}^{d\times d} \to [0,\infty)$ that summarizes the total energy of the system, contracts exponentially (or linearly in log-space), and is positive unless $(x_t, p_t) = (x_{\min}, 0)$. Ultimately we are trying to prove a result of the form $\mathcal{V}_t' \leq -\lambda \mathcal{V}_t$ for some rate $\lambda > 0$. As the energy $\mathcal{H}_t$ is decreasing, it suggests using $\mathcal{H}_t$ as a Lyapunov function. Unfortunately, this will not suffice, as $\mathcal{H}_t$ plateaus instantaneously ($\mathcal{H}_t' = 0$) at points on the trajectory where $p_t = 0$ despite x_t possibly being far from $x_{\min}$. However, when $p_t = 0$, the momentum field reduces to the term $-\nabla f(x_t)$ and the derivative of $\langle x_t - x_{\min}, p_t\rangle$ in t is instantaneously strictly negative $-\langle x_t - x_{\min}, \nabla f(x_t)\rangle < 0$ for convex f (unless we are at $(x_{\min}, 0)$). This suggests the family of Lyapunov functions that we study in this paper,

$$\mathcal{V}(x,p) = \mathcal{H}(x,p) + \beta \langle x - x_{\min}, p\rangle, \tag{4.14}$$

where $\beta \in (0,\gamma)$ (see the next lemma for conditions that guarantee that it is non-negative). As with $\mathcal{H}$, $\mathcal{V}_t$ is used to indicate $\mathcal{V}(x_t, p_t)$ at time t along a solution to (4.3). Before moving on to the final lemma of the section, we prove two technical lemmas that will give us useful control over $\mathcal{V}$ throughout the paper.

The first lemma describes how β must be constrained for $\mathcal{V}$ to be positive and to track $\mathcal{H}$ closely, so that it is useful for the analysis of the convergence of $\mathcal{H}$ and ultimately f.

Lemma 4.3.5 (Bounding the ratio of $\mathcal{H}$ and $\mathcal{V}$)**.** *Let $x \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ convex with unique minimum $x_{\min}$, $k : \mathbb{R}^d \to \mathbb{R}$ strictly convex with minimum $k(0) = 0$, $\alpha \in (0,1]$ and $\beta \in (0,\alpha]$.*
If $p \in \mathbb{R}^d$ is such that $k(p) \geq \alpha f_c^*(-p)$, then

$$\langle x - x_{\min}, p\rangle \geq -\left(k(p)/\alpha + f(x) - f(x_{\min})\right) \geq -\frac{\mathcal{H}(x,p)}{\alpha}, \tag{4.15}$$

$$\tfrac{\alpha-\beta}{\alpha}\mathcal{H}(x,p) \leq \mathcal{V}(x,p). \tag{4.16}$$

If $p \in \mathbb{R}^d$ is such that $k(p) \geq \alpha f_c^*(p)$, then

$$\langle x - x_{\min}, p\rangle \leq k(p)/\alpha + f(x) - f(x_{\min}) \leq \frac{\mathcal{H}(x,p)}{\alpha}, \tag{4.17}$$

$$\mathcal{V}(x,p) \leq \tfrac{\alpha+\beta}{\alpha}\mathcal{H}(x,p). \tag{4.18}$$

Proof. Assuming that $k(p) \geq \alpha f_c^*(-p)$, we have

$$\begin{aligned} k(p)/\alpha + f_c(x - x_{\min}) &\geq f_c^*(-p) + f_c(x - x_{\min}) \\ &\geq \langle x - x_{\min}, -p \rangle - f_c(x - x_{\min}) + f_c(x - x_{\min}) \\ &= -\langle x - x_{\min}, p \rangle , \end{aligned}$$

hence we have (4.15). (4.16) follows by rearrangement. The proof of (4.17) and (4.18) is similar. □

Lemma 4.3.5 constrains β in terms of α. For a result like $\mathcal{V}_t' \leq -\lambda \mathcal{V}_t$, we will need to control β in terms of the magnitude γ of the dissipation field. The following lemma provides constraints on β and, under those constraints, the optimal β. The proof can be found in Section Ap.1 of the Appendix.

Lemma 4.3.6 (Convergence rates in continuous time for fixed α)**.** *Given $\gamma \in (0, 1)$, $f : \mathbb{R}^d \to \mathbb{R}$ differentiable and convex with unique minimum $x_{\min}$, $k : \mathbb{R}^d \to \mathbb{R}$ differentiable and strictly convex with minimum $k(0) = 0$. Let $x_t, p_t \in \mathbb{R}^d$ be the value at time t of a solution to the system (4.3) such that there exists $\alpha \in (0, 1]$ where $k(p_t) \geq \alpha f_c^*(-p_t)$. Define*

$$\lambda(\alpha, \beta, \gamma) = \min\left(\frac{\alpha\gamma - \alpha\beta - \beta\gamma}{\alpha - \beta}, \frac{\beta(1-\gamma)}{1-\beta}\right). \tag{4.19}$$

If $\beta \in (0, \min(\alpha, \gamma)]$, then

$$\mathcal{V}_t' \leq -\lambda(\alpha, \beta, \gamma)\mathcal{V}_t.$$

Finally,

1. *The optimal $\beta \in (0, \min(\alpha, \gamma)]$, $\beta^\star = \arg\max_\beta \lambda(\alpha, \beta, \gamma)$ and $\lambda^\star = \lambda(\alpha, \beta^\star, \gamma)$ are given by,*

$$\beta^\star = \tfrac{1}{1+\alpha}\left(\alpha + \tfrac{\gamma}{2} - \sqrt{(1-\gamma)\alpha^2 + \tfrac{\gamma^2}{4}}\right), \tag{4.20}$$

$$\lambda^\star = \begin{cases} \tfrac{1}{1-\alpha}\left((1-\gamma)\alpha + \tfrac{\gamma}{2} - \sqrt{(1-\gamma)\alpha^2 + \tfrac{\gamma^2}{4}}\right) & \text{for } 0 < \alpha < 1, \\ \tfrac{\gamma(1-\gamma)}{2-\gamma} \text{ for } \alpha = 1, \end{cases} \tag{4.21}$$

2. *If $\beta \in (0, \alpha\gamma/2]$, then*

$$\lambda(\alpha, \beta, \gamma) = \frac{\beta(1-\gamma)}{1-\beta}, \text{ and} \tag{4.22}$$

$$\begin{aligned} &-\left(\gamma - \beta - \gamma^2(1-\gamma)/4\right) k(p_t) - \beta\gamma \langle x_t - x_{\min}, p_t \rangle - \beta \langle x_t - x_{\min}, \nabla f(x_t) \rangle \\ &\leq -\beta(1-\gamma)(k(p_t) + f(x_t) - f(x_{\min}) + \beta \langle x_t - x_{\min}, p_t \rangle). \end{aligned} \tag{4.23}$$

These two lemmas are sufficient to prove the linear contraction of $\mathcal{V}$ and the contraction $f(x_t) - f(x_{\min}) \leq \frac{\alpha}{\alpha-\beta^\star}\mathcal{H}_0 \exp(-\lambda^\star t)$ under the assumption of constant α and β. Still, the constant α, which controls our approximation of f_c^* may be quite pessimistic if it must hold globally along x_t, p_t as the system converges to its minimum. Instead, in the final lemma that collects the convergence result for this section, we consider the case where α may increase as convergence proceeds. To support an improving α, our constant β will now have to vary with time and we will be forced to take slightly suboptimal β and λ given by (4.22) of Lemma 4.3.6. Still, the improving α will be important in future sections for ensuring that we are able to achieve position independent step sizes.

We are now ready to present the central result of this section. Under Assumptions A we show linear convergence of (4.3). In general, the dependence of the rate of linear convergence on f is via the function α and the constant $C_{\alpha,\gamma}$ in our analysis.

Assumption A. *A.1* $f : \mathbb{R}^d \to \mathbb{R}$ *differentiable and convex with unique minimum* $x_{\min}$.

A.2 $k : \mathbb{R}^d \to \mathbb{R}$ *differentiable and strictly convex with minimum* $k(0) = 0$.

A.3 $\gamma \in (0, 1)$.

A.4 There exists some differentiable non-increasing convex function $\alpha : [0, \infty) \to (0, 1]$ *and constant* $C_{\alpha,\gamma} \in (0, \gamma\,]$ *such that for every* $p \in \mathbb{R}^d$,

$$k(p) \geq \alpha(k(p)) \max(f_c^*(p), f_c^*(-p)) \tag{4.24}$$

and that for every $y \in [0, \infty)$

$$-C_{\alpha,\gamma}\alpha'(y)y < \alpha(y). \tag{4.25}$$

In particular, if $k(p) \geq \alpha_\star \max(f_c^*(p), f_c^*(-p))$ *for a constant* $\alpha_\star \in (0, 1]$, *then the constant function* $\alpha(y) = \alpha_\star$ *serves as a valid, but pessimistic choice.*

Remark 4.3.7. Assumption A.4 can be satisfied if a symmetric lower bound on f is known. For example, strong convexity implies

$$f(x + x_{\min}) - f(x_{\min}) \geq \frac{\mu}{2} \|x\|_2^2 .$$

This in turn implies $f_c^*(p) \leq \|p\|_2^2 / (2\mu)$. Because $k(p) = \|p\|_2^2 / (2\mu)$ is symmetric, it satisfies A.4 which explains why conditions relating to strong convexity are necessary for linear convergence of Polyak's heavy ball.

Theorem 4.3.8 (Convergence bound in continuous time with general α)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$ satisfying Assumptions A. Let (x_t, p_t) be a solution to the system (4.3) with initial states $(x_0, p_0) = (x, 0)$ where $x \in \mathbb{R}^d$. Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, $\lambda = \frac{(1-\gamma)C_{\alpha,\gamma}}{4}$, and $\mathcal{W} : [0,\infty) \to [0,\infty)$ be the solution of*

$$\mathcal{W}_t' = -\lambda \cdot \alpha(2\mathcal{W}_t)\mathcal{W}_t,$$

with $\mathcal{W}_0 := \mathcal{H}_0 = f(x_0) - f(x_{\min})$. Then for every $t \in [0,\infty)$, we have

$$f(x_t) - f(x_{\min}) \leq 2\mathcal{H}_0 \exp\left(-\lambda \int_0^t \alpha(2\mathcal{W}_t)\right) \leq 2\mathcal{H}_0 \exp\left(-\lambda\alpha_\star t\right). \tag{4.26}$$

Proof. By (4.24) in assumption A.4, the conditions of Lemma 4.3.5 hold, and by (4.15) and (4.17) we have

$$|\langle x_t - x_{\min}, p_t\rangle| \leq k(p_t)/\alpha(k(p_t)) + f(x_t) - f(x_{\min}) \leq \frac{\mathcal{H}_t}{\alpha(k(p_t))}. \tag{4.27}$$

Instead of defining the Lyapunov function $\mathcal{V}_t$ exactly as in (4.14) we take a time-dependent β_t. Specifically, for every $t \geq 0$ let $\mathcal{V}_t$ be the unique solution v of the equation

$$v = \mathcal{H}_t + \frac{C_{\alpha,\gamma}\alpha(2v)}{2}\langle x_t - x_{\min}, p_t\rangle \tag{4.28}$$

in the interval $v \in [\mathcal{H}_t/2, 3\mathcal{H}_t/2]$. To see why this equation has a unique solution in $v \in [\mathcal{H}_t/2, 3\mathcal{H}_t/2]$, note that from (4.27) it follows that

$$|\alpha(2v)\langle x_t - x_{\min}, p_t\rangle| \leq \mathcal{H}_t \text{ for every } v \geq \frac{\mathcal{H}_t}{2},$$

and hence for any such v, we have

$$\frac{\mathcal{H}_t}{2} \leq \mathcal{H}_t + \frac{C_{\alpha,\gamma}\alpha(2v)}{2}\langle x_t - x_{\min}, p_t\rangle \leq \frac{3}{2}\mathcal{H}_t. \tag{4.29}$$

This means that for $v = \frac{\mathcal{H}_t}{2}$, the left hand side of (4.28) is smaller than the right hand side, while for $v = \frac{3\mathcal{H}_t}{2}$, it is the other way around. Now using (4.25) in assumption A.4 and (4.27), we have

$$|C_{\alpha,\gamma}\alpha'(2\mathcal{V}_t)\langle x_t - x_{\min}, p_t\rangle| \leq \left|C_{\alpha,\gamma}\frac{\alpha'(2\mathcal{V}_t)2\mathcal{V}_t}{\alpha(2\mathcal{V}_t)}\right| < 1, \tag{4.30}$$

Thus, by differentiation, we can see that (4.30) implies that

$$\frac{\partial}{\partial v}\left(v - \mathcal{H}_t - \tfrac{C_{\alpha,\gamma}}{2}\alpha(2v)\langle x_t - x_{\min}, p_t\rangle\right) > 0,$$

which implies that (4.28) has a unique solution $\mathcal{V}_t$ in $[\frac{\mathcal{H}}{2}, \frac{3\mathcal{H}_t}{2}]$. Let $\alpha_t = \alpha(2\mathcal{V}_t)$ and $\beta_t = \frac{C_{\alpha,\gamma}}{2}\alpha(2\mathcal{V}_t)$. By the implicit function theorem, it follows that $\mathcal{V}_t$ is differentiable in t. Morover, since

$$\mathcal{V}_t = \mathcal{H}_t + \frac{C_{\alpha,\gamma}\alpha(2\mathcal{V}_t)}{2}\langle x_t - x_{\min}, p_t\rangle \tag{4.31}$$

for every $t \geq 0$, by differentiating both sides, we obtain that

$$\begin{aligned}\mathcal{V}_t' = &-(\gamma - \beta_t)\langle \nabla k(p_t), p_t\rangle - \beta_t\gamma\langle x_t - x_{\min}, p_t\rangle - \beta_t\langle x_t - x_{\min}, \nabla f(x_t)\rangle \\ &+ \beta_t'\langle x_t - x_{\min}, p_t\rangle\end{aligned}$$

The first three terms are equivalent to the temporal derivative of $\mathcal{V}_t$ with constant $\beta = \beta_t$. Since $\alpha_t \leq \alpha(k(p_t))$ and $\beta_t \leq \gamma$, the assumptions of Lemma 4.3.6 are satisfied locally for α_t, β_t and we get

$$\mathcal{V}_t' \leq -\lambda(\alpha_t, \beta_t, \gamma)\mathcal{V}_t + \beta_t'\langle x_t - x_{\min}, p_t\rangle = -\lambda(\alpha_t, \beta_t, \gamma)\mathcal{V}_t + C_{\alpha,\gamma}\alpha_t'\langle x_t - x_{\min}, p_t\rangle\,\mathcal{V}_t'.$$

Using (4.22) of Lemma 4.3.6 for α_t, β_t, we have $\lambda(\alpha_t, \beta_t, \gamma) = \frac{\beta_t(1-\gamma)}{1-\beta_t} \geq \beta_t(1-\gamma)$ and

$$\mathcal{V}_t' \leq -\beta_t(1-\gamma)\mathcal{V}_t + C_{\alpha,\gamma}\alpha_t'\langle x_t - x_{\min}, p_t\rangle\,\mathcal{V}_t'.$$

Using (4.30) we have $\mathcal{V}_t' \leq -\frac{\beta_t(1-\gamma)}{2}\mathcal{V}_t$. Notice that $\mathcal{V}_0 = \mathcal{H}_0$ since we have assumed that $p_0 = 0$, and the claim of the lemma follows by Grönwall's inequality. The final inequality (4.26) follows from the fact that $\alpha(2\mathcal{V}_t) \geq \alpha(3\mathcal{H}_0) = \alpha_\star$. □

4.3.4 Partial Lower Bounds

In this section we consider a partial converse of Proposition 4.3.8, showing in a simple setting that if the assumption $k(p) \geq \alpha\max(f_c^*(p), f_c^*(-p))$ of A.4 is violated, then the ODE (4.3) contracts sub-linearly. Figure 4.4 considers the example $f(x) = x^4/4$. If $k(p) = |p|^a/a$, then assumptions A cannot be satisfied for small p unless $b \geq 4/3$. Figure 4.4 shows that an inappropriate choice of $k(p) = p^2/2$ leads to sub-linear convergence both in continuous time and for one of the discretizations of Section 4.4. In contrast, the choice of $k(p) = 3p^{4/3}/4$ results in linear convergence, as expected.

Let $b, a > 1$ and $\gamma > 0$. For $d = 1$ dimension, with the choice $f(x) := |x|^b/b$ and $k(p) := |p|^a/a$, (4.3) takes the following form,

$$\begin{aligned}x_t' &= |p_t|^{a-1}\operatorname{sign}(p_t), \\ p_t' &= -|x_t|^{b-1}\operatorname{sign}(x_t) - \gamma p_t.\end{aligned} \tag{4.32}$$

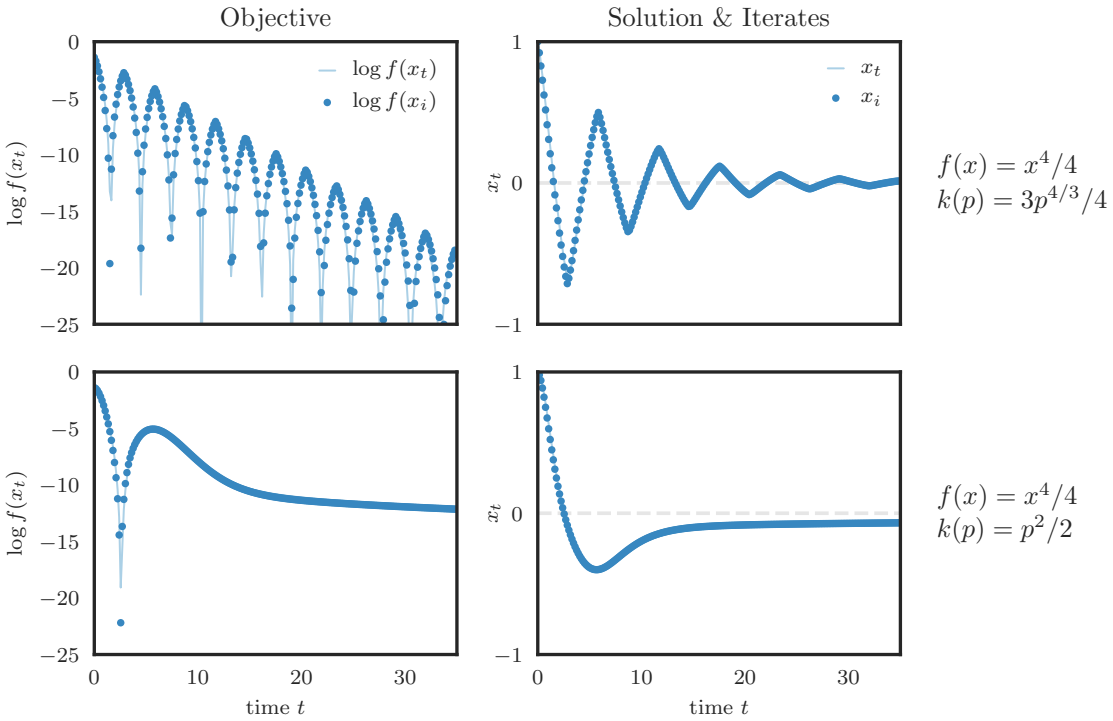


Figure 4.4: Importance of Assumptions A. Solutions x_t and iterates x_i of our first explicit method on $f(x) = x^4/4$ with two different choices of k. Notice that $f_c^*(p) = 3p^{4/3}/4$ and thus $k(p) = p^2/2$ cannot be made to satisfy assumption A.4.

Since $f(x)$ takes its minimum at 0, (x_t, p_t) are expected to converge to $(0, 0)$ as $t \to \infty$. There is a trivial solution: $x_t = p_t = 0$ for every $t \in \mathbb{R}$. The following Lemma shows an existence and uniqueness result for this equation. The proof is included in Section Ap.2 of the Appendix.

Lemma 4.3.9 (Existence and uniqueness of solutions of the ODE)**.** *Let $a, b, \gamma \in (0, \infty)$. For every $t_0 \in \mathbb{R}$ and $(x, p) \in \mathbb{R}^2$, there is a unique solution $(x_t, p_t)_{t\in\mathbb{R}}$ of the ODE (4.32) with $x_{t_0} = x$, $p_{t_0} = p$. Either $x_t = p_t = 0$ for every $t \in \mathbb{R}$, or $(x_t, p_t) \neq (0, 0)$ for every $t \in \mathbb{R}$.*

Note that if (x_t, p_t) is a solution, and $\Delta \in \mathbb{R}$, then $(x_{t+\Delta}, p_{t+\Delta})$ is also a solution (time translation), and $(-x_t, -p_t)$ is also a solution (central symmetry).

Note also that $f^*(p) = f^*(-p) = |p|^{b^*}/b^*$ for $b^* := (1 - \frac{1}{b})^{-1}$. Hence if $a \leq b^*$, or equivalently, if $\frac{1}{b} + \frac{1}{a} \geq 1$, the conditions of Proposition 4.3.8 are satisfied for some $\alpha > 0$ (in particular, if $a = b^*$, then $\alpha = 1$ independently of x_0, p_0). Hence in

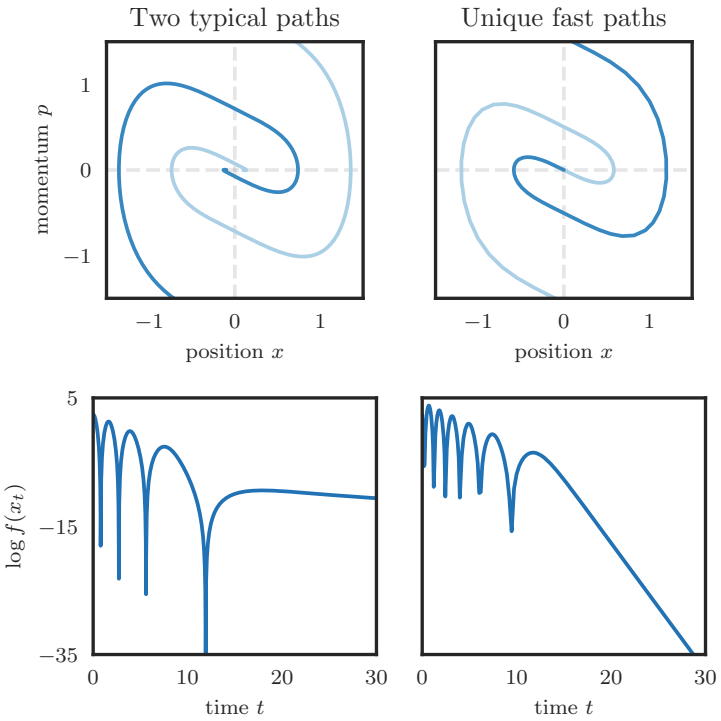


Figure 4.5: Solutions to the Hamiltonian descent system with $f(x) = x^4/4$ and $k(p) = x^2/2$. The right plots show a numerical approximation of $(x_t^{(\eta)}, p_t^{(\eta)})$ and $(-x_t^{(\eta)}, -p_t^{(\eta)})$. The left plots show a numerical approximation of $(x_t^{(\theta)}, p_t^{(\theta)})$ and $(-x_t^{(\theta)}, -p_t^{(\theta)})$ for $\theta = \eta + \delta \in \mathbb{R}$, which represent typical paths.

such cases, the speed of convergence is linear. For $a > b^*$, $\lim_{p\to 0} \frac{K(p)}{f^*(p)} = 0$, so the conditions of Proposition 4.3.8 are violated.

Now we are ready to state the main result in this section, a theorem characterizing the convergence speeds of (x_t, p_t) to $(0, 0)$ in this situation. The proof is included in Section Ap.2 of the Appendix.

Proposition 4.3.10 (Lower bounds on the convergence rate in continuous time)**.** *Suppose that* $\frac{1}{b} + \frac{1}{a} < 1$. *For any* $\theta \in \mathbb{R}$, *we denote by* $(x_t^{(\theta)}, p_t^{(\theta)})$ *the unique solution of* (4.32) *with* $x_0 = \theta, p_0 = 0$. *Then there exists a constant* $\eta \in (0, \infty)$ *depending on* a *and* b *such that the path* $(x_t^{(\eta)}, p_t^{(\eta)})$ *and its mirrored version* $(x_t^{(-\eta)}, p_t^{(-\eta)})$ *satisfy that*

$$|x_t^{(-\eta)}| = |x_t^{(\eta)}| \leq \mathcal{O}(\exp(-\alpha t)) \text{ for every } \alpha < \gamma(a-1) \text{ as } t \to \infty.$$

For any path (x_t, p_t) *that is not a time translation of* $(x_t^{(\eta)}, p_t^{(\eta)})$ *or* $(x_t^{(-\eta)}, p_t^{(-\eta)})$, *we have*

$$\left|x_t^{-1}\right| = O(t^{\frac{1}{ba-b-a}}) \text{ as } t \to \infty,$$

so the speed of convergence is sub-linear and not linearly fast.

Figure 4.5 illustrates the two paths where the convergence is linearly fast for $a = 2, b = 4$. The main idea in the proof of Proposition 4.3.10 is that we establish the existence of a class of trapping sets, i.e. once the path of the ODE enters one of them, it never escapes. Convergence rates within such sets can be shown to be logarithmic, and it is established that only two paths (which are symmetric with respect to the origin) avoid each one of the trapping sets, and they have linear convergence rate.

4.4 Optimization Algorithms

In this section we consider three discretizations of the continuous system (4.3), one implicit and two explicit. For these discretizations we must assume more about the relationship between f and k. The implicit method defines the iterates as solution of a local subproblem. The first and second explicit methods are fully explicit, and we must again make stronger assumptions on f and k. The proofs of all of the results in this section are given in Section Ap.3 of the Appendix.

4.4.1 Implicit Method

Consider the following discrete approximation (x_i, p_i) to the continuous system, making the fixed $\epsilon > 0$ finite difference approximation, $\frac{x_{i+1}-x_i}{\epsilon} = x_t'$ and $\frac{p_{i+1}-p_i}{\epsilon} = p_t'$, which approximates the field at the forward points.

$$\begin{aligned} \frac{x_{i+1} - x_i}{\epsilon} &= \nabla k(p_{i+1}) \\ \frac{p_{i+1} - p_i}{\epsilon} &= -\gamma p_{i+1} - \nabla f(x_{i+1}). \end{aligned} \tag{4.33}$$

Since $\nabla\, k^*(\nabla\, k(p)) = p$, this system of equations corresponds to the stationary condition of the following subproblem iteration, which we introduce as our implicit method.

Method 1 (Implicit Method)**.** *Given $f, k : \mathbb{R}^d \to \mathbb{R}$, $\epsilon, \gamma \in (0, \infty)$, $x_0, p_0 \in \mathbb{R}^d$.*
Let $\delta = (1 + \gamma\epsilon)^{-1}$ and

$$\begin{aligned} x_{i+1} &= \arg\min_{x \in \mathbb{R}^d} \left\{ \epsilon k^*\left(\tfrac{x - x_i}{\epsilon}\right) + \epsilon\delta f(x) - \delta \left\langle p_i, x \right\rangle \right\} \\ p_{i+1} &= \delta p_i - \epsilon\delta \nabla f(x_{i+1}). \end{aligned} \tag{4.34}$$

The following lemma shows that the formulation (4.34) is well defined. The proof is included in Section Ap.3 of the Appendix.

Lemma 4.4.1 (Well-definedness of the implicit scheme)**.** *Suppose that f and k satisfy assumptions A.1 and A.2, and $\epsilon, \gamma \in (0, \infty)$. Then* (4.34) *has a unique solution for every $x_i, p_i \in \mathbb{R}^d$, and this solution also satisfies* (4.33)*.*

As this discretization involves solving a potentially costly subproblem at each iteration, it requires a relatively light assumption on the compatibility of f and k.

Assumption B. *B.1 There exists $C_{f,k} \in (0, \infty)$ such that for all $x, p \in \mathbb{R}^d$,*

$$|\langle \nabla f(x), \nabla k(p) \rangle| \leq C_{f,k} \mathcal{H}(x, p). \tag{4.35}$$

Remark 4.4.2. Smoothness of f implies $\frac{1}{2}\|\nabla f(x)\|_2^2 \leq L(f(x) - f(x_{\min}))$ (see (2.1.7) of Theorem 2.1.5 of [193]). Thus, if f is smooth and $k(p) = \frac{1}{2}\|p\|_2^2$, then the assumption B.1 can be satisfied by $C_{f,k} = \max\{1, L\}$, since

$$|\langle \nabla f(x), \nabla k(p) \rangle| \leq \tfrac{1}{2}\|\nabla f(x)\|_2^2 + \tfrac{1}{2}\|\nabla k(p)\|_2^2 \leq L(f(x) - f(x_{\min})) + k(p).$$

The following proposition shows a convergence result for the implicit scheme.

Proposition 4.4.3 (Convergence bound for the implicit scheme)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$, and $C_{f,k}$ satisfying assumptions A and B. Suppose that $\epsilon < \frac{1-\gamma}{2\max(C_{f,k},1)}$. Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, and let $\mathcal{W}_0 = f(x_0) - f(x_{\min})$ and for $i \geq 0$,*

$$\mathcal{W}_{i+1} = \mathcal{W}_i \left[1 + \epsilon C_{\alpha,\gamma}(1 - \gamma - 2C_{f,k}\epsilon)\alpha(2\mathcal{W}_i)/4\right]^{-1}.$$

Then for any (x_0, p_0) with $p_0 = 0$, the iterates of (4.33) *satisfy for every $i \geq 0$,*

$$f(x_i) - f(x_{\min}) \leq 2\mathcal{W}_i \leq 2\mathcal{W}_0[1 + \epsilon C_{\alpha,\gamma}(1 - \gamma - 2C_{f,k}\epsilon)\alpha_\star/4]^{-i}.$$

Remark 4.4.4. Proposition 4.4.3 means that we can fix any step size $0 < \epsilon < \frac{1-\gamma}{2\max(C_{f,k},1)}$ independently of the initial point, and have linear convergence with contraction rate that is proportional to $\alpha(3\mathcal{H}_0)$ initially and possibly increasing as we get closer to the optimum. In Section 4.5 we introduce kinetic energies $k(p)$ that behave like $\|p\|_2^a$ near 0 and $\|p\|_2^A$ in the tails. We will show that for functions $f(x)$ that behave like $\|x - x_{\min}\|_2^b$ near their minima and $\|x - x_{\min}\|_2^B$ in the tails the conditions of assumptions B are satisfied as long as $\frac{1}{a} + \frac{1}{b} = 1$ and $\frac{1}{A} + \frac{1}{B} \geq 1$. In particular, if we choose $k(p) = \sqrt{\|p\|_2^2 + 1} - 1$ (relativistic kinetic energy), then $a = 2$ and $A = 1$, and assumptions B can be shown to hold for every f that has quadratic behavior near its minimum and no faster than exponential growth in the tails.

4.4.2 First Explicit Method, with Analysis via the Hessian of f

The following discrete approximation (x_i, p_i) to the continuous system makes a similar finite difference approximation, $\frac{x_{i+1}-x_i}{\epsilon} = x_t'$ and $\frac{p_{i+1}-p_i}{\epsilon} = p_t'$ for $\epsilon > 0$. In contrast to the implicit method, it approximates the field at the point (x_i, p_{i+1}), making it fully explicit without any costly subproblem,

$$\begin{aligned}\frac{x_{i+1} - x_i}{\epsilon} &= \nabla k(p_{i+1}) \\ \frac{p_{i+1} - p_i}{\epsilon} &= -\gamma p_{i+1} - \nabla f(x_i).\end{aligned}$$

This method can be rewritten as our first explicit method.

Method 2 (First Explicit Method)**.** *Given $f, k : \mathbb{R}^d \to \mathbb{R}$, $\epsilon, \gamma \in (0, \infty)$, $x_0, p_0 \in \mathbb{R}^d$. Let $\delta = (1 + \gamma\epsilon)^{-1}$ and*

$$\begin{aligned}p_{i+1} &= \delta p_i - \epsilon\delta\nabla f(x_i) \\ x_{i+1} &= x_i + \epsilon\nabla k(p_{i+1}).\end{aligned} \tag{4.36}$$

This discretization exploits the convexity of k by approximating the continuous dynamics at the forward point p_{i+1}, but is made explicit by approximating at the backward point x_i. Because this method approximates the field at the backward point x_i it requires a kind of smoothness assumption to prevent f from changing too rapidly between iterates. This assumption is in the form of a condition on the Hessian of f, and thus we require twice differentiability of f for the first explicit method. Because the accumulation of gradients of f in the form of p_i are modulated by k, this condition in fact expresses a requirement on the interaction between ∇k and $\nabla^2 f$, see assumption C.3.

Assumption C. *C.1 There exists $C_k \in (0, \infty)$ such that for every $p \in \mathbb{R}^d$,*

$$\langle \nabla k(p), p\rangle \leq C_k k(p). \tag{4.37}$$

C.2 $f : \mathbb{R}^d \to \mathbb{R}$ convex with a unique minimum at $x_{\min}$ and twice continuously differentiable for every $x \in \mathbb{R}^d \setminus \{x_{\min}\}$.

C.3 There exists $D_{f,k} \in (0, \infty)$ such that for every $p \in \mathbb{R}^d$, $x \in \mathbb{R}^d \setminus \{x_{\min}\}$,

$$\left\langle \nabla k(p), \nabla^2 f(x)\nabla k(p)\right\rangle \leq D_{f,k}\alpha(3\mathcal{H}(x,p))\mathcal{H}(x,p). \tag{4.38}$$

Remark 4.4.5. If f smooth and twice differentiable then $\langle v, \nabla^2 f(x) v \rangle$ is everywhere bounded by L for $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$ (see Theorem 2.1.6 of [193]). Thus, using $k(p) = \frac{1}{2}\|p\|_2^2$, this allows us to satisfy assumption C.3 with $D_{f,k} = \max\{1, 2L\}$, since

$$\left\langle \nabla k(p), \nabla^2 f(x) \nabla k(p) \right\rangle \leq L \|\nabla k(p)\|_2^2 = 2Lk(p) \leq f(x) - f(x_{\min}) + 2Lk(p).$$

Assumption C.1 is clearly satisfied in this case by $C_k = 2$.

The following lemma shows a convergence result for this discretization.

Proposition 4.4.6 (Convergence bound for the first explicit scheme)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$, $C_{f,k}$, C_k, $D_{f,k}$ satisfying assumptions A, B, and C, and that $0 < \epsilon < \min\left(\frac{1-\gamma}{2\max(C_{f,k}+6D_{f,k}/C_{\alpha,\gamma},1)}, \frac{C_{\alpha,\gamma}}{10C_{f,k}+5\gamma C_k}\right)$. Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, $\mathcal{W}_0 := f(x_0) - f(x_{\min})$, and for $i \geq 0$, let*

$$\mathcal{W}_{i+1} = \mathcal{W}_i \left(1 + \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1 - \gamma - 2\epsilon(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\right]\alpha(2\mathcal{W}_i)\right)^{-1}.$$

Then for any (x_0, p_0) with $p_0 = 0$, the iterates (4.36) satisfy for every $i \geq 0$,

$$f(x_i) - f(x_{\min}) \leq 2\mathcal{W}_i \leq 2\mathcal{W}_0 \left(1 + \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1 - \gamma - 2\epsilon(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\right]\alpha_\star\right)^{-i}.$$

Remark 4.4.7. Similar to Remark 4.4.4, Proposition 4.4.6 implies that, under suitable assumptions and position independent step sizes, the first explicit method can achieve linear convergence with contraction rate that is proportional to $\alpha(3\mathcal{H}_0)$ initially and possibly increasing as we get closer to the optimum. In particular, again as remarked in Remark 4.4.4, for $f(x)$ that behave like $\|x - x_{\min}\|_2^b$ near their minima and $\|x - x_{\min}\|_2^B$ in the tails the conditions of assumptions C can be satisfied for kinetic energies that grow like $\|p\|_2^a$ in the body and $\|p\|_2^A$ in the tails as long as $\frac{1}{a} + \frac{1}{b} = 1$, $\frac{1}{A} + \frac{1}{B} \geq 1$. The distinction here is that for the first explicit method we will require $b, B \geq 2$.

4.4.3 Second Explicit Method, with Analysis via the Hessian of k

Our second explicit method inverts relationship between f and k from the first. Again, it makes a fixed ϵ step approximation $\frac{x_{i+1}-x_i}{\epsilon} = x_t'$ and $\frac{p_{i+1}-p_i}{\epsilon} = p_t'$. In contrast to the implicit (4.33) and first explicit (4.36) methods, it approximates the field at the point (x_{i+1}, p_i).

Method 3 (Second Explicit Method)**.** *Given* $f, k : \mathbb{R}^d \to \mathbb{R}$, $\epsilon, \gamma \in (0, \infty)$, $x_0, p_0 \in \mathbb{R}^d$. *Let,*

$$\begin{aligned} x_{i+1} &= x_i + \epsilon \nabla k(p_i) \\ p_{i+1} &= (1 - \epsilon\gamma)p_i - \epsilon \nabla f(x_{i+1}). \end{aligned} \tag{4.39}$$

This discretization exploits the convexity of f by approximating the continuous dynamics at the forward point x_{i+1}, but is made explicit by approximating at the backward point p_i. As with the other explicit method, it requires a smoothness assumption to prevent k from changing too rapidly between iterates, which is expressed as a requirement on the interaction between ∇f and $\nabla^2 k$, see assumption D.5. These assumptions can be satisfied for k that have quadratic or higher power growth and are suitable for f that may have unbounded second derivatives at their minima (for such f, Assumptions C can not hold).

Assumption D. *D.1* $k : \mathbb{R}^d \to \mathbb{R}$ *strictly convex with minimum* $k(0) = 0$ *and twice continuously differentiable for every* $p \in \mathbb{R}^d \setminus \{0\}$.

D.2 There exists $C_k \in (0, \infty)$ *such that for every* $p \in \mathbb{R}^d$,

$$\langle \nabla k(p), p \rangle \leq C_k k(p). \tag{4.40}$$

D.3 There exists $D_k \in (0, \infty)$ *such that for every* $p \in \mathbb{R}^d \setminus \{0\}$,

$$\left\langle p, \nabla^2 k(p) p \right\rangle \leq D_k k(p). \tag{4.41}$$

D.4 There exists $E_k, F_k \in (0, \infty)$ *such that for every* $p, q \in \mathbb{R}^d$,

$$k(p) - k(q) \leq E_k k(q) + F_k \left\langle \nabla k(p) - \nabla k(q), p - q \right\rangle. \tag{4.42}$$

D.5 There exists $D_{f,k} \in (0, \infty)$ *such that for every* $x \in \mathbb{R}^d$, $p \in \mathbb{R}^d \setminus \{0\}$,

$$\left\langle \nabla f(x), \nabla^2 k(p) \nabla f(x) \right\rangle \leq D_{f,k} \alpha(3\mathcal{H}(x,p)) \mathcal{H}(x,p). \tag{4.43}$$

Remark 4.4.8. Smoothness of f implies $\frac{1}{2} \|\nabla f(x)\|_2^2 \leq L(f(x) - f(x_{\min}))$ (see (2.1.7) of Theorem 2.1.5 of [193]). Thus, if f is smooth and $k(p) = \frac{1}{2} \|p\|_2^2$, then the assumption D.5 can be satisfied by $D_{f,k} = \max\{1, 2L\}$, since $\nabla^2 k(p) = I$ and

$$\left\langle \nabla f(x), \nabla^2 k(p) \nabla f(x) \right\rangle = \|\nabla f(x)\|_2^2 \leq 2L(f(x) - f(x_{\min})) \leq 2L(f(x) - f(x_{\min})) + k(p).$$

The k-specific assumptions D.2 and D.3 can clearly be satisfied with $C_k = D_k = 2$ in this case. We show that D.4 can be satisfied in Section 4.5.

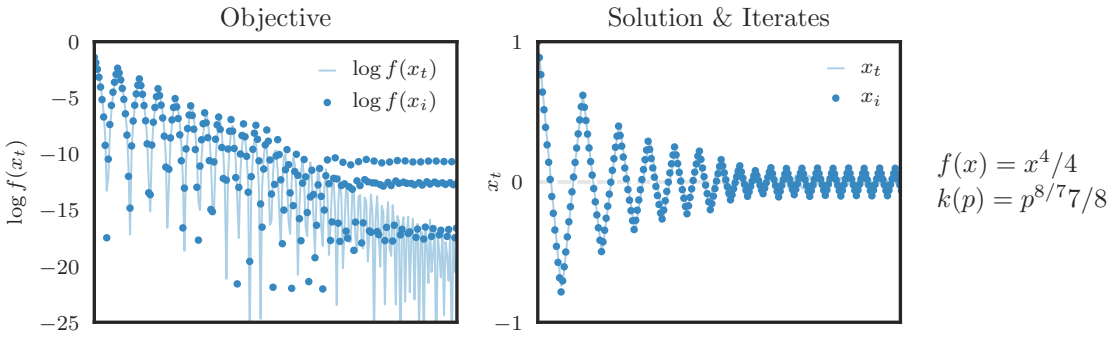


Figure 4.6: Importance of discretization assumptions. Solutions x_t and iterates x_i of our first explicit method on $f(x) = x^4/4$. With an inappropriate choice of kinetic energy, $k(p) = p^{8/7}7/8$, the continuous solution converges at a linear rate but the iterates do not.

Proposition 4.4.9 (Convergence bound for the second explicit scheme)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$, $C_{f,k}$, C_k, D_k, $D_{f,k}$, E_k, F_k satisfying assumptions A, B, and D, and that*

$$0 < \epsilon < \min\left(\frac{1-\gamma}{2(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})}, \frac{1-\gamma}{8D_k(1+E_k)}, \frac{C_{\alpha,\gamma}}{6(5C_{f,k} + 2\gamma C_k) + 12\gamma C_{\alpha,\gamma}}, \sqrt{\frac{1}{6\gamma^2 D_k F_k}}\right).$$

Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, $\mathcal{W}_0 := f(x_0) - f(x_{\min})$, and for $i \geq 0$, let

$$\mathcal{W}_{i+1} = \mathcal{W}_i\left(1 - \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1 - \gamma - 2\epsilon(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\right]\alpha(2\mathcal{W}_i)\right).$$

Then for any (x_0, p_0) with $p_0 = 0$, the iterates (4.39) satisfy for every $i \geq 0$,

$$f(x_i) - f(x_{\min}) \leq 2\mathcal{W}_i \leq 2\mathcal{W}_0 \cdot \left(1 - \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1 - \gamma - 2\epsilon(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\right]\alpha_\star\right)^i.$$

Remark 4.4.10. Similar to Remark 4.4.4, Proposition 4.4.9 implies that, under suitable assumptions and for a fixed step size independent of the initial point, the second explicit method can achieve linear convergence with contraction rate that is proportional to $\alpha(3\mathcal{H}_0)$ initially and possibly increasing as we get closer to the optimum. In particular, again as remarked in Remark 4.4.4, for $f(x)$ that behave like $\|x - x_{\min}\|_2^b$ near their minima and $\|x - x_{\min}\|_2^B$ in the tails the conditions of assumptions D can be satisfied for kinetic energies that grow like $\|p\|_2^a$ in the body and $\|p\|_2^A$ in the tails as long as $\frac{1}{a} + \frac{1}{b} = 1$, $\frac{1}{A} + \frac{1}{B} \geq 1$. The distinction here is that for the second explicit method we will require $b, B \leq 2$.

To conclude the analysis of our methods on convex functions, consider the example $f(x) = x^4/4$ from Figure 4.4. If we take $k(p) = |p|^a/a$, then assumption A.4 requires that $a \leq 4/3$. Assumptions B and C cannot be satisfied as long as $a < 4/3$, which suggests that $k(p) = f^*(p)$ is the only suitable choice in this case. Indeed, in Figure 4.6, we see that the choice of $k(p) = p^{8/7}7/8$ results in a system whose continuous dynamics converge at a linear rate and whose discrete dynamics fail to converge. Note that as the continuous systems converge the oscillation frequency increases dramatically, making it difficult for a fixed step size scheme to approximate.

4.4.4 First Explicit Method on Non-Convex f

We close this section with a brief analysis of the convergence of the first explicit method on non-convex f. A traditional requirement of discretizations is some degree of smoothness to prevent the function changing too rapidly between points of approximation. The notion of Lipschitz smoothness is the standard one, but the use of the kinetic map ∇k to select iterates allows Hamiltonian descent methods to consider the broader definition of uniform smoothness, as discussed in [277, 13, 278] but specialized here for our purposes.

Uniform smoothness is defined by a norm $\|\cdot\|$ and a convex non-decreasing function $\sigma : [0, \infty) \to [0, \infty]$ such that $\sigma(0) = 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is σ-uniformly smooth, if for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \sigma(\|y - x\|). \tag{4.44}$$

Lipschitz smoothness corresponds to $\sigma(t) = \frac{1}{2}t^2$, and generally speaking there exist non-trivial uniformly smooth functions for $\sigma(t) = \frac{1}{b}t^b$ for $1 < b \leq 2$, see, e.g., [192, 277, 13, 278].

Assumption E. *E.1* $f : \mathbb{R}^d \to \mathbb{R}$ *differentiable.*

E.2 $\gamma \in (0, \infty)$.

E.3 There exists a norm $\|\cdot\|$ *on* $\mathbb{R}^d$, $b \in (1, \infty)$, $D_k \in (0, \infty)$, $D_f \in (0, \infty)$, $\sigma : [0, \infty) \to [0, \infty]$ *non-decreasing convex such that* $\sigma(0) = 0$ *and* $\sigma(ct) \leq c^b\sigma(t)$ *for* $c, t \in (0, \infty)$*; for all* $p \in \mathbb{R}^d$,

$$\sigma(\|\nabla k(p)\|) \leq D_k k(p); \tag{4.45}$$

and for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(y) + \langle \nabla f(x), y - x \rangle + D_f\sigma(\|y - x\|). \tag{4.46}$$

Lemma 4.4.11 (Convergence of the first explicit scheme without convexity)**.** *Given* $\|\cdot\|$, f, k, γ, b, D_k, D_f, σ *satisfying assumptions E and A.2. If* $\epsilon \in (0, \sqrt[b-1]{\gamma/D_f D_k}]$, *then the iterates* (4.36) *of the first explicit method satisfy*

$$\mathcal{H}_{i+1} - \mathcal{H}_i \leq (\epsilon^b D_f D_k - \epsilon\gamma)k(p_{i+1}) \leq 0, \tag{4.47}$$

and $\|\nabla f(x_i)\|_2 \to 0$.

Remark 4.4.12. L-Lipschitz continuity of the gradients $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for $L > 0$ with Euclidean norm $\|\cdot\|_2$ implies both $f(y) \leq f(y) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2$ and $\frac{1}{2}\|\nabla f(x)\|_2^2 \leq L(f(x) - f(x_{\min}))$. Thus, if f, k are L_f, L_k smooth, respectively, then the condition for convergence simplifies to $\epsilon \leq \gamma/L_f L_k$.

4.5 Kinetic Maps for Functions with Power Behavior

In this section we design a family of kinetic maps ∇k suitable for a class of functions f that exhibit power growth, which we will describe precisely as a set of assumptions. This class includes strongly convex and smooth functions. However, it is much broader, including functions with possibly non-quadratic power behavior and singular or unbounded Hessians. First, we show that this family of kinetic energies satisfies the k-specific assumptions of Section 4.4. Then we use the generic analysis of Section 4.4 to provide a specific set of assumptions on fs and their match to the choice of k. As a consequence, this analysis greatly extends the class of functions for which linear convergence is possible with fixed step size first order computation. Still, this analysis is not meant to be an exhaustive catalogue of possible kinetic energies for Hamiltonian descent. Instead, it serves as an example of how known properties of f can be used to design k. Note that, with a few exceptions, the proofs of all of our results in this section are deferred to Section Ap.4 of the Appendix.

4.5.1 Power Kinetic Energies

We assume a given norm $\|x\|$ and its dual $\|p\|_* = \sup\{\langle x, p\rangle : \|x\| \leq 1\}$ for $x, p \in \mathbb{R}^d$. Define the family of power kinetic energies k,

$$k(p) = \varphi_a^A(\|p\|_*) \text{ where } \varphi_a^A(t) = \tfrac{1}{A}\left(t^a + 1\right)^{\frac{A}{a}} - \tfrac{1}{A} \text{ for } t \in [0, \infty) \text{ and } a, A \in [1, \infty). \tag{4.48}$$

For $a = A$ we recover the standard power functions, $\varphi_a^a(t) = t^a/a$. For distinct $a \neq A$, we have $(\varphi_a^A)'(t) \sim t^{A-1}$ for large t and $(\varphi_a^A)'(t) \sim t^{a-1}$ for small t. Thus,

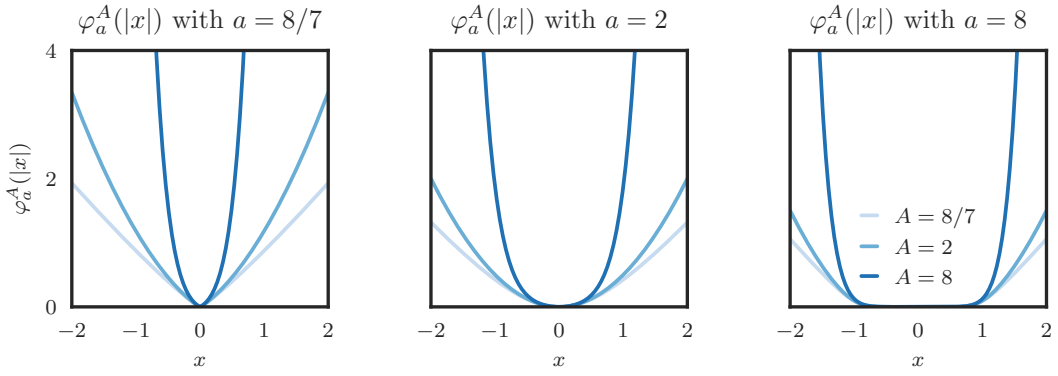


Figure 4.7: Power kinetic energies in one dimension.

$k(p) \sim \|p\|_*^A / A$ as $\|p\|_* \uparrow \infty$ and $k(p) \sim \|p\|_*^a / a$ as $\|p\|_* \downarrow 0$. See Figure 4.7 for examples from this family in one dimension.

Broadly speaking, this family of kinetic energies must be matched in a conjugate fashion to the body and tail behavior of f. Informally, for this choice of k we will require conditions on f that correspond to requiring that it grows like $\|x - x_{\min}\|^b$ in the body (as $\|x - x_{\min}\| \to 0$) and $\|x - x_{\min}\|^B$ in the tails (as $\|x - x_{\min}\| \to \infty$) for some $b, B \in (1, \infty)$. In particular, our growth conditions in the case of f "growing like" $\|x\|_2^2 = \langle x, x \rangle$ everywhere will be necessary conditions of strong convexity and smoothness. More generally, a, A, b, B will be well-matched if $1/a + 1/b = 1/A + 1/B = 1$, but other scenarios are possible. Of these, the conjugate relationship between a and b is the most critical; it captures the asymptotic match between f and k as $(x_i, p_i) \to (x_{\min}, 0)$, and our analysis requires that $1/a + 1/b = 1$. The match between A and B is less critical. In the ideal case, B is known and $A = B/(B-1)$. In this case, the discretizations will converge at a constant fast linear rate. If B is not known, it suffices for $1/A + 1/B \geq 1$. The consequence of underestimating $A < B/(B-1)$ will be reflected in a linear, but non-constant, rate of convergence (via α of Assumption A.4), which depends on the initial x_0 and slowly improves towards a fast rate as the system converges and the regime switches. We present a complete analysis and set of conditions on f for two of the most useful scenarios. In Proposition 4.5.8 we consider the case that f grows like $\varphi_b^B(\|x - x_{\min}\|)$ where $b, B > 1$ are exactly known. In this case convergence proceeds at a fast constant linear rate when matched with $k(p) = \varphi_a^A(\|p\|_*)$ where $a = b/(b-1)$ and $A = B/(B-1)$. In Proposition 4.5.10 we consider the case that f grows like $\varphi_2^B(\|x - x_{\min}\|)$ where $B \geq 2$ is unknown. Here, the convergence is linear with a non-constant rate when matched with the relativistic kinetic energy $k(p) = \varphi_2^1(\|p\|_*)$. The case covered by relativistic kinetic

	$f(x)$ grows like $\varphi_b^B(\|x\|)$			appropriate $k(p) = \varphi_a^A(\|p\|_*)$	
method	powers known?	body power b	tail power B	body power a	tail power A
implicit	known	$b > 1$	$B > 1$	$a = b/(b-1)$	$A = B/(B-1)$
	unknown	$b = 2$	$B \geq 2$	$a = 2$	$A = 1$
1st explicit	known	$b \geq 2$	$B \geq 2$	$a = b/(b-1)$	$A = B/(B-1)$
	unknown	$b = 2$	$B \geq 2$	$a = 2$	$A = 1$
2nd explicit	known	$1 < b \leq 2$	$1 < B \leq 2$	$a = b/(b-1)$	$A = B/(B-1)$

Table 4.1: A summary of the conditions on f and power kinetic k considered in this section that satisfy the assumptions of Section 4.4. Here "grows like" is an imprecise term meaning that fs growth can be bounded in an appropriate way by $\varphi_b^B(\|x\|)$ (φ_b^B is defined in (4.48)). The full precise assumptions on f are laid out in Propositions 4.5.8 and 4.5.10. In particular, $b = B = 2$ corresponds to assumptions similar in spirit to strong convexity and smoothness. Other combinations of b, B and a, A are possible.

∇k is particularly valuable, as it covers a large class of globally non-smooth, but strongly convex functions. Table 4.1 summarizes this, and throughout the remaining subsections we flesh out the details of these claims.

For these kinetic energies to be suitable in our analysis, they must at minimum satisfy assumptions A.2, C.1, D.1, D.3, and D.4. Assumptions C.1 and D.3 are clearly satisfied by $k(p) = |p|^a/a$ for $p \in \mathbb{R}$ with constants $C_k = a$ and $D_k = a(a-1)$. In the remainder of this subsection, we provide conditions on the norms and a, A under which assumptions like these hold for φ_a^A with multiple power behavior in any finite dimension.

In general, the problematic terms of $\nabla k(p)$ and $\nabla^2 k(p)$ that arise in high dimensions involve the gradient and Hessian of the norm. The gradient of norm can be dealt with cleanly, but our analysis requires additional control on the Hessian of the norm. To control terms involving $\nabla^2 \|p\|_*$ we define a generalization of the maximum eigenvalue induced by the norm $\|\cdot\|$. Let $\lambda_{\max}^{\|\cdot\|} : \mathbb{R}^{d\times d} \to \mathbb{R}$ be the function defined by

$$\lambda_{\max}^{\|\cdot\|}(M) = \sup\{\langle v, Mv\rangle : v \in \mathbb{R}^d, \|v\| = 1\}. \tag{4.49}$$

For symmetric $M \in \mathbb{R}^{d\times d}$ and Euclidean $\|\cdot\|$ this is exactly the maximum eigenvalue of M. Now we are able to state our lemma analyzing power kinetic energies.

Lemma 4.5.1 (Verifying assumptions on k)**.** *Given a norm $\|p\|_*$ on $p \in \mathbb{R}^d$, $a, A \in [1, \infty)$, and φ_a^A in (4.48). Define the constant,*

$$C_{a,A} = \left(1 - \left(\tfrac{a-1}{A-1}\right)^{\frac{a-1}{A-a}} + \left(\tfrac{a-1}{A-1}\right)^{\frac{A-1}{A-a}}\right)^{\frac{B-b}{b}}. \tag{4.50}$$

$k(p) = \varphi_a^A(\|p\|_*)$ satisfies the following.

1. *Convexity. If $a > 1$ or $A > 1$, then k is strictly convex with a unique minimum at $0 \in \mathbb{R}^d$.*

2. *Conjugate. For all $x \in \mathbb{R}^d$, $k^*(x) = (\varphi_a^A)^*(\|x\|)$.*

3. *Gradient. If $\|p\|_*$ is differentiable at $p \in \mathbb{R}^d \setminus \{0\}$ and $a > 1$, then k is differentiable for all $p \in \mathbb{R}^d$, and for all $p \in \mathbb{R}^d$,*

$$\langle \nabla k(p), p \rangle \leq \max\{a, A\} k(p), \tag{4.51}$$
$$(\varphi_a^A)^*(\|\nabla k(p)\|) \leq (\max\{a, A\} - 1) k(p). \tag{4.52}$$

   Additionally, if $a, A > 1$, define $B = A/(A-1)$, $b = a/(a-1)$, and then

$$\varphi_b^B(\|\nabla k(p)\|) \leq C_{a,A}(\max\{a, A\} - 1) k(p). \tag{4.53}$$

   Additionally, if $a, A \geq 2$, then for all $p, q \in \mathbb{R}^d$,

$$k(p) \leq \langle \nabla k(q), q \rangle + \langle \nabla k(p) - \nabla k(q), p - q \rangle. \tag{4.54}$$

4. *Hessian. If $\|p\|_*$ is twice continuously differentiable at $p \in \mathbb{R}^d \setminus \{0\}$, then k is twice continuously differentiable for all $p \in \mathbb{R}^d \setminus \{0\}$, and for all $p \in \mathbb{R}^d \setminus \{0\}$,*

$$\left\langle p, \nabla^2 k(p) p \right\rangle \leq \max\{a, A\}(\max\{a, A\} - 1) k(p). \tag{4.55}$$

   Additionally, if $a, A \geq 2$ and there exists $N \in [0, \infty)$ such that $\|p\|_* \lambda_{\max}^{\|\cdot\|_*}(\nabla^2 \|p\|_*) \leq N$ for $p \in \mathbb{R}^d \setminus \{0\}$, then for all $p \in \mathbb{R}^d \setminus \{0\}$

$$(\varphi_{a/2}^{A/2})^* \left(\frac{\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 k(p))}{\max\{a, A\} - 1 + N} \right) \leq (\max\{a, A\} - 2) k(p). \tag{4.56}$$

Remark 4.5.2. (4.51), (4.54), and (4.55) together directly confirm that these k satisfy C.1, D.3, and D.4 with constants $C_k = \max\{a, A\}$, $D_k = \max\{a, A\}(\max\{a, A\} - 1)$, $E_k = \max\{a, A\} - 1$, and $F_k = 1$. The other results (4.52), (4.53), and (4.56) will be used in subsequent lemmas along with assumptions on f to satisfy the remaining assumptions of discretization.

The assumption that $\|p\|_* \lambda_{\max}^{\|\cdot\|_*}(\nabla^2 \|p\|_*) \leq N$ in Lemma 4.5.1 is satisfied by b-norms for $b \in [2, \infty)$, as the following lemma confirms. It implies that if $\|p\|_* = \|p\|_b$ for $b \geq 2$, we can take $N = b - 1$ in (4.56).

Lemma 4.5.3 (Bounds on $\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 \|p\|_*)$ for b-norms)**.** *Given $b \in [2, \infty)$, let $\|x\|_b = \left(\sum_{n=1}^d |x^{(n)}|^b\right)^{1/b}$ for $x \in \mathbb{R}^d$. Then for $x \in \mathbb{R}^d \setminus \{0\}$,*

$$\|x\|_b \, \lambda_{\max}^{\|\cdot\|_b}\big(\nabla^2 \|x\|_b\big) \leq (b-1).$$

The remaining assumptions B.1, C.3, and D.5 involve inner products between derivatives of f and k. To control these terms we will use the Fenchel-Young inequality. To this end, the conjugates of φ_a^A will be a crucial component of our analysis.

Lemma 4.5.4 (Convex conjugates of φ_a^A)**.** *Given $a, A \in (1, \infty)$ and φ_a^A in (4.48). Define $B = A/(A-1)$, $b = a/(a-1)$. The following hold.*

1. *Near Conjugate. φ_b^B upper bounds the conjugate $(\varphi_a^A)^*$ for all $t \in [0, \infty)$,*

$$(\varphi_a^A)^*(t) \leq \varphi_b^B(t). \tag{4.57}$$

2. *Conjugate. For all $t \in [0, \infty)$,*

$$(\varphi_a^a)^*(t) = \varphi_b^b(t). \tag{4.58}$$

$$(\varphi_1^A)^*(t) = \begin{cases} 0 & t \in [0,1] \\ \frac{1}{B} t^B - t + \frac{1}{A} & t \in (1, \infty) \end{cases}. \tag{4.59}$$

$$(\varphi_a^1)^*(t) = \begin{cases} 1 - (1 - t^b)^{\frac{1}{b}} & t \in [0,1] \\ \infty & t \in (1, \infty) \end{cases}. \tag{4.60}$$

$$(\varphi_1^1)^*(t) = \begin{cases} 0 & t \in [0,1] \\ \infty & t \in (1, \infty) \end{cases}. \tag{4.61}$$

4.5.2 Matching power kinetic ∇k with assumptions on f

In this subsection and the next we study assumptions on f that imply the suitability of $k(p) = \varphi_a^A(\|p\|_*)$ with the discretizations of Section 4.4. The preceding subsection is an analysis that verifies that such k satisfy the k-specific assumptions A, C, and D. We now consider the remaining assumptions of A, B, C, and D, which require an appropriate match between f and k. This includes the derivation of α and control of terms of the form $\langle \nabla f(x), \nabla k(p) \rangle$ and $\langle \nabla k(p), \nabla^2 f(x) \nabla k(p) \rangle$ by the total energy

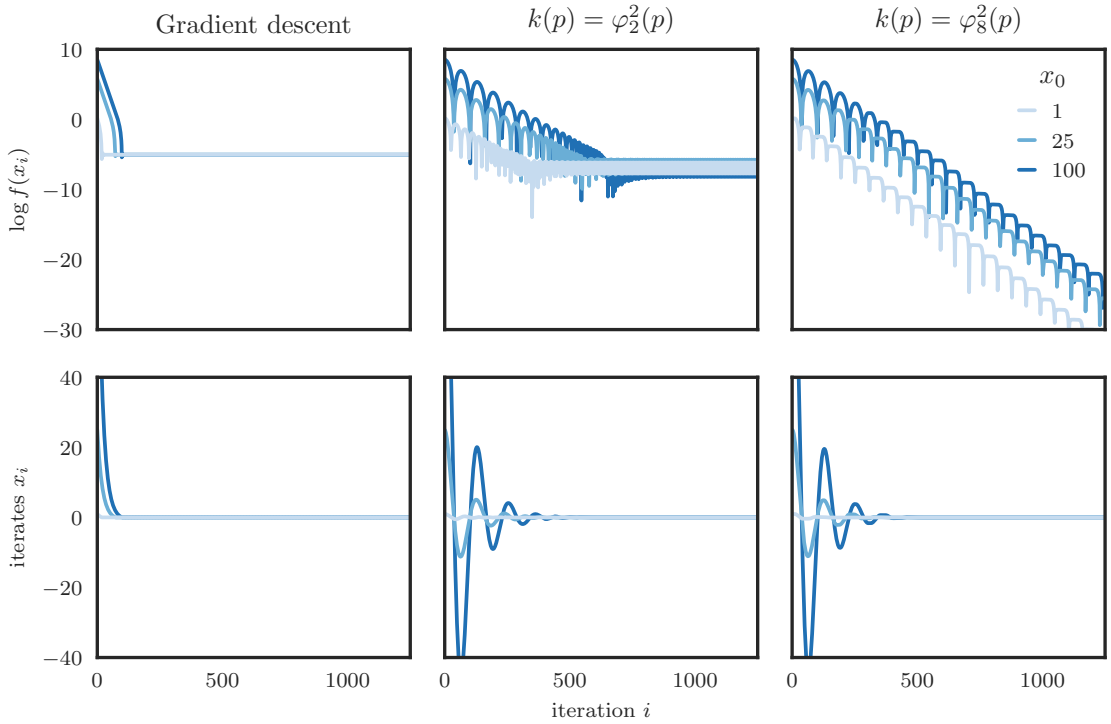


Figure 4.8: Optimizing $f(x) = \varphi_{8/7}^2(x)$ with three different methods with fixed step size: gradient descent, classical momentum, and our second explicit method. Because the second derivative of f is infinite at its minimum, only second explicit method with $k(p) = \varphi_8^2(p)$ is able to converge with a fixed step size.

$\mathcal{H}(x, p)$. Here we consider the case that f exhibits power behavior with known, but possibly distinct, powers in the body and the tails.

To see a complete example of this type of analysis, take the case $f(x) = |x|^b/b$ and $k(p) = |p|^a/a$ with $x, p \in \mathbb{R}$, $b > 2$, $a < 2$, and $1/a + 1/b = 1$. For α the strategy will be to find a lower bound on f that is symmetric about f's minimum. The conjugate of the centered lower bound can be used to construct an upper bound on f_c^* with which the gap between k and f_c^* can be studied. In this case it is simple, as we have $f_c^*(p) = k(p)$ and $\alpha = 1$. The strategy for terms of the form $\langle \nabla f(x), \nabla k(p) \rangle$ and $\langle \nabla k(p), \nabla^2 f(x) \nabla k(p) \rangle$ will be a careful application of the Fenchel-Young inequality. Using $a - 1 = a/b$, the conjugacy of b and $b/(b-1)$, and the Fenchel-Young inequality,

$$|\langle \nabla f(x), \nabla k(p) \rangle| = |x|^{b-1}|p|^{a/b} \leq \tfrac{b-1}{b}(|x|^{b-1})^{\frac{b}{b-1}} + \tfrac{1}{b}(|p|^{a/b})^b = (b-1)f(x) + (a-1)k(p)$$
$$\leq (\max\{a, b\} - 1)\mathcal{H}(x, p).$$

Finally, using the conjugacy of $b/2$ and $b/(b-2)$ and again the Fenchel-Young

inequality,

$$\begin{aligned}\left\langle \nabla k(p), \nabla^2 f(x) \nabla k(p) \right\rangle = (b-1)|x|^{b-2}|p|^{2a/b} &\leq \tfrac{(b-1)(b-2)}{b}(|x|^{b-2})^{\frac{b}{b-2}} + \tfrac{(b-1)2}{b}(|p|^{2a/b})^{\frac{b}{2}} \\ &= (b-1)(b-2)f(x) + 2(b-1)(a-1)k(p) \\ &\leq (b-1)\max\{b-2, 2(a-1)\}\mathcal{H}(x,p).\end{aligned}$$

Along with Lemma 4.5.1, this covers Assumptions A, B, and C. Thus, we can justify the use of the first explicit method for this f, k. All of the analyses of this section essentially follow this outline.

Remark 4.5.5. These strategies apply naturally when f is twice differentiable and smooth. In this case, we have $\frac{1}{2}\|\nabla f(x)\|_2^2 \leq L(f(x) - f(x_{\min}))$ and $\lambda_{\max}^{\|\cdot\|_2}(\nabla^2 f(x)) \leq L$. Thus, using $k(p) = \frac{1}{2}\|p\|_2^2$ is appropriate and $\langle \nabla f(x), \nabla k(p) \rangle \leq \max\{L, 1\}\mathcal{H}(x,p)$ and $\langle \nabla k(p), \nabla^2 f(x) \nabla k(p) \rangle \leq 2Lk(p)$.

We are now ready to consider the case of f growing like $\varphi_b^B(\|x - x_{\min}\|)$ matched with $k(p) = \varphi_a^A(\|p\|_*)$ for $1/a + 1/b = 1/A + 1/B = 1$. Assumptions F, below, will be used in different combinations to confirm that the assumptions of the different discretizations are satisfied. Assumptions F.1, F.2, F.3, and F.4 are required for all methods. The explicit methods each require an additional assumption: F.5 for the first explicit method and F.6 for the second. Thus, for $f : \mathbb{R}^d \to \mathbb{R}$ and $k(p) = \varphi_2^1(\|p\|_*)$, Proposition 4.5.8 can be summarised as

$$\begin{aligned}\text{F.1} \wedge \text{F.2} \wedge \text{F.3} \wedge \text{F.4} &\Rightarrow \text{A} \wedge \text{B}, \\ \text{F.1} \wedge \text{F.2} \wedge \text{F.3} \wedge \text{F.4} \wedge \text{F.5} &\Rightarrow \text{A} \wedge \text{B} \wedge \text{C}, \\ \text{F.1} \wedge \text{F.2} \wedge \text{F.3} \wedge \text{F.4} \wedge \text{F.6} &\Rightarrow \text{A} \wedge \text{B} \wedge \text{D}.\end{aligned}$$

Note, that Lemma 4.5.1 implies that the power kinetic energies are themselves examples of functions satisfying Assumptions F. Figure 4.8 illustrates a consequence of this proposition; $f(x) = \varphi_{8/7}^2(x)$ for $x \in \mathbb{R}$ is a difficult function to optimize with a first order method using a fixed step size; the second derivative grows without bound as $x \to 0$. As shown, Hamiltonian descent with the matched $k(p) = \varphi_8^2(p)$ converges, while gradient descent and classical momentum do not.

Assumption F. *F.1* $f : \mathbb{R}^d \to \mathbb{R}$ *differentiable and convex with unique minimum* $x_{\min}$.

F.2 $\|p\|_*$ *is differentiable at* $p \in \mathbb{R}^d \setminus \{0\}$ *with dual norm* $\|x\| = \sup\{\langle x, p\rangle : \|p\|_* = 1\}$.

F.3 $B = A/(A-1)$, *and* $b = a/(a-1)$.

F.4 There exist $\mu, L \in (0,\infty)$ *such that for all* $x \in \mathbb{R}^d$

$$\begin{aligned} f(x) - f(x_{\min}) &\geq \mu\varphi_b^B(\|x - x_{\min}\|) \\ \varphi_a^A(\|\nabla f(x)\|_*) &\leq L(f(x) - f(x_{\min})). \end{aligned} \tag{4.62}$$

F.5 $b \geq 2$ *and* $B \geq 2$. $f : \mathbb{R}^d \to \mathbb{R}$ *is twice continuously differentiable for all* $x \in \mathbb{R}^d \backslash \{x_{\min}\}$ *and there exists* $L_f, D_f \in (0,\infty)$ *such that for all* $x \in \mathbb{R}^d \backslash \{x_{\min}\}$

$$\left(\varphi_{b/2}^{B/2}\right)^* \left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right) \leq D_f(f(x) - f(x_{\min})). \tag{4.63}$$

F.6 $b \leq 2$ *and* $B \leq 2$. $\|p\|_*$ *is twice continuously differentiable at* $p \in \mathbb{R}^d \setminus \{0\}$, *and there exists* $N \in (0,\infty)$ *such that* $\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 \|p\|_*) \leq N \|p\|_*^{-1}$ *for all* $p \in \mathbb{R}^d \setminus \{0\}$.

Remark 4.5.6. Assumption F.4 can be read as the requirement that f is bounded above and below by φ_b^B, and in the $b = B = 2$ case it is a necessary condition of strong convexity and smoothness.

Remark 4.5.7. Assumption F.5 generalizes a sufficient condition for smoothness. Consider for simplicity the Euclidean norm case $\|\cdot\| = \|\cdot\|_2$ and let $\lambda_{\max}(M)$ be the maximum eigenvalue of $M \in \mathbb{R}^{d\times d}$. If $b = B = 2$, then $(\varphi_{b/2}^{B/2})^*$ is finite only on $[0,1]$ where it is zero. Moreover, F.5 simplifies to there existing $L_f \in (0,\infty)$ such that $\lambda_{\max}(\nabla^2 f(x)) \leq L_f$ everywhere, the standard smoothness condition. When $b > 2, B = 2$, $(\varphi_{b/2}^{B/2})^*$ is finite on $[0,1]$ where it behaves like a power $b/(b-2)$ function for small arguments. Thus, F.5 can be satisfied in the Euclidean norm case by a function whose maximum eigenvalue is shrinking like $\|x - x_{\min}\|_2^{b-2}$ as $x \to x_{\min}$; the balance of where the behavior switches can be controlled by L_f. When $b = 2, B > 2$, the role is switched and F.5 can be satisfied by a function whose maximum eigenvalue is bounded near the minimum and grows like $\|x - x_{\min}\|_2^{B-2}$ as $\|x - x_{\min}\|_2 \to \infty$. When $b, B > 2$, this can be satisfied by a function whose maximum eigenvalue shrinks like $\|x - x_{\min}\|_2^{b-2}$ in the body and grows like $\|x - x_{\min}\|_2^{B-2}$ in the tail.

Proposition 4.5.8 (Verifying assumptions for f with known power behavior and appropriate k)**.** *Given a norm* $\|\cdot\|_*$ *satisfying F.2 and* $a, A \in (1,\infty)$, *take*

$$k(p) = \varphi_a^A(\|p\|_*).$$

with φ_a^A *defined in* (4.48). *The following cases hold with this choice of* k *on* $f : \mathbb{R}^d \to \mathbb{R}$ *convex.*

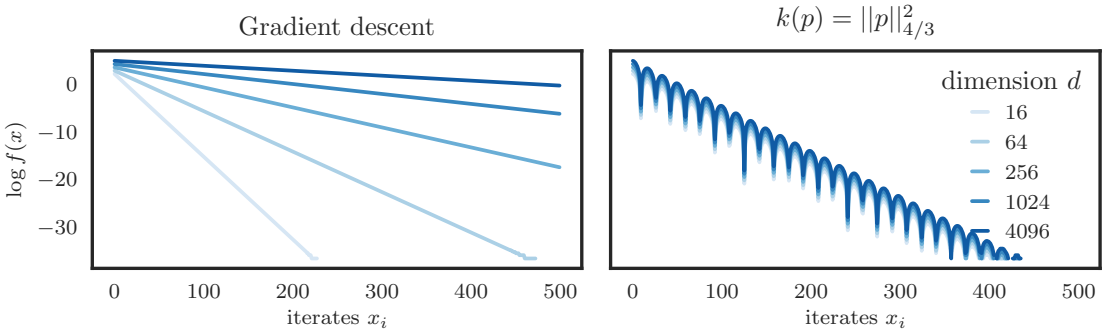


Figure 4.9: Dimension dependence on $f(x) = \|x\|_4^2/2$ initialized at $x_0 = (2, 2, \ldots, 2) \in \mathbb{R}^{\mathrm{d}}$. Left: Gradient descent with fixed step size equal to the inverse of the smoothness coefficient, $L = 3$. Right: Hamiltonian descent with $k(p) = \|p\|_{4/3}^2/2$ and a fixed step size (same for all dimensions). Gradient descent convergences linearly with rate $\lambda = 3/(3 - 1/\sqrt{d})$ while Hamiltonian descent converges with dimension-free linear rates.

1. *For the implicit method* (4.33)*, assumptions A, B hold with constants*

$$\alpha = \min\{\mu^{a-1}, \mu^{A-1}, 1\} \qquad C_{\alpha,\gamma} = \gamma \qquad C_{f,k} = \max\{a-1, A-1, L\}, \quad (4.64)$$

if $f, a, A, \mu, L, \|\cdot\|_*$ *satisfy assumptions F.1, F.2, F.3, F.4.*

2. *For the first explicit method* (4.36)*, assumptions A, B, and C hold with constants* (4.64) *and*

$$C_k = \max\{a, A\} \qquad D_{f,k} = L_f \alpha^{-1} \max\left\{D_f,\, 2C_{a,A}(\max\{a, A\} - 1)\right\}, \quad (4.65)$$

if $f, a, A, \mu, L, L_f, D_f, \|\cdot\|_*$ *satisfy assumptions F.1, F.2, F.3, F.4, and F.5.*

3. *For the second explicit method* (4.39)*, assumptions A, B, and D hold with constants* (4.64) *and*

$$\begin{aligned} C_k &= \max\{a, A\} \qquad D_k = \max\{a, A\}(\max\{a, A\} - 1) \\ E_k &= \max\{a, A\} - 1 \qquad F_k = 1 \\ D_{f,k} &= \alpha^{-1}(\max\{a, A\} - 1 + N) \max\left\{2L, a-2, A-2\right\}, \end{aligned} \quad (4.66)$$

if $f, a, A, \mu, L, N, \|\cdot\|_*$ *satisfy assumptions F.1, F.2, F.3, F.4, and F.6.*

We highlight an interesting consequence of Proposition 4.5.8 for high dimensional problems. For many first-order methods using standard gradients on smooth f, linear convergence can be guaranteed by the Polyak-Łojasiewicz (PL) inequality,

$\|\nabla f(x)\|_2^2/2 \geq \mu(f(x) - f(x_{\min}))$, see e.g., [133]. The rates of convergence generally depend on μ and the smoothness constant L. Unfortunately, for some functions the constant L or the constant μ may depend on the dimensionality of the space. Although smoothness and the PL inequality can be defined with respect to non-Euclidean norms, this does not generally overcome the issue of dimension dependence if standard gradients are used, see [131, 191] for a discussion and methods using non-standard gradients. The situation is distinct for Hamiltonian descent. If f is smooth with respect to a non-Euclidean norm $\|\cdot\|$, then, by taking $k(p) = \|p\|_*^2/2$, Proposition 4.5.8 may guarantee, under appropriate assumptions, dimension independent rates when using standard gradients (dependence on the dimensionality is mediated by the constant N). For example, consider $f(x) = \|x\|_4^2/2 = (\sum_{n=1}^d (x^{(n)})^4)^{1/2}/2$ defined for d-dimensional vectors $x \in \mathbb{R}^d$. It is possible to show that f is smooth with respect to $\|\cdot\|_2$ with constant $L = 3$ (our Lemma 4.5.3 together with an analysis analogous to Lemma 14 in Appendix A of [232] and the fact that $\|x\|_4 \leq \|x\|_2$) and that f satisfies the PL inequality with $\mu = 1/\sqrt{d}$ (the fact that $\|x\|_{4/3}^2 \leq \sqrt{d}\,\|x\|_2^2$ and Lemma 4.5.1). The iterates of a gradient descent algorithm with fixed step size $1/L$ on this f will therefore satisfy the following,

$$f(x_{i+1}) - f(x_i) \leq -\frac{1}{6}\|\nabla f(x_i)\|_2^2 \leq -\frac{1}{3\sqrt{d}} f(x_i).$$

From which we conclude that gradient descent converges linearly with rate $\lambda = 3/(3 - 1/\sqrt{d})$, worsening as $d \to \infty$. Figure 4.9 illustrates this, along with a comparison to Hamiltonian descent with $k(p) = \|p\|_{4/3}^2$, which enjoys dimension independence as $N = 3$.

4.5.3 Matching relativistic kinetic ∇k with assumptions on f

The strongest assumption of Proposition 4.5.8 is that the power behaviour of f captured in the constant b is exactly known. This is generally not the case and usually hard to determine. The only possible exception is $b = 2$, which can be guaranteed by lower bounding the eigenvalues of the Hessian. In our second analysis, we consider a kinetic energy generically suitable for such strongly convex functions that may not be smooth. Crucially, less information needs to be known about f for this kinetic energy to be applicable. The cost imposed by this lack of knowledge is a non-constant rate of linear convergence, which begins slowly and improves towards a faster rate as $(x_i, p_i) \to (x_{\min}, 0)$.

In particular, we consider the use of the relativistic kinetic energy,

$$k(p) = \varphi_2^1(\|p\|_*) = \sqrt{\|p\|_*^2 + 1} - 1,$$

which was studied by Lu *et al.* [157] and Livingstone *et al.* [149] in the context of Hamiltonian Monte Carlo. Consider for the moment the Euclidean norm case. In this case, we have

$$\nabla k(p) = \frac{p}{\sqrt{\|p\|_2^2 + 1}}.$$

As noted by Lu *et al.* [157], this kinetic map resembles the iterate updates of popular adaptive gradient methods [75, 274, 90, 135]. Because the iterate updates $x_{i+1} - x_i$ of Hamiltonian descent are proportional to $\nabla k(p)$ from some $p \in \mathbb{R}^d$, this suggests that the relativistic map may have favorable properties. Notice that $\|\nabla k(p)\|_2^2 = \|p\|_2^2 / (\|p\|_2^2 + 1) < 1$, implying that $\|x_{i+1} - x_i\|_2 < \epsilon$ uniformly and regardless of the magnitudes of ∇f. The fact that the magnitude of iterate updates is uniformly bounded makes the relativistic map suitable for functions with very fast growing tails, even if the rate of growth is not exactly known.

More precisely, we consider the case of f growing like $\varphi_2^B(\|x - x_{\min}\|)$ matched with $k(p) = \varphi_2^1(\|p\|_*)$ for $B \geq 2$. Assumptions G, below, will be used in different combinations to confirm that the assumptions of the different discretizations are satisfied. Assumptions G.1, G.2, G.3, and G.4 are required for all methods. The first explicit method requires additional assumptions G.5 for $B > 2$ and G.6 for $B = 2$. We do not include an analysis for the second explicit method. Thus, for $f : \mathbb{R}^d \to \mathbb{R}$ and $k(p) = \varphi_2^1(\|p\|_*)$, Proposition 4.5.10 can be summarised as

$$\text{G.1} \wedge \text{G.2} \wedge \text{G.3} \wedge \text{G.4} \Rightarrow \text{A} \wedge \text{B}$$
$$\text{G.1} \wedge \text{G.2} \wedge \text{G.3} \wedge \text{G.4} \wedge \text{G.5} \Rightarrow \text{A} \wedge \text{B} \wedge \text{C}$$
$$\text{G.1} \wedge \text{G.2} \wedge \text{G.3} \wedge \text{G.4} \wedge \text{G.6} \Rightarrow \text{A} \wedge \text{B} \wedge \text{C}$$

Figure 4.10 illustrates a consequence of this proposition; $f(x) = \varphi_8^2(x)$ for $x \in \mathbb{R}$ is a difficult function to optimize with a first order method using a fixed step size; the second derivative grows without bound as $|x| \to \infty$. Thus if the initial point is taken to be very large, gradient descent must take a very conservative choice of step size. As shown, Hamiltonian descent with the matched $k(p) = \varphi_{8/7}^2(p)$ converges quickly and uniformly, while gradient descent with a fixed step size suffers a very slow rate for $|x_0| \gg 0$. In the middle panel, the relativistic choice converges slowly at first, but speeds up as convergence proceeds, making it a suitable agnostic choice in cases such as this.

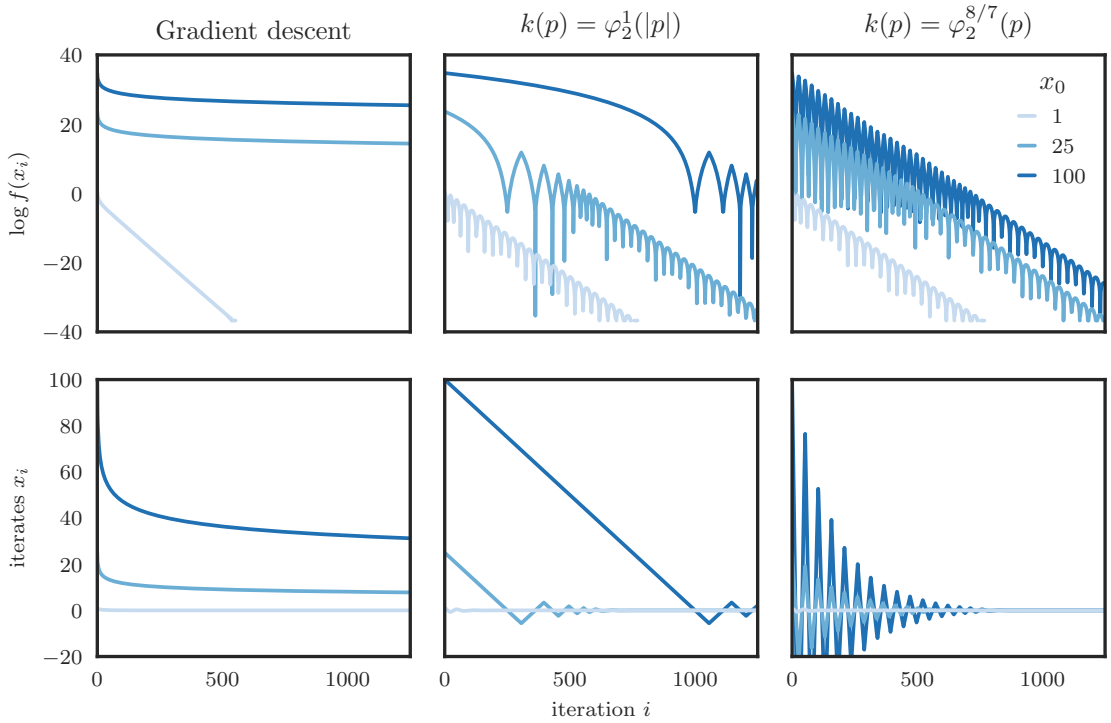


Figure 4.10: $f(x) = \varphi_2^8(x)$ with three different methods: gradient descent with the optimal fixed step size, Hamiltonian descent with relativistic kinetic energy, and Hamiltonian descent with the near dual kinetic energy.

Assumption G. *G.1* $f : \mathbb{R}^d \to \mathbb{R}$ *differentiable and convex with unique minimum* $x_{\min}$.

G.2 $\|p\|_*$ *is differentiable at* $p \in \mathbb{R}^d \setminus \{0\}$ *with dual norm* $\|x\| = \sup\{\langle x, p\rangle : \|p\|_* = 1\}$.

G.3 $B \in [2, \infty)$ *and* $A = B/(B-1)$.

G.4 There exist $\mu, L \in (0, \infty)$ *such that for all* $x \in \mathbb{R}^d$

$$\begin{aligned} f(x) - f(x_{\min}) &\geq \mu \varphi_2^B(\|x - x_{\min}\|) \\ \varphi_2^1(\|\nabla f(x)\|_*) &\leq L(f(x) - f(x_{\min})). \end{aligned} \tag{4.67}$$

G.5 $B > 2$. *Define*

$$\psi(t) = \begin{cases} 0 & 0 \leq t < 1 \\ t - 3t^{\frac{1}{3}} + 2 & 1 \leq t \end{cases}. \tag{4.68}$$

$f : \mathbb{R}^d \to \mathbb{R}$ *is twice continuously differentiable for all* $x \in \mathbb{R}^d \setminus \{x_{\min}\}$ *and there exists* $L_f \in (0, \infty)$ *such that for all* $x \in \mathbb{R}^d \setminus \{x_{\min}\}$

$$\psi\left(\tfrac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}\left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right)\right) \leq 3(f(x) - f(x_{\min})). \tag{4.69}$$

G.6 $B = 2$. $f : \mathbb{R}^d \to \mathbb{R}$ *is twice continuously differentiable for all* $x \in \mathbb{R}^d \setminus \{x_{\min}\}$ *and there exists* $L_f \in (0, \infty)$ *such that for all* $x \in \mathbb{R}^d \setminus \{x_{\min}\}$

$$\lambda_{\max}^{\|\cdot\|}\big(\nabla^2 f(x)\big) \leq L_f. \tag{4.70}$$

Remark 4.5.9. Assumptions G hold in general for convex functions f that grow quadratically at their minimum, and as power B in the tails, for some $B \geq 2$.

We include the proof of this proposition below, as it highlights every aspect of our analysis, including non-constant α.

Proposition 4.5.10 (Verifying assumptions for f with unknown power behavior and relativistic k)**.** *Given a norm* $\|\cdot\|_*$ *satisfying G.2, take*

$$k(p) = \varphi_2^1(\|p\|_*)$$

with φ_a^A *in* (4.48)*. The following cases hold with this choice of kinetic energy* k *on* $f : \mathbb{R}^d \to \mathbb{R}$ *convex.*

1. *For the implicit method* (4.33)*, assumptions A, B hold with constants*

$$C_{\alpha,\gamma} = \gamma \qquad C_{f,k} = \max\{1, L\}, \tag{4.71}$$

and α *non-constant, equal to*

$$\alpha(y) = \min\{\mu^{A-1}, \mu, 1\}(y+1)^{1-A}, \tag{4.72}$$

if $f, B, \mu, L, \|\cdot\|_*$ *satisfy assumptions G.1, G.2, G.3, G.4.*

2. *For the first explicit method* (4.36)*, assumptions A, B, and C hold with constants* (4.71)*,* α *equal to* (4.72)*, and*

$$C_k = 2 \qquad D_{f,k} = \frac{3L_f}{\min\{\mu^{A-1}, \mu, 1\}}, \tag{4.73}$$

if $f, B, \mu, L, L_f, \|\cdot\|_*$ *satisfy assumptions G.1, G.2, G.3, G.4, and G.5.*

3. *For the first explicit method* (4.36)*, assumptions A,B, and C hold with constants* (4.71)*,* α *equal to* (4.72)*, and*

$$C_k = 2 \qquad D_{f,k} = \frac{6L_f}{\min\{\mu, 1\}}, \tag{4.74}$$

if $f, B, \mu, L, L_f, \|\cdot\|_*$ *satisfy assumptions G.1,G.2, G.3, G.4, and G.6.*

Proof of Proposition 4.5.10. First, by Lemma 4.5.1, this choice of k satisfies assumptions A.2 and C.1 with constant $C_k = 2$. We consider the remaining assumptions of A, B, and C.

1. Our first goal is to derive α. By assumption G.4, we have $\mu\varphi_b^B(\|x\|) \leq f_c(x)$. Lemma Ap.4.2 in Appendix Ap.4 implies that $\varphi_2^A(\mu^{-1}t) \leq \max\{\mu^{-2}, \mu^{-A}\}\varphi_2^A(t)$ for $t \geq 0$. Since $(\mu\varphi_b^B(\|\cdot\|))^* = \mu(\varphi_b^B)^*(\mu^{-1}\|\cdot\|_*)$ by Lemma 4.5.1 and the results discussed in the review of convex analysis, we have by assumption G.3 and Lemma 4.5.4,

$$f_c^*(p) \leq \mu(\varphi_2^B)^*\left(\mu^{-1}\|p\|_*\right) \leq \mu\varphi_2^A\left(\mu^{-1}\|p\|_*\right) \leq \max\{\mu^{-1}, \mu^{1-A}\}\varphi_2^A(\|p\|_*).$$

Since $\varphi_2^A(0) = \varphi_2^1(0)$, any α satisfies (4.24) for $p = 0$. Assume $p \neq 0$. First, for $y \in [0, \infty)$, we have by rearrangement and convexity,

$$\varphi_2^A((\varphi_2^1)^{-1}(y)) = \tfrac{1}{A}(y+1)^A - \tfrac{1}{A} \leq y(y+1)^{A-1}.$$

Thus,

$$\frac{k(p)}{\varphi_2^A(\|p\|_*)} = \frac{k(p)}{\varphi_2^A((\varphi_2^1)^{-1}(k(p)))} = \frac{Ak(p)}{(k(p)+1)^A - 1} \geq (k(p)+1)^{1-A}.$$

From this we conclude

$$k(p) \geq (k(p)+1)^{1-A}\varphi_2^A(\|p\|_*) \geq \alpha(k(p))f_c^*(p).$$

Since k is symmetric, we have (4.24) of assumption A.4. To see that α satisfies the remaining conditions of assumption A.4, note that $(y+1)^{1-A}$ is convex and decreasing for $A > 1$; $(y+1)^{1-A}$ is non-negative and $\alpha(0) = \min\{\mu^{A-1}, \mu, 1\} \leq 1$. Finally, G.3 implies $1 < A \leq 2$, for which,

$$-\alpha'(y)y = \min\{\mu^{A-1}, \mu, 1\}(A-1)(y+1)^{-A}y < (A-1)\alpha(y) \leq \alpha(y). \tag{4.75}$$

So we can take $C_{\alpha,\gamma} = \gamma$ and α satisfies assumptions A. This implies that k satisfies assumptions A. Assumption G.1 is the same as assumption A.1, therefore f and k satisfy assumptions A.

Now by Fenchel-Young, the symmetry of norms, Lemma 4.5.1, and assumption G.4,

$$|\langle \nabla k(p), \nabla f(x)\rangle| \leq (\varphi_2^1)^*(\|\nabla k(p)\|) + \varphi_2^1(\|\nabla f(x)\|_*) \leq C_{f,k}\mathcal{H}(x,p),$$

where $C_{f,k} = \max\{1, L\}$ for assumptions B.

2. Assume $B > 2$, so that $A < 2$. The analysis of case 1. follows and therefore assumptions A and B hold along with the constants just derived. (4.75) implies

$$[\alpha(y)y]' = \alpha'(y)y + \alpha(y) = \alpha'(y)y + (2-A)\alpha(y) + (A-1)\alpha(y) \geq (2-A)\alpha(y).$$

Thus, $((y+1)^{2-A} - 1) \leq \alpha(y)y$. Since $2 - A = \frac{B-2}{B-1}$ and $\frac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}(y)$ is the inverse function of $(y+1)^{2-A} - 1$, it would be enough to show for $p \in \mathbb{R}^d$ and $x \in \mathbb{R}^d \setminus \{x_{\min}\}$ that

$$\tfrac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}\left(\frac{\|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right) \leq 3\mathcal{H}(x,p),$$

for assumptions C to hold with constant $D_{f,k} = 3L_f/\min\{\mu^{A-1}, \mu, 1\}$. First, for ψ in 4.68 note that for $t \in [0,1)$,

$$\psi^*(t) = 2(1-t)^{-\frac{1}{2}} - 2, \tag{4.76}$$

and that $\psi^*((\varphi_2^1)'(t)^2) = 2\varphi_2^1(t)$. Furthermore, by Lemma Ap.4.1 of Appendix Ap.4, we have that $\|\nabla k(p)\| = (\varphi_2^1)'(\|p\|_*) < 1$. Lemma Ap.4.2 in Appendix Ap.4 implies that $\varphi_1^{\frac{B-1}{B-2}}(\epsilon t) \leq \epsilon\varphi_1^{\frac{B-1}{B-2}}(t)$ for $\epsilon < 1$ and $t \geq 0$. All together this implies,

$$\begin{aligned}
&\tfrac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}\left(\frac{\|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right) \\
&\leq \|\nabla k(p)\|^2 \tfrac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}\left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right) \\
&\leq \psi^*(\|\nabla k(p)\|^2) + \psi\left(\tfrac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}\left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right)\right) \\
&\leq 2k(p) + \psi\left(\tfrac{B-1}{B-2}\varphi_1^{\frac{B-1}{B-2}}\left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f}\right)\right) \leq 3\mathcal{H}(x,p).
\end{aligned}$$

3. For $B = 2$, the analysis of case 1. follows and therefore assumptions A and B hold along with the constants just derived.. Here α is equal to

$$\alpha(y) = \frac{\min(\mu, 1)}{y + 1}. \tag{4.77}$$

Considering that $z/(1-z)$ is the inverse function of $y/(y+1)$ for $z \in [0, 1)$, it would be enough to show for $p \in \mathbb{R}^d$ and $x \in \mathbb{R}^d \setminus \{x_{\min}\}$ that

$$\frac{\|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{2L_f - \|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))} \leq 3\mathcal{H}(x, p),$$

for assumptions C to hold with constant $D_{f,k} = 6L_f / \min\{\mu, 1\}$. Indeed, taking ψ, ψ^* from (4.68) and (4.76), we have again $\psi^*((\varphi_2^1)'(t)^2) = 2\varphi_2^1(t)$. Again, by Lemma Ap.4.1 of Appendix Ap.4, we have that $\|\nabla k(p)\| = (\varphi_2^1)'(\|p\|_*) < 1$. Moreover $z/(2L - z) \leq 1$ for $z \leq L$. All together,

$$\begin{aligned}
\frac{\|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{2L_f - \|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))} &\leq \|\nabla k(p)\|^2 \frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{2L_f - \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))} \\
&\leq \psi^*(\|\nabla k(p)\|^2) + \psi\left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{2L_f - \lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}\right) \\
&\leq 2k(p) \leq 3\mathcal{H}(x, p).
\end{aligned}$$

□

4.6 Conclusion

The conditions of strong convexity and smoothness guarantee the linear convergence of most first-order methods. For a convex function f these conditions are essentially quadratic growth conditions. In this work, we introduced a family of methods, which require only first-order computation, yet extend the class of functions on which linear convergence is achievable. This class of functions is broad enough to capture non-quadratic power growth, and, in particular, functions f whose Hessians may be singular or unbounded. Although our analysis provides ranges for the step size and other parameters sufficient for linear convergence, it does not necessarily provide the optimal choices. It is a valuable open question to identify those choices.

The insight motivating these methods is that the first-order information of a second function, the kinetic energy k, can be used to incorporate global bounds on the convex conjugate f^* in a manner that achieves linear convergence on f. This opens

a series of theoretical questions about the computational complexity of optimization. Can meaningful lower bounds be derived when we assume access to the first order information of two functions f and k? Clearly, any meaningful answer would restrict k—otherwise the problem of minimizing f could be solved instantly by assuming first-order access to $k = f^*$ and evaluating $\nabla k(0) = \nabla f^*(0) = x_{\min}$. Exactly what that restriction would be is unclear, but a satisfactory answer would open yet more questions: is there a meaningful hierarchy of lower bounds when access is given to the first-order information of $N > 2$ functions? When access is given to the second-order information of $N > 1$ functions?

From an applied perspective, first-order methods are playing an increasingly important role in the era of large datasets and high-dimensional non-convex problems. In these contexts, it is often impractical for methods to require exact first-order information. Instead, it is frequently assumed that access is limited to unbiased estimators of derivative information. It is thus important to investigate the properties of the Hamiltonian descent methods described in this paper under such stochastic assumptions. For non-convex functions, the success of adaptive gradient methods, which bear a resemblance to our methods using a relativistic kinetic energy, suggests there may be gains from an exploration of other kinetic energies. Can kinetic energies be designed to condition Hamiltonian descent methods when the Hessian of f is not positive semi-definite everywhere and to encourage iterates to escape saddle points? Finally, the main limitation of the work presented herein is the requirement that a practitioner have knowledge about the behavior of f near its minimum. Therefore, it would be valuable to investigate adaptive methods that do not require such knowledge, but instead estimate it on-the-fly.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

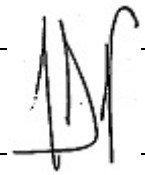
Title of Paper	Hamiltonian Descent Methods
Publication Status	□Published □ Accepted for Publication □Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style
Publication Details	Chris J. Maddison*, Daniel Paulin*, Yee Whye Teh, Arnaud Doucet. Hamiltonian Descent Methods. In *Preprint*, 2018.

Student Confirmation

Student Name:	Chris J. Maddison		
Contribution to the Paper	• I proposed the Hamiltonian descent differential equation and the use of the convex conjugate in the design of the kinetic energy. • Daniel Paulin and I collaborated jointly on the proofs of convergence in Hamiltonian descent. • I designed a Lyapunov function that proved the convergence of the continuous Hamiltonian descent ODE in the strongly convex case. • Daniel modified this Lyapunov function and proved the general convergence results presented in Sections 2 and 3 of the paper. • Daniel is completely responsible for the lower bounds proofs. • I proved the results of Section 4. • All authors contributed to the development of the paper through discussions and ideas, and all authors reviewed the final draft.		
Signature 		Date	05 May 2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Arnaud Doucet		
Supervisor comments I agree that the candidate has made a substantial contribution to the publication.		
Signature 	Date	05 May 2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Dual Space Preconditioning for Gradient Descent

5.1 Abstract

The conditions of relative smoothness and relative strong convexity were recently introduced for the analysis of Bregman gradient methods for convex optimization. In this chapter, we introduce a fully explicit descent scheme with relative smoothness in the dual space between the convex conjugate of the objective function and a designed dual reference function. For Legendre type convex functions under this dual relative smoothness, our scheme naturally remains in the interior of the domain, despite being fully explicit. We obtain linear convergence under dual relative strong convexity with a condition number that is invariant under horizontal translations. Our method is a non-linear preconditioning of gradient descent that can improve the conditioning of explicit first-order methods on problems with non-smooth or non-strongly convex structure. We show how this method can be applied to p-norm regression and exponential penalty function minimization.

5.2 Introduction

We study the minimization of a proper, lower semi-continuous (lsc), strictly convex, and differentiable function $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$,

$$\min_{x \in \operatorname{dom} f} f(x), \tag{P}$$

where $\operatorname{dom} f = \{x \in \mathbb{R}^d : f(x) < \infty\}$. Our primary focus is on functions f with Legendre structure: $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$ and $\|\nabla f(x_i)\| \to \infty$ for x_i converging to the boundary of $\operatorname{dom} f$. For such functions, a global minimizer $x_{\min}$, if it exists, is unique

Algorithm 2.1 Dual preconditioned gradient descent.

Given $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex, $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex with $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$ and $0 = \arg\min_{x^*} k(x^*)$, $x_0 \in \operatorname{int}(\operatorname{dom} f)$, and $L^* > 0$. For all $i \geq 0$,

$$x_{i+1} = x_i - \frac{1}{L^*} \nabla k(\nabla f(x_i)).$$

and in $\operatorname{int}(\operatorname{dom} f)$. We introduce an iterative first-order method (Algorithm 2.1) for (P). Iterative first-order methods produce a sequence of iterates $x_i \in \operatorname{int}(\operatorname{dom} f)$ converging to $x_{\min}$ using only the ability to compute $f(x)$ or the gradient vector $\nabla f(x)$ of first partial derivatives at any point $x \in \operatorname{int}(\operatorname{dom} f)$. Our method may be seen as a non-linear preconditioning of the classical gradient descent method. We show that the convergence of our method is guaranteed under a generalization of the standard Lipschitz continuity condition on ∇f, and develop two applications that show how this generalization can be used in practice.

In the analysis of first-order methods, it is standard to assume that the derivatives of f at some order are globally bounded by constants. For example, consider the classical gradient descent method, whose iterates satisfy

$$x_{i+1} = \operatorname*{arg\,min}_{x \in \operatorname{dom} f} \left\{ \langle \nabla f(x_i), x \rangle + \tfrac{L}{2} \|x - x_i\|^2 \right\}, \tag{5.1}$$

where $L > 0$ and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. A classical analysis shows that the iterates of gradient descent converge linearly in i, i.e., $f(x_i) - f(x_{\min}) = \mathcal{O}(\lambda^i)$ for $\lambda = 1 - \mu/L$, when f is assumed to be $\mu > 0$ strongly convex and ∇f is assumed to be L-Lipschitz continuous (traditionally called smoothness). Taken together for twice continuously differentiable f, these conditions are equivalent to the conditions that the eigenvalues of the Hessian matrix of second-order partial derivatives $\nabla^2 f(x)$ are everywhere lower bounded by the constant $\mu > 0$ (strong convexity) and upper bounded by the constant $L > 0$ (smoothness),

$$\mu I \preceq \nabla^2 f(x) \preceq LI \text{ for all } x \in \operatorname{int}(\operatorname{dom} f), \tag{5.2}$$

where $\preceq$ indicates the partial order of positive semi-definite matrices. Under these classical assumptions, closely matching lower and upper bounds are available on the number of gradient evaluations needed for a certain level of precision (see, e.g., [194]).

Analyses of first-order methods using only non-constant models of the derivatives of f have recently been discovered [16, 244, 156, 153]. In particular, [16] studied the following generalized gradient method that takes a designed Legendre convex reference

function $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ with $\operatorname{int}(\operatorname{dom} f) \subseteq \operatorname{int}(\operatorname{dom} h)$. Given $x_0 \in \operatorname{int}(\operatorname{dom} f)$, this method's iterates satisfy

$$x_{i+1} = \operatorname*{arg\,min}_{x \in \operatorname{dom} f} \{\langle \nabla f(x_i), x \rangle + L D_h(x, x_i)\} \tag{5.3}$$

where $L > 0$, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, and $D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$ for $x, y \in \operatorname{int}(\operatorname{dom} h)$. (5.3) is due to [189] and falls in a family of so-called Bregman proximal gradient methods. A standard analysis (see, e.g., [19]) of (5.3) makes the "absolute" assumptions that f is Lipschitz continuous and that h is strongly convex. In contrast, Bauschke, Bolte, and Teboulle [16] show that the following relative smoothness condition between f and h is sufficient for the convergence of (5.3),

$$\nabla^2 f(x) \preceq L \nabla^2 h(x) \text{ for all } x \in \operatorname{int}(\operatorname{dom} f). \tag{5.4}$$

It is possible for (5.4) to hold for f and h that are both non-smooth. For example, [16] study a Poisson inverse problem f whose derivatives of all orders are unbounded as $x \to 0$. They design an appropriate h, whose Hessian is also unbounded at 0, but which satisfies (5.4). [244, 156] extend the analysis of (5.3) to show that lower bounding the Hessian of $f(x)$ with the Hessian of $\mu h(x)$ for $\mu > 0$ (relative strong convexity) is sufficient for the linear convergence of $f(x_i) - f(x_{\min})$. To summarize, (5.4) generalizes smoothness and admits optimization methods for non-smooth differentiable f provided that the kind of non-smoothness can be captured by (5.4) and that (5.3) has a solution that can be efficiently computed.

In this paper we introduce a method (Algorithm 2.1) that exploits an application of relative smoothness in the dual space through a Legendre convex dual reference function $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ with $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$ and $0 = \arg\min_{x^*} k(x^*)$. The method is a generalization of gradient descent, in which the update direction of the iterates is preconditioned by ∇k. In section 5.5 we consider in detail the conditions under which we can provide convergence rates for Algorithm 2.1. The central condition is the existence of $L^* > 0$ such that

$$\nabla^2 f(x) \preceq L^* [\nabla^2 k(\nabla f(x))]^{-1} \text{ for all } x \in \operatorname{int}(\operatorname{dom} f). \tag{5.5}$$

Under this condition we show that along the iterates of Algorithm 2.1, $k(\nabla f(x_i)) - k(0)$ converges sub-linearly with rate $\mathcal{O}(i^{-1})$ (and thus $x_i \to x_{\min}$ for Legendre f). When the lower bound analog of (5.5) holds up to a constant factor $\mu^* > 0$, we show that $f(x_i) - f(x_{\min})$ converges linearly with rate $\lambda^* = 1 - \mu^*/L^*$. As we show in section

5.5.2, (5.5) is a relative smoothness condition in the dual space. It relates the growth of the second derivatives of f to the growth of the first derivatives of f, modulated by the choice of preconditioner ∇k. In contrast to the primal application of relative smoothness (5.4), the class of f satisfying (5.5) for a fixed k is closed under horizontal translations. With the exception of quadratic k or h, (5.4) and (5.5) are generally not equivalent conditions and $\mu \neq \mu^*$, $L \neq L^*$. Thus, the global information encoded in the dual reference function k is distinct from the information encoded in the reference function h. In section 5.6, we design ks and globally convergent methods for p-norm regression (see [48, 2] and references therein) and exponential penalty functions (see, e.g., [60, 59]).

5.3 Related literature

Dual preconditioned gradient descent requires of f only the ability to evaluate ∇f locally. The complexity of optimization under such assumptions is well-understood within the local oracle model of computation [190], which restricts access to information about f. First-order methods are those that use only local evaluations of f or ∇f (see [18, 194] for excellent and recent introductions). [190] first derived sub-linear lower bounds for first-order methods, i.e., $f(x_i) - f(x_{\min}) \geq \Omega(i^{-2})$, within the class of smooth convex functions. Shortly thereafter [195] obtained upper bounds of matching order.

The recent works on relative smoothness [16, 156, 113] derive first-order methods that do not require classical smoothness. As far as we know these conditions were first studied in [31] and rediscovered multiple times, e.g. [256]. Bauschke *et al.* [16] provided a general analysis of mirror descent under these generalized smoothness conditions for first-order methods. [156] provided the proof of linear convergence of the primal gradient and dual averaging schemes under both relative strong convexity and smoothness. Analyses of first-order methods under relative smoothness have been extended to non-convex f [34, 69], and an analogous notion of relative Lipschitz continuity has been developed for continuous convex optimization [153]. Accelerated versions of the primal schemes have been proposed in [113] and [112]. Relative smoothness has been used in the analysis of stochastic composite least-squares problems [81], symmetric non-negative matrix factorization [69], and the Sinkhorn algorithm [171]. These results do not contradict the classical lower bounds [190], because relative smoothness is a global condition that provides non-black-box information about f.

Dual preconditioned gradient descent extends linear preconditioning of gradient descent (see, e.g., [42, sect. 9.4]). Linear preconditioning improves dual gradient methods [94, 95], and is a classical tool in the study of iterative methods for linear systems [255, Chap. 13]. Non-linear preconditioning methods have recently been shown to stabilize Euler discretization schemes of stochastic differential equations [122, 226]. In fact, the non-linear preconditioning of [122] is the same as the one we consider for exponential penalty functions. We discuss the relationship of dual preconditioned gradient descent to existing methods in more detail in section 5.7.1.

5.4 Convex analysis background

5.4.1 Convex conjugate and Legendre functions

In this section we review some basic facts of convex analysis that will be used throughout. Let $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be a proper, lsc, convex function with domain $\operatorname{dom} h = \{x : \mathbb{R}^d : h(x) < \infty\}$. To indicate $\operatorname{dom} h = \mathbb{R}^d$, we simply define $h : \mathbb{R}^d \to \mathbb{R}$ as ranging only over the reals. Let $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ indicate the Euclidean norm and inner product, respectively, unless otherwise specified. The convex conjugate $h^* : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ of a proper, lsc convex function h is given by

$$h^*(x^*) = \sup\{\langle x, x^* \rangle - h(x) : x \in \operatorname{dom} h\}. \tag{5.6}$$

h^* is also a proper, lsc, convex function, and $(h^*)^* = h$ [222, Cor. 12.2.1]. If $g : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ is another proper, lsc, convex function and $g(x) \leq h(x)$ for all $x \in \mathbb{R}^d$, then $h^*(x^*) \leq g^*(x^*)$ for all $x^* \in \mathbb{R}^d$ follows by definition. For h differentiable on $\operatorname{int}(\operatorname{dom} h)$, we have by [222, Thm. 26.4] for $x \in \operatorname{int}(\operatorname{dom} h)$,

$$\langle x, \nabla h(x) \rangle = h(x) + h^*(\nabla h(x)). \tag{5.7}$$

For more on h^*, we refer readers to [222, 42, 36].

We make heavy use of Legendre type convex functions [222, Chap. 25]. Intuitively, these functions can be thought of as generalizations of positive definite quadratics and their gradient maps as generalizations of positive definite linear maps.

Definition 1 (Legendre convex functions)**.** *Let $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be proper, lsc, and convex. h is Legendre, if*

1. $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$,

2. *h is differentiable on* $\operatorname{int}(\operatorname{dom} h)$*, with* $\|\nabla h(x_i)\| \to \infty$ *for every sequence* $x_i \in \operatorname{int}(\operatorname{dom} h)$ *converging to a boundary point* $x \in \partial(\operatorname{dom} h)$*,*

3. *h is strictly convex on* $\operatorname{int}(\operatorname{dom} h)$*.*

A key consequence of property 2 of Legendre convex functions is that they can only be minimized in their interior. We confirm this in Lemma 5.4.1 below.

Lemma 5.4.1. *Let $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be a Legendre convex function with a minimum at $x_{\min} \in \operatorname{dom} h$. $x_{\min}$ is unique and furthermore $x_{\min} \in \operatorname{int}(\operatorname{dom} h)$.*

Proof. First, we argue that $x_{\min}$ cannot be found on the boundary by contradiction. Suppose that $x_{\min}$ is a boundary point. Since $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$, by convexity there exists a line segment connecting the boundary point $x_{\min}$ and any other interior point a. However, by [222, Lem. 26.2], we know that the directional derivative converges to $-\infty$ as we tend towards the boundary point on this line segment, hence $x_{\min}$ could not be a minimum of h. Thus we conclude that $x_{\min} \in \operatorname{int}(\operatorname{dom}(h))$. By property 3., Legendre functions are strictly convex on their interior, and thus $x_{\min}$ is unique. □

Property 2 together with Lemma 5.4.1 implies that Legendre convex functions grow without bound for sequences $x_i \in \mathbb{R}^d$ where $\|x_i - x_{\min}\| \to \infty$. We present a specialization of this fact in Lemma 5.4.2, which will be used in our analysis to show that the dual reference function k is radially unbounded.

Lemma 5.4.2. *Let $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be a Legendre convex function achieving its minimum at $0 \in \operatorname{dom} h$. Then h is radially unbounded, i.e., if $x_i \in \mathbb{R}^d$ is a sequence such that $\|x_i\| \to \infty$, then $h(x_i) \to \infty$.*

Proof. First, by Lemma 5.4.1 it follows that $0 \in \operatorname{int}(\operatorname{dom} h)$ and it is the unique minimum of h. Thus, we can define the sphere $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\| = r\}$ for some $r > 0$ such that $\mathcal{S} \in \operatorname{int}(\operatorname{dom} h)$. By continuity of h in the interior of its domain, and the uniqueness of the minimum at zero, we have $\inf_{x \in \mathcal{S}} h(x) > h(0)$. Now, assume without loss of generality that $\|x_i\| > r$. By strict convexity of Legendre functions, property 3, we have

$$h(0) + \frac{\|x_i\|}{r}\left(h\left(\frac{r x_i}{\|x_i\|}\right) - h(0)\right) < h(0) + (h(x_i) - h(0)) \tag{5.8}$$

and thus

$$h(x_i) > h(0) + \frac{\|x_i\|}{r}\left(\inf_{x \in \mathcal{S}} h(x) - h(0)\right). \tag{5.9}$$

Our result follows by taking $i \to \infty$. □

A second key consequence of Legendre structure is that the gradient map ∇h is invertible and given by $(\nabla h)^{-1} = \nabla h^*$, which also gives a characterization of the inverse of $\nabla^2 h(x)$. We summarize both of these properties in Lemma 5.4.3.

Lemma 5.4.3. *Let $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be a Legendre convex function, then h^* is also Legendre. The gradient map ∇h is one-to-one and onto from the open set $\operatorname{int}(\operatorname{dom} h)$ onto the open set $\operatorname{int}(\operatorname{dom} h^*)$, continuous in both directions, and for all $x \in \operatorname{int}(\operatorname{dom} h)$*

$$\nabla h^*(\nabla h(x)) = x. \tag{5.10}$$

If h is twice continuously differentiable on an open set containing x, then

$$\nabla^2 h^*(\nabla h(x)) \nabla^2 h(x) = \nabla^2 h(x) \nabla^2 h^*(\nabla h(x)) = I. \tag{5.11}$$

Proof. For the first part see Rockafellar [222, Thm. 26.5]. For (5.11), note that, by the inverse function theorem, ∇h^* is continuously differentiable at $\nabla h(x)$ under the assumption that ∇h is continuously differentiable on an open set containing x. The remainder follows by the chain rule applied to (5.10). □

5.4.2 Relative smoothness and relative strong convexity

Analyses of first-order methods for differentiable optimization typically require that ∇f is Lipschitz continuous (smooth). Recently [16, 156] discovered that certain so-called Bregman proximal gradient methods (mirror descent due to [189] is the first such method) require a generalized "relative" smoothness condition, which admits f that have non-Lipschitz ∇f. These relative smoothness conditions generalize the classical smoothness condition and are defined via the Bregman divergence [44] of a given a proper, lsc, convex function $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ that is differentiable on the interior of its domain. The Bregman divergence (see [15] for a review) is defined $\forall x \in \operatorname{dom} h, \forall y \in \operatorname{int}(\operatorname{dom} h)$,

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle . \tag{5.12}$$

In the special case of $h(x) = \|x\|_2^2 / 2$, D_h is the classical Euclidean distance squared. The relative conditions of relative strong convexity and relative smoothness [16, 156] relate two convex functions via their respective Bregman divergences.

Definition 2 (Relative smoothness and strong convexity)**.** *Let $g, h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be proper, lsc, convex functions that are differentiable on the interior of their domains.*

g is L-smooth relative to h on a convex set Q if $Q \subseteq \operatorname{int}(\operatorname{dom} h) \cap \operatorname{int}(\operatorname{dom} g)$, and there exists $L > 0$ such that for every $x, y \in Q$,

$$D_g(x, y) \leq L D_h(x, y).$$

g is μ-strongly convex relative to h on a convex set Q if $Q \subseteq \operatorname{int}(\operatorname{dom} g) \cap \operatorname{int}(\operatorname{dom} h)$, and there exists $\mu > 0$ such that for every $x, y \in Q$,

$$D_g(x, y) \geq \mu D_h(x, y).$$

Here, again, the special cases of relative strong convexity and smoothness with respect to $h(x) = \|x\|_2^2 / 2$ are exactly the classical conditions of strong convexity and smoothness, the first-order equivalents of (5.2). Lemma 5.4.4 (a re-statement of [156, Prop. 1.1]) below describes a variety of equivalent definitions for relative strong convexity and smoothness.

Lemma 5.4.4. (Equivalent definitions of relative conditions [156, Prop. 1.1]). *Let $g, h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be proper, lsc, convex functions that are differentiable on the interior of their domains. The following are equivalent*

1. *g is L-smooth relative to h on Q.*
2. *$Lh - g$ is convex on Q.*
3. *$\langle \nabla g(x) - \nabla g(y), x - y \rangle \leq L \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in Q$.*

The following are equivalent

1. *g is μ-strongly convex relative to h on Q.*
2. *$g - \mu h$ is convex on Q.*
3. *$\mu \langle \nabla h(x) - \nabla h(y), x - y \rangle \leq \langle \nabla g(x) - \nabla g(y), x - y \rangle$ for all $x, y \in Q$.*

Just as Lipschitz continuity of ∇g can be characterized by a bound on $\nabla^2 g$, relative smoothness and strong convexity can be characterized by the second derivatives of g and h. In particular, $Lh(x) - g(x)$ is convex if and only if $L\nabla^2 h(x) - \nabla^2 g(x)$ is positive semi-definite for all x. In this way, it is clear how relative smoothness generalizes classical smoothness (the $\nabla^2 h(x) = I$ case). We present the second-order characterization of the relative conditions in Lemma 5.4.5, generalized slightly to allow the bound to fail at a single point. This is useful for cases in which $\nabla^2 f$ is not continuous at $x_{\min}$. Typically, it is easiest to prove relative smoothness or strong convexity via these second-order equivalents.

Lemma 5.4.5 (Second-order characterizations of relative conditions)**.** *Let $g, h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be proper, lsc, convex functions that are differentiable on the interior of their domains. Let g, h be twice continuously differentiable at all $x \in Q \setminus \{z\}$ for $z \in Q \subseteq \operatorname{int}(\operatorname{dom} g) \cap \operatorname{int}(\operatorname{dom} h)$.*

1. *g is L-smooth relative to h on Q iff $\exists L > 0$,*

$$\nabla^2 g(x) \preceq L\nabla^2 h(x) \qquad \forall x \in Q \setminus \{z\}.$$

2. *g is μ-strongly convex relative to h on Q iff $\exists \mu > 0$,*

$$\mu\nabla^2 h(x) \preceq \nabla^2 g(x) \qquad \forall x \in Q \setminus \{z\}.$$

Proof. For relative smoothness, ($\Rightarrow$) follows directly from part one of [194, Thm. 2.1.4] applied to $f(x) = Lh(x) - g(x)$. For ($\Leftarrow$), we have $f(x) = Lh(x) - g(x)$ convex. Now, let $x, y \in Q$ and $t \in [0, 1]$ and define $x_t = y + t(x - y)$. There can be at most one time $a \in [0, 1]$ such that $x_a = z$. Take a to be that time, if it exists, or some arbitrary $a \in [0, 1]$, otherwise. We have

$$\begin{aligned}
&\langle L\nabla h(x) - \nabla g(x) - L\nabla h(y) + \nabla g(y), x - y\rangle \\
&= \langle \nabla f(x) - \nabla f(y), x - y\rangle \\
&\overset{(a)}{=} \lim_{\tau \downarrow a} \langle \nabla f(x) - \nabla f(x_\tau), x - y\rangle + \lim_{\tau \uparrow a} \langle \nabla f(x_\tau) - \nabla f(y), x - y\rangle \\
&\overset{(b)}{=} \lim_{\tau \downarrow a} \int_\tau^1 \left\langle x - y, \nabla^2 f(x_t)(x - y)\right\rangle \, dt + \lim_{\tau \uparrow a} \int_0^\tau \left\langle x - y, \nabla^2 f(x_t)(x - y)\right\rangle \, dt \geq 0,
\end{aligned}$$

(a) follows by the continuity of ∇f and (b) by the fundamental theorem of calculus. The relative strong convexity result follows analogously. □

The analyses of Bregman proximal gradient methods under relative smoothness rely on some standard manipulations of Bregman divergences. In Lemma 5.4.6 we summarize the ones used in our analysis.

Lemma 5.4.6. *Let $h : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be a convex Legendre function.*

1. *(Dual divergence) For all $x, y \in \operatorname{int}(\operatorname{dom} h)$,*

$$D_h(x, y) = D_{h^*}(\nabla h(y), \nabla h(x)).$$

2. *(Three-point property) [54, Lem. 3.1] For all $x, y, z \in \operatorname{int}(\operatorname{dom} h)$,*

$$D_h(x, y) = D_h(x, z) + D_h(z, y) - \langle x - z, \nabla h(y) - \nabla h(z)\rangle .$$

3. *(Bregman proximal inequality) [54, Lem. 3.2] Given a proper, lsc, convex* $\phi : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ *with* $\operatorname{int}(\operatorname{dom} h) \subseteq \operatorname{dom} \phi$*, and for* $y \in \operatorname{int}(\operatorname{dom} h)$*,*

$$z_{\min} = \arg \min_{z \in \operatorname{int}(\operatorname{dom} h)} \{\phi(z) + D_h(z, y)\}.$$

Then for all $x \in \operatorname{int}(\operatorname{dom} h)$

$$\phi(x) + D_h(x, y) \geq \phi(z_{\min}) + D_h(z_{\min}, y) + D_h(x, z_{\min}).$$

Proof. For the dual divergence property,

$$\begin{aligned} h(x) - h(y) - \langle \nabla h(y), x - y \rangle \\ &\overset{(a)}{=} -h^*(\nabla h(x)) + h^*(\nabla h(y)) - \langle \nabla h(y) - \nabla h(x), x \rangle \\ &\overset{(b)}{=} h^*(\nabla h(y)) - h^*(\nabla h(x)) - \langle \nabla h^*(\nabla h(x)), \nabla h(y) - \nabla h(x) \rangle , \end{aligned}$$

Here (a) follows from (5.7) and (b) follows from Lemma 5.4.3. The other two properties are given in [54, Lem. 3.1, Lem. 3.2]. □

5.5 Analysis of the dual preconditioned scheme

5.5.1 Motivation

Relative smoothness (Def. 2) is the key condition under which [16, 244, 156] analyzed the convergence of Bregman proximal gradient methods. In this section we show that the dual space preconditioned gradient descent method (Algorithm (2.1)) converges under a distinct relative smoothness condition. To motivate this, we consider two idealizations: one of the Bregman proximal gradient method and another of the dual space preconditioned gradient method.

First, consider the Bregman proximal gradient method update (5.3), which can be rewritten in the following form.

$$x_{i+1} = \arg \min_{x \in \operatorname{dom} f} \{\langle \nabla f(x_i) - L \nabla h(x_i), x \rangle + L h(x)\} \tag{5.13}$$

In this form, it is clear that, if $h = f$ and $L = 1$, then the iteration would converge in a single step. This is an idealization, because a single iteration would be as expensive to compute as the original problem. The spirit behind relative smoothness is that the condition $h = f$ can be relaxed to admit h for which the update (5.13) is efficiently solvable and the iterates still converge.

Now, consider the case that f is Legendre convex with a minimum at $x_{\min}$, and let $f_c^*(x^*) = f^*(x^*) - \langle x^*, x_{\min}\rangle$ for $x^* \in \mathbb{R}^d$. Notice that $\nabla f_c^*(\nabla f(x)) = x - x_{\min}$ by Lemma 5.4.3 and that Algorithm 2.1 with $k = f_c^*$ and $L_i = 1$ would converge in a single step. Thus, in analogy to relative smoothness analysis of [16] in the primal space, the spirit behind our analysis under relative smoothness in the dual space is that the requirement $k = f_c^*$ can be relaxed while maintaining the convergence of Algorithm 2.1. In particular, the essential features of f_c^* that we require of k are that it is minimized at 0 and smooth relative to f^*.

5.5.2 Relative conditions in the dual space

Our analysis guaranteeing the convergence of the dual preconditioned method uses the relative smoothness (Def. 2) of k relative to f^*. We call this condition dual relative smoothness, to contrast it with the typical application of relative smoothness [16, 156, 244], which we henceforth call primal relative smoothness. Similarly, we distinguish dual relative strong convexity from the condition of relative strong convexity applied in [156, 244] (henceforth called primal relative smoothness).

Definition 3 (Dual relative smoothness and dual relative strong convexity)**.** *Let $f, k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be Legendre convex functions. We say f is dual L^*-smooth (dual μ^*-strongly convex, resp.) relative to k on $Q \subseteq \operatorname{int}(\operatorname{dom} f)$, if k is L^*-smooth (μ^*-strongly convex, resp.) relative to f^* on $\nabla f(Q) \subseteq \operatorname{int}(\operatorname{dom} f^*)$. We abbreviate this condition to dual relative smoothness (dual relative strong convexity, resp.).*

Our dual relative conditions are defined via the convex conjugate f^*, which is generally inaccessible. Lemma 5.5.1 below gives equivalent definitions of dual relative smoothness and strong convexity in terms of objects that are more accessible.

Lemma 5.5.1 (Equivalent definitions of dual relative conditions)**.** *Let $f, k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be Legendre convex functions. The following are equivalent.*

1. *f is dual L^*-smooth relative to k on $\operatorname{int}(\operatorname{dom} f)$.*
2. *For all $x, y \in \operatorname{int}(\operatorname{dom} f)$,*
$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y).$$
3. *For all $x, y \in \operatorname{int}(\operatorname{dom} f)$,*
$$\langle \nabla k(\nabla f(x)) - \nabla k(\nabla f(y)), \nabla f(x) - \nabla f(y)\rangle \leq L^* \langle \nabla f(x) - \nabla f(y), x - y\rangle .$$

The following are equivalent.

1. *f is dual μ^*-strongly convex relative to k on* $\operatorname{int}(\operatorname{dom} f)$.

2. *For all $x, y \in \operatorname{int}(\operatorname{dom} f)$,*

$$\mu^* D_f(x, y) \leq D_k(\nabla f(y), \nabla f(x)).$$

3. *For all $x, y \in \operatorname{int}(\operatorname{dom} f)$,*

$$\mu^* \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \langle \nabla k(\nabla f(x)) - \nabla k(\nabla f(y)), \nabla f(x) - \nabla f(y) \rangle$$

Proof. We prove the relative smoothness results, and the relative strong convexity ones follow similarly. First, notice that that $\nabla f(\operatorname{int}(\operatorname{dom} f)) = \operatorname{int}(\operatorname{dom} f^*)$ by Lemma 5.4.3. So, by definition of dual relative smoothness we can apply relative smoothness in the dual space over $\operatorname{int}(\operatorname{dom} f^*)$. For $(1 \Rightarrow 2)$, we have by Lemma 5.4.4 that for all $x^*, y^* \in \operatorname{int}(\operatorname{dom} f^*)$,

$$D_k(y^*, x^*) \leq L^* D_{f^*}(y^*, x^*). \tag{5.14}$$

By Lemmas 5.4.3 and 5.4.6, this implies

$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_{f^*}(\nabla f(y), \nabla f(x)) = L^* D_f(x, y), \tag{5.15}$$

for all $x, y \in \operatorname{int}(\operatorname{dom} f)$. For $(2 \Rightarrow 3)$, simply sum over a permutation of x, y in 2. For $(3 \Rightarrow 1)$, we have by Lemma 5.4.3 for all $x^*, y^* \in \operatorname{int}(\operatorname{dom} f^*)$,

$$\langle \nabla k(x^*) - \nabla k(y^*), x^* - y^* \rangle \leq L^* \langle \nabla f^*(x^*) - \nabla f^*(y^*), x^* - y^* \rangle \tag{5.16}$$

This is equivalent to k being L^*-smooth relative to f^* on $\operatorname{int}(\operatorname{dom} f^*)$ by Lemma 5.4.4. □

The dual relative conditions have a natural second-order characterization, which reveals the structure of the difference between them and primal relative conditions. Again, typically it is easiest to prove dual relative smoothness (or strong convexity) via these second-order conditions.

Lemma 5.5.2 (Second-order characterizations of dual relative conditions)**.** *Let $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be Legendre convex, minimized at $x_{\min} \in \operatorname{int}(\operatorname{dom} f)$, and twice continuously differentiable at all $x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$. Let $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be Legendre convex, and twice continuously differentiable at all $x^* \in \operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$.*

1. *f is dual L^*-smooth relative to k on* $\operatorname{int}(\operatorname{dom} f)$ *iff* $\exists L^* > 0$,

$$\nabla^2 f(x) \preceq L^* [\nabla^2 k(\nabla f(x))]^{-1} \qquad \forall x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}.$$

2. *f is dual μ^*-strongly convex relative to k on* $\operatorname{int}(\operatorname{dom} f)$ *iff* $\exists \mu^* > 0$,

$$\mu^* [\nabla^2 k(\nabla f(x))]^{-1} \preceq \nabla^2 f(x) \qquad \forall x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}.$$

Remark 5.5.3. It is well-known that the primal and dual relative conditions are equivalent in the case of $\nabla^2 h(x) = I = \nabla^2 k(x^*)$ (see, e.g., [278, 232, 132, 276]). In particular, if f is μ-strongly convex and L-smooth on $\operatorname{int}(\operatorname{dom} f)$, then its convex conjugate f^* is $(1/L)$-strongly convex and $(1/\mu)$-smooth on $\operatorname{int}(\operatorname{dom} f^*)$. In fact, for twice continuously differentiable f, the equivalence is a simple consequence of Lemmas 5.4.5 and 5.5.2. However, this is equivalence is not true in general.

Given a Legendre convex $g : \mathbb{R} \to \{\mathbb{R}, \infty\}$ define the following sets of functions

$$\mathscr{F}_g = \{f : f \text{ is smooth and strongly convex relative to } g\}, \tag{5.17}$$
$$\mathscr{F}_g^* = \{f : f \text{ is dual smooth and dual strongly convex relative to } g\}. \tag{5.18}$$

Let $k(x^*) = |x^*|^q/q$ for $x^* \in \mathbb{R}$ and $1 < q < 2$. A simple argument by contradiction shows that $\mathscr{F}_k^* \nsubseteq \mathscr{F}_h$ for all twice continuously differentiable $h : \mathbb{R} \to \mathbb{R}$, implying that the primal and dual relative conditions are not equivalent in general. Consider

$$f_b(x) = |x - b|^p/p, \tag{5.19}$$

for $p = \frac{q}{q-1}$ and $x \in \mathbb{R}$. First $f_b \in \mathcal{F}_k^*$ for all b, which follows from $[k''(f_b'(x))]^{-1} = (p-1)|x-b|^{p-2} = f_b''(x)$ and Lemma 5.5.2. On the other hand, suppose there is some twice continuously differentiable $h : \mathbb{R} \to \mathbb{R}$ such that $f_b \in \mathscr{F}_h$ for all b. Then there exists $\mu > 0$ such that $\mu h''(b) \leq f_b''(b) = 0$ for all b. This implies that $h''(x) \equiv 0$ and thus $h(x) \equiv 0$. However, this leads to a contradiction, because smoothness is violated: $f_b''(b + \epsilon) > 0 = Lh''(x)$ for any $L, \epsilon > 0$.

Proof of Lemma 5.5.2. Again, we prove the relative smoothness result, and the relative strong convexity one follows similarly. By Lemma 5.4.3, if ∇f is continuously differentiable for $x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$, then ∇f^* is continuously differentiable for $x^* \in \operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$ by the inverse function theorem. Thus, by Lemma 5.4.5 relative smoothness in the dual space is equivalent to: for all $x^* \in \operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$,

$$\nabla^2 k(x^*) \preceq L^* \nabla^2 f^*(x^*). \tag{5.20}$$

By (5.11) of Lemma 5.4.3, these matrices are invertible (and thus positive definite). Thus, (5.20) is equivalent to for all $x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$,

$$\nabla^2 k(\nabla f(x)) \preceq L^* [\nabla^2 f(x)]^{-1}. \tag{5.21}$$

Since $A^{-1} \preceq B^{-1}$ is equivalent to $B \preceq A$ for positive definite matrices, we are done. □

As Remark 5.5.3 shows, the primal relative conditions (Def. 2) and the dual relative conditions (Def. 3) are not generally equivalent concepts. One major difference is the fact that dual relative conditions are invariant under horizontal translations of f. To see why, let f, k satisfy the dual relative smoothness condition (Def. 3) with constant L^* on $\operatorname{int}(\operatorname{dom} f)$. Define $g(x) = f(x - z)$ for $z \in \mathbb{R}^d$. Then, by Theorem 12.3 of [222], $g^*(x^*) = f^*(x^*) + \langle z, x^* \rangle$. First, note that $\operatorname{dom} g^* = \operatorname{dom} f^*$. Bregman divergences of functions that differ only in affine terms are identical [15], so we have for all $x^*, y^* \in \operatorname{int}(\operatorname{dom} g^*) = \operatorname{int}(\operatorname{dom} f^*)$

$$D_k(x^*, y^*) \leq L^* D_{f^*}(x^*, y^*) = L^* D_{g^*}(x^*, y^*). \tag{5.22}$$

Thus g is dual L^*-smooth relative to k on $\operatorname{int}(\operatorname{dom} g)$. Invariance under horizontal translation is clearly easy to violate in the case of primal relative smoothness (see previous remark).

Even if h is allowed to translate with f, the primal and dual relative conditions can lead to distinct conditioning. Given a positive definite $A \succ 0$, let

$$f(x) = \|Ax - b\|^p / p, \qquad h(x) = \left\|x - A^{-1}b\right\|^p / p, \qquad k(x^*) = \|x^*\|^q / q, \tag{5.23}$$

for $1/p + 1/q = 1$ and $p > 2$. It is not hard to show that f satisfies both the dual (with respect to k) and primal (with respect to h) relative conditions. Nonethless, the condition numbers are distinct. A simple calculation reveals that for this choice of k and h,

$$\frac{L}{\mu} = p^2 \left(\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \right)^p \qquad \text{vs.} \qquad \frac{L^*}{\mu^*} = (p-1)^2 \left(\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \right)^{4-q}, \tag{5.24}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the smallest and largest singular values of A, respectively. Thus, the primal condition number is larger than the dual number (since $4 - q = 3 - (p-1)^{-1} < p$ when $p > 2$). Similarly, the example $f(x) = \|Ax-b\|_4^4/4 + \|Cx-d\|_2^2/2$ of [156, p. 339] can be shown to have better conditioning under the dual preconditioned method than under the Bregman proximal gradient method.

To close this subsection, we consider a natural sufficient condition for dual relative smoothness: the Lipschitz continuity of the composition $\nabla k \circ \nabla f$.

Lemma 5.5.4. *Let $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be a proper function with $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$ that is twice continuously differentiable on $\operatorname{int}(\operatorname{dom} f)$. Let $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be Legendre convex and twice continuously differentiable on $\nabla f(\operatorname{int}(\operatorname{dom} f))$. If $\nabla k \circ \nabla f$ is L^*-Lipschitz continuous, then for all $x \in \operatorname{int}(\operatorname{dom} f)$,*

$$\nabla^2 f(x) \preceq L^* [\nabla^2 k(\nabla f(x))]^{-1}. \tag{5.25}$$

In particular, if f is Legendre convex, then f is dual L^*-smooth relative to k on $\operatorname{int}(\operatorname{dom} f)$.

Proof. Let $x \in \operatorname{int}(\operatorname{dom} f)$, $v \in \mathbb{R}^d$, $K(x) = \nabla^2 k(\nabla f(x))$, and $F(x) = \nabla^2 f(x)$. Note that ∇k is continuously differentiable at $\nabla f(x)$. Hence, $K(x)$ is invertible and thus positive definite by (5.11) of Lemma 5.4.3. By L^*-Lipschitz continuity we also have

$$\|K(x)F(x)v\| = \lim_{t \to 0} \frac{\|\nabla k(\nabla f(x + tv)) - \nabla k(\nabla f(x))\|}{t} \leq L^* \|v\|. \tag{5.26}$$

Thus, $\|K(x)F(x)\| \leq L^*$ for the induced matrix norm. Now,

$$\begin{aligned} \left\langle v, [K(x)]^{1/2} F(x) [K(x)]^{1/2} v \right\rangle &\leq \rho([K(x)]^{1/2} F(x) [K(x)]^{1/2}) \|v\|^2 \\ &= \rho(K(x)F(x)) \|v\|^2 \\ &\leq \|K(x)F(x)\| \|v\|^2 \leq L^* \|v\|^2 \end{aligned} \tag{5.27}$$

where $\rho(A)$ is the spectral radius of A. The result follows because $B^{1/2} A B^{1/2} \preceq I$ implies $A \preceq B^{-1}$ for positive definite B. □

5.5.3 Convergence rates under dual relative smoothness

In this subsection we provide conditions under which convergence rates for Algorithm 2.1 can be established for Legendre convex f. The first key ingredient is the condition of dual relative smoothness between f and k with constant L^*, developed in the previous section. The second is the requirement that $0 = \arg\min_{x^*} k(x^*)$. In this case, we find that $k(\nabla f(x_i))) - k(0)$ converges with rate $\mathcal{O}(i^{-1})$. When f is also dual $\mu^* > 0$ strongly convex relative to k, we find that $f(x_i) - f(x_{\min})$ converges with rate $\mathcal{O}((1 - \mu^*/L^*)^i)$. Both of these results are derived from the following descent lemma.

Lemma 5.5.5 (Descent lemma)**.** *Given $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex, $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex with $0 = \arg\min_{x^* \in \operatorname{dom} k} k(x^*)$, and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. If f is dual L^*-smooth relative to k on $\operatorname{int}(\operatorname{dom} f)$, then for all $i > 0$, the iterates of Algorithm 2.1 satisfy*

1. *$x_i \in \operatorname{int}(\operatorname{dom} f)$,*

2. for all $x \in \operatorname{int}(\operatorname{dom} f)$,

$$k(\nabla f(x_i)) \leq k(\nabla f(x)) - D_k(\nabla f(x), \nabla f(x_{i-1})) + L^* D_f(x_{i-1}, x) - L^* D_f(x_i, x). \quad (5.28)$$

In particular, we have for $i > 0$,

3. $k(\nabla f(x_i)) + L^* D_f(x_i, x_{i-1}) \leq k(\nabla f(x_{i-1}))$,

4. $L^* D_f(x_i, x_{\min}) + D_k(\nabla f(x_i), 0) + D_k(0, \nabla f(x_{i-1})) \leq L^* D_f(x_{i-1}, x_{\min})$.

Proof. First, note $\operatorname{int}(\operatorname{dom} f^*) \subseteq \operatorname{int}(\operatorname{dom} k)$ by relative smoothness. Thus, by Lemma 5.4.3 we have $\nabla f(x) \in \operatorname{int}(\operatorname{dom} k)$ for all $x \in \operatorname{int}(\operatorname{dom} f)$.

We proceed by induction. For $i = 0$ we have $x_0 \in \operatorname{int}(\operatorname{dom} f)$ by assumption. Now, for $i > 0$, assume the induction hypothesis for x_{i-1}. First, define

$$x_\lambda = x_{i-1} - \frac{1}{\lambda} \nabla k(\nabla f(x_{i-1})) \quad (5.29)$$

for $\lambda > 0$. Because $x_{i-1} \in \operatorname{int}(\operatorname{dom} f) \neq \emptyset$, the following set is not empty,

$$S = \{\lambda \geq L^* : x_\lambda \in \operatorname{int}(\operatorname{dom} f)\}. \quad (5.30)$$

Let $x^*_{i-1} = \nabla f(x_{i-1})$ and $x^*_\lambda = \nabla f(x_\lambda)$ for all $\lambda \in S$. By Lemma 5.4.3, we have

$$\nabla f^*(x^*_\lambda) = \nabla f^*(x^*_{i-1}) - \frac{1}{\lambda} \nabla k(x^*_{i-1}). \quad (5.31)$$

Therefore x^*_λ satisfies the stationary condition of the following subproblem,

$$\min_{x^* \in \operatorname{int}(\operatorname{dom} f^*)} \left\{ \frac{1}{\lambda} \left\langle \nabla k(x^*_{i-1}), x^* - x^*_{i-1} \right\rangle + D_{f^*}(x^*, x^*_{i-1}) \right\}. \quad (5.32)$$

From the Bregman proximal inequality of Lemma 5.4.6 applied with $h = f^*$, $\phi(x^*) = \frac{1}{\lambda} \left\langle \nabla k(x^*_{i-1}), x^* - x^*_{i-1} \right\rangle$, $x = x^*$, $y = x^*_{i-1}$ and $z_{\min} = x^*_\lambda$, we have

$$\begin{aligned} &\left\langle \nabla k(x^*_{i-1}), x^* - x^*_{i-1} \right\rangle + \lambda D_{f^*}(x^*, x^*_{i-1}) \geq \\ &\left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle + \lambda D_{f^*}(x^*_\lambda, x^*_{i-1}) + \lambda D_{f^*}(x^*, x^*_\lambda). \end{aligned} \quad (5.33)$$

Putting everything together, we have for all $x^* \in \operatorname{int}(\operatorname{dom} f^*)$

$$\begin{aligned} k(x^*_\lambda) &\overset{(a)}{\leq} k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle + L^* D_{f^*}(x^*_\lambda, x^*_{i-1}) \\ &\overset{(b)}{\leq} k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle + \lambda D_{f^*}(x^*_\lambda, x^*_{i-1}) \\ &\overset{(c)}{\leq} k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^* - x^*_{i-1} \right\rangle + \lambda D_{f^*}(x^*, x^*_{i-1}) - \lambda D_{f^*}(x^*, x^*_\lambda) \\ &\overset{(d)}{\leq} k(x^*) - D_k(x^*, x^*_{i-1}) + \lambda D_{f^*}(x^*, x^*_{i-1}) - \lambda D_{f^*}(x^*, x^*_\lambda). \end{aligned} \quad (5.34)$$

(a) follows from dual L^*-smoothness, (b) from $L^* \leq \lambda$ and the non-negativity of the Bregman divergence, (c) from (5.33), and (d) by definition and simple algebra. Taking $x^* = x^*_{i-1}$ and recalling the definition of x^*_{i-1} and x^*_λ reveals that

$$k(\nabla f(x_\lambda)) + \lambda D_{f^*}(\nabla f(x_{i-1}), \nabla f(x_\lambda)) \leq k(\nabla f(x_{i-1})). \tag{5.35}$$

Now, our goal is to show that $x_i = x_{L^*} \in \operatorname{int}(\operatorname{dom} f)$ by showing that $L^* \in S$. We proceed by contradiction, so suppose $L^* \notin S$. Then $x_{L^*} \in \mathbb{R}^d \setminus \operatorname{int}(\operatorname{dom} f)$. Hence we can find $\Lambda \geq L^*$ such that $x_\Lambda \in \partial(\operatorname{dom} f)$. Now take a sequence $\lambda_j \to \Lambda$ such that $\lambda_j > \Lambda$. By the above discussion for all $j \geq 0$ we have $k(\nabla f(x_{\lambda_j})) \leq k(\nabla f(x_{i-1}))$. k being minimized at 0 means it satisfies Lemma 5.4.2 and thus is radially unbounded. This implies that $\left\|\nabla f(x_{\lambda_j})\right\| \leq C$ for some $C > 0$ and all $j \geq 0$. But this contradicts the requirement from property 2 of Legendre functions that $\left\|\nabla f(x_{\lambda_j})\right\| \to \infty$ since $x_{\lambda_j} \to x_\Lambda \in \partial(\operatorname{dom} f)$ by assumption. This completes the proof that $x_i = x_{L^*} \in \operatorname{int}(\operatorname{dom} f)$. Since $L^* \in S$, (5.34) along with the dual divergence property of Lemma 5.4.6 ensures that 2. holds. Taking $x = x_{i-1}$ in 2. ensures that 3. holds while 4. follows by taking $x = x_{\min}$. □

We are ready to analyze the convergence rates of the dual preconditioned gradient descent method.

Theorem 5.5.6. *Given $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex, $x_{\min} = \arg\min_x f(x)$, $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex, $0 = \arg\min_{p \in \operatorname{dom} k} k(p)$, and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. If f is dual L^*-smooth relative to k on $\operatorname{int}(\operatorname{dom} f)$, then for all $i > 0$ and $x \in \operatorname{int}(\operatorname{dom} f)$ the iterates of Algorithm 2.1 satisfy*

$$k(\nabla f(x_i)) - k(0) \leq \frac{L^*}{i}(f(x_0) - f(x_{\min})). \tag{5.36}$$

In particular, $\nabla f(x_i) \to 0$. If additionally f is dual μ^*-strongly convex relative to k on $\operatorname{int}(\operatorname{dom} f)$ with $\mu^* > 0$, then for all $i > 0$ the iterates of Algorithm 2.1 satisfy

$$f(x_i) - f(x_{\min}) \leq \left(1 - \frac{\mu^*}{L^*}\right)^i (f(x_0) - f(x_{\min})). \tag{5.37}$$

Remark 5.5.7. Ensuring that k is minimized at 0 is not difficult. Let l satisfy the requirements on k in Theorem 5.5.6 and $0 \in \operatorname{int}(\operatorname{dom} l)$, but with a minimum other than 0. Then $k(p) = l(p) - \langle \nabla l(0), p \rangle$ will suffice for Theorem 5.5.6.

Proof of Theorem 5.5.6. First, we have $x_i \in \operatorname{int}(\operatorname{dom} f)$ and $k(\nabla f(x_i))$ is non-increasing by 1. and 3. of the Descent Lemma 5.5.5. Thus, by 2. of the same lemma, for all $x \in \operatorname{int}(\operatorname{dom} f)$ and $i > 0$

$$\begin{aligned} i(k(\nabla f(x_i)) - k(\nabla f(x))) &\leq \sum_{j=1}^{i} k(\nabla f(x_i)) - k(\nabla f(x)) \\ &\leq L^* D_f(x_0, x) - L^* D_f(x_i, x). \end{aligned} \tag{5.38}$$

Dropping the negative term on the right hand side, dividing by i, and taking $x = x_{\min}$ gives our first result. This implies that $k(\nabla f(x_i)) \to k(0)$, which implies that $\nabla f(x_i) \to 0$ by continuity and the uniqueness of k's minimum. Now, assume that f is dual μ^*-strongly convex relative to k on $\operatorname{int}(\operatorname{dom} f)$ with $\mu^* > 0$. For all $i > 0$,

$$\begin{aligned} L^*(f(x_i) - f(x_{\min})) &\overset{(a)}{\leq} L^*(f(x_{i-1}) - f(x_{\min})) - D_k(0, \nabla f(x_{i-1})) \\ &\overset{(b)}{\leq} L^*(f(x_{i-1}) - f(x_{\min})) - \mu^*(f(x_{i-1}) - f(x_{\min})), \end{aligned} \tag{5.39}$$

where (a) follows as an implication of 4. in the Descent Lemma 5.5.5 and (b) follows from dual relative strong convexity. This inequality implies our desired result. □

Theorem 5.5.6 guarantees the convergence of the iterates of Algorithm 2.1 under the assumption that dual relative smoothness hold globally for a fixed L^*. Unfortunately it may be difficult to to derive a tight bound on L^* or small L^* may be appropriate locally. In this case, it may be useful to use a line search to choose L^*. Consider the following generalization of the update rule of Algorithm 2.1,

$$x_{i+1} = x_i - \frac{1}{L_i^*} \nabla k(\nabla f(x_i)) \tag{5.40}$$

where $L_i^* > 0$ is allowed to depend on the iteration. The next proposition shows that, under suitable assumptions, (5.40) converges with rates analogous to Theorem 5.5.6.

Proposition 5.5.8 (Adaptive step sizes)**.** *Given $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex, $k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ Legendre convex with $0 = \arg\min_{x^* \in \operatorname{dom} k} k(x^*)$, and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. If, for all $i > 0$ the iterates defined by (5.40) satisfy*

1. *$x_i \in \operatorname{int}(\operatorname{dom} f)$,*
2. *$k(\nabla f(x_i)) \leq k(\nabla f(x_{i-1}))$,*
3. *$k(\nabla f(x_i)) - k(0) \leq L_{i-1}^*(f(x_{i-1}) - f(x_i))$,*

then we have

$$k(\nabla f(x_i)) - k(0) \leq \frac{\max_{0\leq j\leq i-1} L_j^*}{i}(f(x_0) - f(x_{\min})). \tag{5.41}$$

Remark 5.5.9. In practice, a possible choice of step sizes is

$$L_{i-1}^* = \min\{2^r, r \in \mathbb{Z} : \text{ 1., 2., and 3. of Proposition 5.5.8 are satisfied}\}. \tag{5.42}$$

If L^* is the smallest real number such that f is dual L^*-smooth relative to k (see Lemma 5.5.1 for an equivalent condition), then this scheme satisfies that $L_{i-1}^* < 2L^*$ for every $i > 0$ (hence we are making steps that are almost as large or larger as if we would use the smallest possible fixed L^*, without knowing the value of L^* in advance). The search through the set in (5.42) for finding L_i^* can be initialized at L_{i-1}^*.

Proof of Proposition 5.5.8. The proof follows similar lines as in the previous case. First, by summing up the inequalities from 3, we obtain that

$$\sum_{1\leq j\leq i}[k(\nabla f(x_j)) - k(0)] \leq \sum_{1\leq j\leq i} L_{i-1}^*(f(x_{i-1}) - f(x_i)) \leq (f(x_0) - f(x_{\min})) \max_{0\leq j\leq i-1} L_j^*,$$

and using 2., it follows that $\sum_{1\leq j\leq i}[k(\nabla f(x_j)) - k(0)] \geq i(k(\nabla f(x_i)) - k(0))$. The result follows directly. □

An important question that we do not address in this section is whether the sublinear convergence of $k(\nabla f(x_i)) - k(0)$ implies specific rates of convergence of other quantities of interest. These might be, for example, $\|x_i - x_{\min}\|$ or $f(x_i) - f(x_{\min})$. Rates for these will likely depend on both f and k.

5.6 Applications

5.6.1 Exponential Penalty Functions

Consider the following linear programming problem.

$$\min_{x\in\mathbb{R}^d}\{c^T x : Ax \leq b\}, \tag{LP}$$

where $c \in \mathbb{R}^d$, $b \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n\times d}$. Associate with this linear program the following relaxation into an unconstrained problem: $\min_{x\in\mathbb{R}^d} f_\tau(x)$ for

$$f_\tau(x) = c^T x + \tau \sum_{i=1}^{n} \exp((A_i x - b_i)/\tau), \tag{5.43}$$

where $\tau > 0$ and A_i is the ith row of A (a row vector). This approximation of (LP) with exponential penalty functions was studied by several authors (see [251, 60, 209, 12]) and is directly useful in the machine learning literature for boosting (see, e.g., [167]). Derivatives of all orders for this problem are unbounded as $\|x\| \to \infty$, and analyses of optimization methods, which rely on global smoothness constants, do not provide global convergence rates. In this section we design a dual reference function for f_τ under the following assumptions on (LP).

Assumption H. *Suppose that the following hold for problem* (LP).

1. $\|A_i\| = 1$ *for* $1 \leq i \leq n$.
2. $A \in \mathbb{R}^{n \times d}$ *is of full rank* $d \leq n$.
3. $P = \{x \in R^n : Ax \leq b\}$ *is a polytope, which is contained in a Euclidean ball of radius* $R > 0$ *and contains a Euclidean ball of radius* $r > 0$.

The dual reference function will be designed so that f_τ is dual smooth relative to it and Algorithm 2.1, with appropriate step-size choices, converges with global guarantees.

Define the dual reference function $k : \mathbb{R}^d \to \mathbb{R}$,

$$k(x^*) = \|x^*\| - \log(\|x^*\| + 1). \tag{5.44}$$

This behaves like a quadratic $\|x^*\|^2/2$ near its minimum $x^* = 0$ and like $\|x^*\|$, i.e., grows linearly, at infinity. It is also possible to verify that k is Legendre convex. Furthermore, we have:

$$\nabla k(x^*) = \frac{x^*}{\|x^*\| + 1}, \qquad \nabla^2 k(x^*) = \frac{I}{\|x^*\| + 1} - \frac{x^* x^{*T}}{(\|x^*\| + 1)^2 \|x^*\|}. \tag{5.45}$$

Hence, $[\nabla^2 k(x^*)]^{-1} \succeq (1 + \|x^*\|)I$. From Lemma 5.5.4 and this inequality it follows that the fact that f is dual L^*-smooth relative to k is implied by

$$\nabla^2 f_\tau(x) \preceq L^* \left[1 + \|\nabla f_\tau(x)\|\right] I \qquad \forall x \in \mathbb{R}^d. \tag{5.46}$$

This is the strategy of the following theorem, which shows that f_τ is dual smooth to this choice of k under our assumptions.

Proposition 5.6.1. *Under Assumption H for* f_τ *defined in* (5.43) *and* k *defined in* (5.44)*, we have that*

$$\nabla^2 f_\tau(x) \preceq L_\tau^* [\nabla^2 k(\nabla f_\tau(x))]^{-1} \qquad \forall x \in \mathbb{R}^d, \tag{5.47}$$

where the dual relative smoothness constant is given by

$$L_\tau^* = \frac{2R}{r}\frac{\left\|A^T A\right\|}{\tau}(\eta + \|c\|). \tag{5.48}$$

Here, $\left\|A^T A\right\|$ *is the induced matrix norm, and*

$$\eta = \sup_{\|s\|_\infty \leq 1} \left\|A^T s\right\| \leq \sqrt{n}\left\|A^T\right\|_\infty. \tag{5.49}$$

Because f_τ *and* k *are Legendre convex,* f *is dual smooth relative to* k *and Theorem 5.5.6 implies that Algorithm 2.1 converges with* $k(\nabla f(x_i))$ *converging at a rate* $\mathcal{O}(1/i)$*.*

Remark 5.6.2. From Theorem 5.5.6, we have

$$k(\nabla f_\tau(x)) \leq \frac{L_\tau^*(f_\tau(x_0) - f_\tau(x_{\min}))}{i}. \tag{5.50}$$

This suggests that, if we can start from an initial point within the polytope, then we can reach a point where $\|\nabla f_\tau(x)\|$ is significantly less than $\|c\|$ (which is expected to be near the minimum) in polynomial amount of steps, depending on the conditioning R/r and the value of τ. The step-size $1/L_i^*$ can also be chosen adaptively, as explained in Proposition 5.5.8. Near the minimum, both $f_\tau(x)$ and $k(x^*)$ behave like quadratic functions, so local linear convergence rates hold. We believe that this iterative scheme is reasonably efficient for high dimensional well-conditioned polytopes, but in other less well conditioned instances it is outperformed by existing algorithms such as multiplicative weights [9] or [59], which is based on Newton's method (hence uses second-order information).

Proof of Proposition 5.6.1. Note that $1 \leq \eta \leq n$, because $\|A_i\| = 1$. Let $\alpha(x) := \max_{i\in[n]}(A_i x - b_i)$. Then $\alpha(x) < 0$ inside the polytope and $\alpha(x) > 0$ outside of it. By differentiation, we have

$$\nabla f_\tau(x) = \sum_{i=1}^n A_i \exp((A_i x - b_i)/\tau) + c, \tag{5.51}$$

$$\nabla^2 f_\tau(x) = \sum_{i=1}^n \frac{A_i^T A_i}{\tau} \exp((A_i x - b_i)/\tau). \tag{5.52}$$

Note that f_τ is defined everywhere and differentiable. Furthermore, under our assumption that $\operatorname{rank}(A^T A) = \operatorname{rank}(A) = d$, it is evidently strictly convex and therefore Legendre.

The Hessian of f_τ satisfies

$$\nabla^2 f_\tau(x) \preceq \exp(\alpha(x)/\tau)\frac{A^T A}{\tau} \preceq \exp\left(\alpha(x)/\tau\right)\frac{\left\|A^T A\right\|}{\tau} I. \tag{5.53}$$

Because $\eta \geq 1$, it is clear that the claim of the theorem holds for every x where $\alpha(x) \leq 0$ (i.e. inside the polytope or on its boundary). From now on we will assume that x is such that $\alpha(x) > 0$ (outside of the polytope). Let x_c be a minimizer of $\alpha(x)$ (at least one exists since the polytope is compact and $\alpha(x)$ is a continuous function), then using the assumption $\|A_i\| = 1$ it follows that $\alpha(x_c) = -r < 0$. Hence $x \neq x_c$. We are going to need an upper bound on $\|x - x_c\|$, which we will obtain as follows. By the definitions, we have $A_i x_c \leq -r + b_i$ and $A_i x = A_i x - b_i + b_i \leq \alpha(x) + b_i$, hence

$$\begin{aligned} A_i \left(x_c + \frac{r}{\alpha(x)+r}(x - x_c) \right) &= \frac{r}{\alpha(x)+r} A_i x + \frac{\alpha(x)}{\alpha(x)+r} A_i x_c \\ &\leq \frac{r}{\alpha(x)+r}(\alpha(x) + b_i) + \frac{\alpha(x)}{\alpha(x)+r}(-r + b_i) = b_i. \end{aligned}$$

Therefore $x_c + \frac{r}{\alpha(x)+r}(x - x_c) \in P \subset B_{x_c}(2R)$, so

$$0 < \|x - x_c\| \leq 2\frac{\alpha(x)+r}{r} R \quad \text{and} \quad \|x - x_c\|^{-1} \geq \frac{r}{\alpha(x)+r}\frac{1}{2R}. \tag{5.54}$$

Let $\mathcal{I} = \{i \in [n];\, A_i x - b_i > 0\}$, $\mathcal{J} = \{i \in [n];\, A_i x - b_i \leq 0\}$, and

$$G_{\mathcal{I}}(x) = \sum_{i \in \mathcal{I}} e^{\frac{1}{r}(A_i x - b_i)} A_i \qquad G_{\mathcal{J}}(x) = \sum_{i \in \mathcal{J}} e^{\frac{1}{r}(A_i x - b_i)} A_i. \tag{5.55}$$

Then $\nabla f_\tau(x) = G_{\mathcal{I}}(x) + G_{\mathcal{J}}(x) + c$. We have

$$\begin{aligned} \|G_{\mathcal{I}}(x)\| \geq \frac{G_{\mathcal{I}}(x)^T (x - x_c)}{\|x - x_c\|} &= \|x - x_c\|^{-1} \sum_{i \in \mathcal{I}} e^{\frac{1}{r}(A_i x - b_i)} A_i (x - x_c) \\ &\overset{(a)}{\geq} \|x - x_c\|^{-1} e^{\frac{\alpha(x)}{\tau}} (\alpha(x) + r) \\ &\overset{(b)}{\geq} \frac{r}{2R} e^{\frac{\alpha(x)}{\tau}}. \end{aligned}$$

Here, (a) follows from the facts that there is a $j \in \mathcal{I}$ such that $A_j(x - x_c) = \alpha(x) + b_j - A_j x_c \geq \alpha(x) + r$ and the fact that $A_i(x - x_c) \geq b_i + r - b_i > 0$ holds for every $i \in \mathcal{I}$. (b) follows from (5.54). From (5.53) we obtain that

$$\begin{aligned} \nabla^2 f_\tau(x) &\preceq \exp\left(\alpha(x)/\tau\right) \frac{\left\|A^T A\right\|}{\tau} I \\ &\preceq \frac{2R}{r} \frac{\left\|A^T A\right\|}{\tau} \|G_{\mathcal{I}}(x)\|\, I \\ &\preceq \frac{2R}{r} \frac{\left\|A^T A\right\|}{\tau} (\|\nabla f_\tau(x)\| + \|G_{\mathcal{J}}(x)\| + \|c\|) I. \end{aligned} \tag{5.56}$$

Hence (5.46) follows from the facts that $\|G_{\mathcal{J}}(x)\| \leq \eta$ and $\eta + \|c\| \geq 1$. As discussed (5.47) follows from $[\nabla^2 k(x^*)]^{-1} \succeq (1 + \|x^*\|) I$. $\square$

5.6.2 p-norm Regression

Consider the following p-norm regression problem,

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_p^p, \tag{pnorm}$$

where $A \in \mathbb{R}^{n \times d}$, $d \ll n$, $b \in \mathbb{R}^n$, and $p \geq 1$. This problem is a useful abstraction for some important graph problems, including Lipschitz learning on graphs [143] and ℓ_p-norm minimizing flows [4]. Algorithms specialized for p-norm regression have recently been studied in the theoretical computer science literature by several authors (see, e.g., [48, 2] and references therein). In this subsection, we design an appropriate dual reference function for (pnorm) under the following assumptions. Let A_i denote the rows of A (as row vectors).

Assumption I. *Suppose that the following hold for problem* (pnorm)*.*

1. $2 \leq p < \infty$.
2. *A is full rank d, and for all $x \in \mathbb{R}^d$ there is a subset $I(x) \subset [n]$ such that $A_i x \neq b_i$ for all $i \in I(x)$, and* $\operatorname{span}\{A_i : i \in I(x)\} = \mathbb{R}^d$.
3. $c_G = \inf_{\|s\|=1} \|As\|_p^p > 0$.
4. $c_H = \inf_{u,v \in \mathbb{R}^d : \|u\|=1, \|v\|=1} \sum_{i=1}^n |A_i u|^{p-2} (A_i v)^2 > 0$.

Remark 5.6.3. Although these assumptions seem restrictive, we can show that, if $n \geq 2d - 1$ and $(A_i)_{1 \leq i \leq n}$ and $(b_i)_{1 \leq i \leq n}$ are chosen as independent random variables with densities that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and $\mathbb{R}$, then the assumptions hold with probability 1. Assumption 2 is implied by the stronger assumption that any d rows of A define a full rank d matrix, and the maximal number of equalities $A_i x = b_i$ that hold for any x is no more than d. This stronger version of Assumption 2, and Assumption 3 holds with probability 1 under the random allocation due to the fact that the set of real valued $d \times d$ matrices with determinant 0 has Lebesgue-measure 0 in $\mathbb{R}^{d \times d}$ (due to the fact that the determinant is a multivariate polynomial of the entries, and the zero set of such polynomials has Lebesgue measure zero unless they are constant 0, see [51]). The minimum in Assumption 4 is achieved for some $u_{\min}$ and $v_{\min}$ due to continuity and compactness of the unit sphere. Since any d rows of A form an independent basis with probability 1, it follows that u and v can be orthogonal to at most $d - 1$ of them, respectively, so using $n \geq 2d - 1$ there exists an i in the sum $\sum_{i=1}^n |A_i u_{\min}|^{p-2} (A_i v_{\min})^2$ that is non-zero, hence Assumption 4 holds.

Consider the dual reference function $k : \mathbb{R}^d \to \mathbb{R}$,

$$k(x^*) = \tfrac{1}{q}\left(\|x^*\|^2 + 1\right)^{\frac{q}{2}} - \tfrac{1}{q}, \tag{5.57}$$

for $q = \frac{p}{p-1}$ (hence $\frac{1}{p} + \frac{1}{q} = 1$). This behaves like a quadratic $\|x^*\|^2/2$ near its minimum $x^* = 0$ and like $\|x^*\|^q/q$ at infinity. For this k, we have

$$\nabla k(x^*) = x^*(1 + \|x^*\|^2)^{\frac{q-2}{2}} \tag{5.58}$$

As the next theorem shows dual relative strong convexity and smoothness of (pnorm) relative to this k hold under our assumptions.

Proposition 5.6.4. *Let $f(x) = \|Ax - b\|_p^p$ be the p-norm objective. Under Assumption I for k defined in* (5.57)*, there exists $\mu^*, L^* > 0$ such that*

$$\mu^*[\nabla^2 k(\nabla f(x))]^{-1} \preceq \nabla^2 f(x) \preceq L^*[\nabla^2 k(\nabla f(x))]^{-1} \qquad \forall x \in \mathbb{R}^d. \tag{5.59}$$

See (5.69) *and* (5.70) *for the definitions of μ^* and L^*. Because f and k are Legendre convex, f is dual smooth and dual strongly convex relative to k and Theorem 5.5.6 implies that Algorithm 2.1 converges with $f(x_i) - f(x_{\min})$ converging at a linear rate $\mathcal{O}((1 - \mu^*/L^*)^i)$.*

To test the empirical performance of this method, we have implemented it with A_i, b, and x_0 i.i.d. as standard normals for power $p = 4$, $d \in \{10^2, 10^3, 10^4\}$, and $n = 10d$. The inverse step-size L_0^* was chosen to be $L_0^* = 1$ initially, and multiplied by 2 if the function value would increase due to too large steps (hence this was chosen adaptively in the beginning, but L_i^* was never decreased later on). As Figure 5.1 shows, empirically our method seems to be performing well, with high precision achieved after 50-80 gradient evaluations, and the convergence rate seems to be mostly unaffected by the dimension d. Hence in this random setting dual space preconditioning is indeed very efficient, and competitive with previous works [48, 2, 4] which had dimension dependent convergence rates. We think that based on Proposition 5.6.4, it can be shown that with high probability, dimension-free convergence rates hold in this random scenario when the number of vectors n tends to infinity (the proof would be based on concentration inequalities for empirical processes, see e.g. [39] for an overview of such inequalities). Note however that we do not believe this always to be the case for general non-random A and b, and there could be instances of very poor conditioning (such as when $n \approx d$) where the homotopy method of [48] or the IRLS method of [3] could perform better. The proof of Proposition 5.6.4 is based on the following two lemmas.

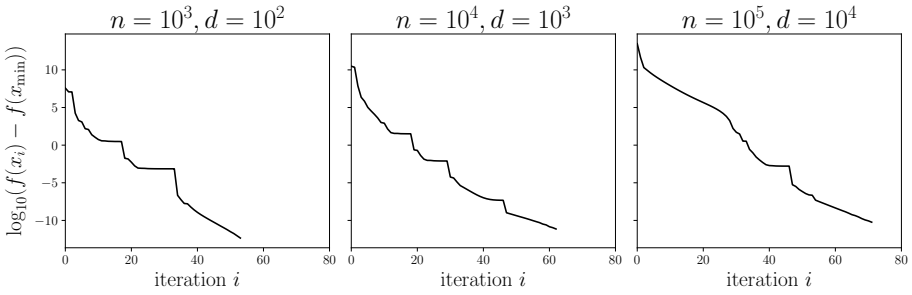


Figure 5.1: Convergence rates for p-norm regression are mostly unaffected by the dimension d for these random instances with $p = 4$.

Lemma 5.6.5 (Bounds on the gradient)**.** *Let $f(x) = \|Ax - b\|_p^p$ be the p-norm objective for* (pnorm)*. Under Assumption I, we have*

$$L_G\|x\|^{p-1} - C_G \leq \|\nabla f(x)\| \leq U_G\|x\|^{p-1} + D_G \tag{5.60}$$

for all $x \in \mathbb{R}^d$, with constants

$$L_G = 2^{-p+1}c_G = 2^{-p+1}\inf_{\|s\|=1}\|As\|_p^p, \quad C_G = \left(\sum_{i=1}^n |b_i|^p\right)^{(p-1)/p} \cdot c_G^{1/p},$$
$$U_G = 2^{p-2}(p+1)\sup_{\|s\|=1}\|As\|_p^p, \quad D_G = 2^{p-2}(p-1)\left(\sum_{i=1}^n |b_i|^p\right).$$

Proof. By differentiation, we have

$$\nabla f(x) = p\sum_{i=1}^n |A_ix - b_i|^{p-2}(A_ix - b_i)A_i, \tag{5.61}$$

thus

$$\begin{aligned}
\|\nabla f(x)\| &= p\left\|\sum_{i=1}^n |A_ix - b_i|^{p-2}(A_ix - b_i)A_i\right\| \\
&\geq \max\left(\frac{p}{\|x\|}\sum_{i=1}^n |A_ix - b_i|^{p-2}(A_ix - b_i)A_ix, 0\right) \\
&= \max\left(\frac{p}{\|x\|}\sum_{i=1}^n \left[|A_ix - b_i|^{p-2}(A_ix - b_i)^2 + |A_ix - b_i|^{p-2}(A_ix - b_i)b_i\right], 0\right) \\
&\geq \max\left(\frac{p}{\|x\|}\sum_{i=1}^n \left(|A_ix - b_i|^p - |A_ix - b_i|^{p-1}|b_i|\right), 0\right),
\end{aligned}$$

now by Young's inequality $|A_ix - b_i|^{p-1}|b_i| \leq |A_ix - b_i|^p\frac{p-1}{p} + \frac{|b_i|^p}{p}$, hence

$$\geq \max\left(\frac{1}{\|x\|}\sum_{i=1}^n \left(|A_ix - b_i|^p - |b_i|^p\right), 0\right)$$

using the fact that $|a+b|^p \leq (|a|+|b|)^p = \left(\frac{2|a|+2|b|}{2}\right)^p \leq 2^{p-1}(|a|^p+|b|^p)$ by convexity (this is so-called the C_p inequality), so $|A_ix-b_i|^p + |b_i|^p \geq 2^{-p+1}|A_ix|^p$, hence

$$\begin{aligned}
&\geq \max\left(\frac{1}{\|x\|}\sum_{i=1}^n \left(2^{-p+1}|A_ix|^p - 2|b_i|^p\right), 0\right) \\
&\geq \max\left(2^{-p+1}\left[\inf_{\|s\|=1}\|As\|_p^p\right]\cdot\|x\|^{p-1} - \frac{2\sum_{i=1}^n |b_i|^p}{\|x\|}, 0\right),
\end{aligned}$$

and the lower bound follows from Assumption I by straightforward rearrangement. For the upper bound, notice that

$$\begin{aligned}
\|\nabla f(x)\| &\leq p\sup_{\|v\|=1}\sum_{i=1}^n |A_ix-b_i|^{p-1}|A_iv| \\
&\leq 2^{p-2}p\sup_{\|v\|=1}\sum_{i=1}^n \left(|A_ix|^{p-1}|A_iv| + |b_i|^{p-1}|A_iv|\right) \\
&\leq 2^{p-2}p\left[\|x\|^{p-1}\sup_{\|s\|=1,\|v\|=1}\sum_{i=1}^n \left(|A_is|^{p-1}|A_iv|\right) + \sup_{\|v\|=1}\sum_{i=1}^n |b_i|^{p-1}|A_iv|\right]
\end{aligned}$$

by Fenchel-Young, and rearrangement

$$\leq 2^{p-2}p\left[\frac{p+1}{p}\sup_{\|s\|=1}\|As\|_p^p + \frac{p-1}{p}\sum_{i=1}^n |b_i|^p\right]$$

hence the result follow. □

Lemma 5.6.6 (Bounds on the Hessian)**.** *Let $f(x) = \|Ax-b\|_p^p$ be the p-norm objective. Suppose that Assumption I holds, and let*

$$R_H = \left\|\sum_{i=1}^n |b_i|^{p-2}A_i^TA_i\right\|^{1/(p-2)} / (c_H 2^{-p})^{1/(p-2)}, \tag{5.62}$$

$$\rho_H = \inf_{\|x\|\leq R_H}\lambda_{\min}(\nabla^2 f(x)) = \inf_{\|x\|\leq 1,\|u\|=1} p(p-1)\sum_{i=1}^n |A_ix-b_i|^{p-2}(A_iu)^2. \tag{5.63}$$

Then $\rho_H > 0$, and we have

$$(L_H\|x\|^{p-2}+C_H)I \preceq \nabla^2 f(x) \preceq (U_H\|x\|^{p-2}+D_H)I \tag{5.64}$$

for all $x \in \mathbb{R}^d$, with constants

$$
\begin{aligned}
L_H &= \min\left(p(p-1)2^{-p-1}c_H, \frac{\rho_H}{2R_H^{p-2}}\right), \\
C_H &= \min\left(\frac{\rho_H}{2}, p(p-1)2^{-p-1}c_H R_H^{p-2}\right), \\
U_H &= 2^{p-3}p(p-1)\sup_{\|u\|=1,\|v\|=1}\sum_{i=1}^{n}|A_i u|^{p-2}(A_i v)^2, \\
D_H &= p(p-1)2^{p-3}\left\|\sum_{i=1}^{n}|b_i|^{p-2}A_i^T A_i\right\|.
\end{aligned}
$$

Proof. We have by differentiation

$$
\nabla^2 f(x) = p(p-1)\sum_{i=1}^{n}|A_i x - b_i|^{p-2}A_i^T A_i. \tag{5.65}
$$

Notice that using the fact that $|a-b|^{p-2} + |b|^{p-2} \geq 2^{-(p-1)}|a|^{p-2}$, we have

$$
\begin{aligned}
\nabla^2 f(x) &= p(p-1)\sum_{i=1}^{n}|A_i x - b_i|^{p-2}A_i^T A_i \\
&\succeq p(p-1)\sum_{i=1}^{n}\left(2^{-(p-1)}|A_i x|^{p-2} - |b_i|^{p-2}\right)A_i^T A_i \\
&\succeq p(p-1)2^{-(p-1)}c_H\|x\|^{p-2} - p(p-1)\left\|\sum_{i=1}^{n}|b_i|^{p-2}A_i^T A_i\right\|.
\end{aligned}
$$

Let R_H be as in (5.62), then using the above bound, we can see that for $\|x\| \geq R_H$, we have

$$
\begin{aligned}
\nabla^2 f(x) &\succeq p(p-1)2^{-p}c_H\|x\|^{p-2}I \\
&\succeq p(p-1)2^{-p-1}c_H\|x\|^{p-2} + p(p-1)2^{-p-1}c_H R_H^{p-2}.
\end{aligned} \tag{5.66}
$$

Since the minimum of the continuous function $\lambda_{\min}(\nabla^2 f(x))$ is achieved on the compact set B_{R_H}, and by the second part of Assumption I, it cannot be zero, and hence $\rho_H > 0$ and $\nabla^2 f(x) \succeq \rho_H I$ for every $x \in B_{R_H}$. The lower bound in (5.64) follows by combining this with (5.66). For the upper bound, using the inequality $|a+b|^{p-2} \leq 2^{p-3}(|a|^{p-2} + |b|^{p-2})$, we obtain that

$$
\begin{aligned}
\nabla^2 f(x) \preceq\ & p(p-1)2^{p-3}\sup_{\|s\|=1}\left\|\sum_{i=1}^{n}|A_i s|^{p-2}A_i^T A_i\right\| \cdot \|x\|^{p-2} \\
&+ p(p-1)2^{p-3}\left\|\sum_{i=1}^{n}|b_i|^{p-2}A_i^T A_i\right\|.
\end{aligned}
$$

□

Now we are ready to prove our main result in this section.

Proof of Proposition 5.6.4. First, both f and k are Legendre convex in this case. This is easy to verify for k, and evidently f is differentiable everywhere. To verify strict convexity of f, note that $\nabla^2 f(x) \succ 0$ under part two of Assumption I. Since both f and k are twice differentiable, by Lemma 5.5.2, it suffices to check that (5.59) holds for the linear convergence of Algorithm 2.1. We have by differentiation,

$$\nabla^2 k(x^*) = (1 + \|x^*\|^2)^{\frac{q-2}{2}} I + (q-2)(1 + \|x^*\|^2)^{\frac{q-4}{2}} x^* x^{*T}. \tag{5.67}$$

Now it is easy to see that for $p \in [2, \infty)$, we have $q = p/(p-1) \in (1, 2]$ and it is not difficult to verify that $\nabla^2 k$ satisfies that for all $x^* \in \mathbb{R}^d$,

$$(1 + \|x^*\|^2)^{\frac{1}{2}\frac{p-2}{p-1}} I \preceq \left[\nabla^2 k(x^*)\right]^{-1} \preceq (p-1)(1 + \|x^*\|^2)^{\frac{1}{2}\frac{p-2}{p-1}} I. \tag{5.68}$$

The claim of the theorem now follows by some straightforward rearrangement using Lemmas 5.6.5 and 5.6.6, with constants

$$\mu^* = \min\left(\frac{C_H}{2(p-1)(2+2D_G)}, \frac{L_H}{4(p-1)U_G^{(p-2)/(p-1)}}\right), \tag{5.69}$$

$$L^* = \min\left(\frac{U_H}{(L_G/2)^{(p-2)/(p-1)}}, 4U_H\left(\frac{C_G}{L_G}\right)^{(p-2)/(p-1)} + 2D_H\right). \tag{5.70}$$

□

5.7 Discussion

5.7.1 Special cases and related methods

Algorithm 2.1 is closely related to a number of existing methods, some of which are subject to the analysis we provide. The most notable of these is the method of steepest descent with respect to a given norm $\|\cdot\|$ (now not necessarily Euclidean). Here we follow the exposition of Boyd and Vandenberghe [42, sect. 4.9]. The steepest descent iteration is given by

$$x_{i+1} = x_i + \frac{1}{L}\|\nabla f(x_i)\|_* d, \qquad \text{where } d \in \arg\max_{\|x\| \le 1} \langle -\nabla f(x_i), x\rangle, \tag{5.71}$$

and $\|x^*\|_* = \sup_{\|x\| \le 1} \langle x, x^*\rangle$ is the dual norm of $\|\cdot\|$. It is possible to verify the following equivalencies for all $x^* \in \mathbb{R}^d$.

$$\begin{aligned}\|x^*\|_* \arg\max\{\langle x^*, x\rangle : \|x\| \le 1\} &= \{x : \langle x^*, x\rangle = \|x^*\|_*^2, \|x\| = \|x^*\|_*\} \\ &= \partial(\|x^*\|_*^2/2)\end{aligned} \tag{5.72}$$

where $\partial k(x^*)$ is the subdifferential of k at x^*. In this form it is clear to see that for strictly convex and differentiable $\|\cdot\|_*$, the steepest descent method (5.71) is a special case of dual preconditioned gradient descent with $k(x^*) = \|x^*\|_*^2/2$. Our analysis does not apply in the case of other norms or normalized steepest descent [42], although they may be seen as close relatives of Algorithm 2.1.

Algorithm 2.1 also generalizes some recent work in machine learning. Using the identity $\nabla k = (\nabla k^*)^{-1}$ and the fact that k^* is Legendre iff k is Legendre, the iterations of dual preconditioned gradient descent may be written as

$$x_{i+1} = x_i - \tfrac{1}{L_i}\nabla k(\nabla f(x_i)) = \operatorname*{arg\,min}_{x\in\mathbb{R}^d}\{\langle \nabla f(x_i), x\rangle + \tfrac{1}{L_i}k^*(L_i(x_i - x))\}. \tag{5.73}$$

In this form it is clear that the rescaled gradient descent method studied in [265, sect. 2.2] is a special case of Algorithm 2.1 with $k(x^*) = 2\langle x^*, B^{-1}x^*\rangle^{q/2}/q$ where B is a positive definite, self-adjoint linear operator and $1 \le q < \infty$ with $p = q/(q-1)$ an integer. [265] study the convergence of this method under smoothness conditions that require bounds on the derivatives of all orders up to p. In contrast, our analysis under dual relative smoothness requires only a relationship between the derivatives of first and second order and is applicable to rescaled gradient descent. To summarize, Algorithm 2.1 may be seen as a generalization of the steepest descent method to general convex regularizers or a generalization of polynomial rescalings of gradient descent.

Dual preconditioning is more distantly related to the dual gradient methods [249, 21]. Dual gradient methods are suitable for the following composite model.

$$\min_{x\in\mathbb{R}^d} f(x) + g(x), \tag{primal}$$

where $f : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ is proper, lsc, and strongly convex and $g : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ is proper, lsc, and convex (see [18, Chap. 12] for a more general model and review). The observation motivating the dual gradient methods is that the dual formulation,

$$\min_{x^*\in\mathbb{R}^d} f^*(x^*) + g^*(-x^*), \tag{dual}$$

admits gradient [249] and accelerated gradient methods [20], because f^* is smooth when f is strongly convex. Similarly, dual preconditioned gradient descent can be seen as a move to the dual space, in which a dual problem $k(x^*) \approx f^*(x^*) - \langle x^*, x_{\min}\rangle$ (dual to $f(x) + \delta_{x=x_{\min}}(x)$) is minimized by a Bregman gradient method. Thus, dual gradient methods and dual preconditioning are most easily applied when the dual structure is

relatively more benign to model than the primal structure, e.g., when f has super-quadratic growth (and thus f_c^* has sub-quadratic growth). Both of the applications considered in this paper are of this kind. However, the two methods differ in terms of what is assumed to be cheap to compute; dual gradient methods assume that it is cheap to find points in $\partial f^*(x^*)$, whereas dual preconditioning explicitly avoids this with two ideas: by designing a dual objective function $k(x^*)$ with a minimum at 0 whose gradient map is cheap to compute and by using f^* as the "reference function" in a dual Bregman gradient scheme. When f is Legendre convex and k is relatively smooth in the dual space to f^*, the primal iterates are cheap to compute and the analysis over the dual iterates closely follows recent work on relative smoothness [16, 156, 244].

5.7.2 Conclusions

In this paper we introduced a non-linear preconditioning scheme for gradient descent on Legendre convex functions f that converges under generalizations of the standard Lipschitz assumption on ∇f. There are at least two interpretations of this method. The first is as a generalization of gradient descent in which the update direction is preconditioned by the gradient map ∇k of a designed dual reference, Legendre convex function k. The second interpretation is as a Bregman gradient method in the dual space, which minimizes the designed k while the conjugate f^* plays the role of the "reference function", see section 5.7.1. The choice of k affects the conditioning of our method, which is made explicit in our analysis through a dual relative smoothness condition between f and k. The dual relative conditions admit non-smooth f and k, and are provably distinct dual cousins of the relative smoothness conditions introduced by [16]. In the first interpretation of dual preconditioning, dual relative smoothness is as a requirement that $\nabla k \circ \nabla f$ is Lipschitz continuous. In the second, k serves as a model of the convex conjugates f^* in a certain problem class. In section 5.6, we show how this method can be applied to exponential penalty functions (see, e.g., [60, 59]) and p-norm regression (see [48, 2] and references therein) with global convergence rate guarantees.

There are natural questions that arise from this work. First, it may be useful to pursue the analogy with dual gradient methods further and to design methods for the general composite model (dual) that exploit dual relative smoothness. Second, it is natural to wonder whether dual relative smoothness can be exploited by an accelerated method, which should be optimal in the class of functions dual smooth relative to a fixed k. [156] raised this question for primal relative smoothness, and

Bregman methods converging at accelerated rates under primal relative smoothness have been designed [113, 112]. Unfortunately, recent work suggests that it is not possible to accelerate mirror descent under the generalized relative conditions [71].

Finally, some caution is warranted. There is no free lunch and the central difficulty of this method is in the design of k. Still, the dual relative conditions studied in this work provide new avenues for improving the conditioning of optimizers via hard-won domain-specific knowledge.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

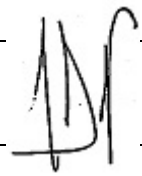
Title of Paper	Dual Space Preconditioning for Gradient Descent
Publication Status	□Published □ Accepted for Publication XSubmitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Chris J. Maddison*, Daniel Paulin*, Yee Whye Teh, Arnaud Doucet. Dual Space Preconditioning for Gradient Descent. In *Review*, 2019.

Student Confirmation

Student Name:	Chris J. Maddison		
Contribution to the Paper	• I proposed the application of relative smoothness and strong convexity for non-linear preconditioning of gradient descent. • I proved all of the convergence results in the paper, with the exception of the adaptive step-size analysis. • I proposed the application to p-norm regression, but Daniel Paulin developed both of the applications and is responsible for all of the analysis in the applications section. • I wrote the entire paper, with some help from Daniel to incorporate his analyses. • All authors contributed to the development of the paper through discussions and ideas, and all authors reviewed the final draft.		
Signature 		Date	05 May 2020

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Arnaud Doucet			
Supervisor comments I agree that the candidate has made a substantial contribution to the publication.			
Signature 		Date	05 May 2020

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 6

Conclusions

We introduced four new methods that address distinct parts of the machine learning workflow: designing objectives, computing gradients, and gradient-based optimization. Future work is discussed in the conclusion sections of each chapter. To conclude, we review the major contributions in turn.

In Chapter 2 we introduced biased gradient estimators for stochastic optimization in the presence of discrete random variables. This is a challenging problem, and the previous unbiased methods generally suffer from high variance. We took a different tack, introducing bias to decrease variance. Our estimators are reparameterization-based, but of a relaxed approximation to the discrete model. The primary advantage of such estimators is that they are easy to implement and computationally inexpensive. We showed improved performance on multiple benchmarks in statistical deep learning.

In Chapter 3 we introduced variational objectives for sequential models, called filtering variational objectives (FIVOs). FIVO is a lower bound on the marginal log-likelihood, constructed by taking the logarithm of a particle filter's estimator for the normalizing constant of the model. We showed that the tightness of FIVO is related to the variance of the estimator from which it is constructed, suggesting that improved bounds may be designed from more sophisticated normalizing constant estimators. We developed a reparameterization scheme for estimating gradients of FIVO. We showed that optimizing FIVO uniformly outperforms baseline objectives on a variety of deep learning benchmarks.

In Chapter 4 we studied the impact of the choice of kinetic energy on the convergence of mometum-based optimizers. We studied conformal Hamiltonian systems, which we called Hamiltonian descent systems, and showed that the kinetic energy acts as a non-linear preconditioner. We studied Hamiltonian descent systems in continuous time and presented three discretization schemes. The choice of kinetic energy in all of these schemes modifies the conditioning of the system to allow linear convergence on

strictly convex functions, even for functions that may be non-smooth or non-strongly convex. In a rough sense, the conditions that we study require that the kinetic energy approximate the curvature of the convex conjugate of the function of interest, a generalization of strong convexity and smoothness. We proved partial lower bounds for simple one-dimensional power functions, showing that when these conditions are not satisfied the Hamiltonian descent system fails to converge linearly. We designed and analyzed kinetic energies for functions with power growth.

In Chapter 5 we presented a non-linear preconditioning scheme for gradient descent. This scheme is closely related to the work of Chapter 4, but its analysis is considerably simpler. We called our scheme dual space preconditioning, and showed that it converges under conditions (related to the conditions studied for Hamiltonian descent) that are related to recent work on relative smoothness and strong convexity [16, 156]. We presented two applications to p-norm regression and exponential barrier functions. We showed how our conditions may be satisfied in practice on these examples.

The central theme of this thesis is on the interplay between the problems of numerical optimization and numerical integration. Although apparently distinct problems, their methodology, analysis, and study are complementary, and there is a productive research agenda at their intersection.

Appendix A

Appendix to Hamiltonian Descent

Ap.1 Proofs for convergence of continuous systems

Lemma 4.3.6 (Convergence rates in continuous time for fixed α)**.** *Given $\gamma \in (0,1)$, $f : \mathbb{R}^d \to \mathbb{R}$ differentiable and convex with unique minimum $x_{\min}$, $k : \mathbb{R}^d \to \mathbb{R}$ differentiable and strictly convex with minimum $k(0) = 0$. Let $x_t, p_t \in \mathbb{R}^d$ be the value at time t of a solution to the system (4.3) such that there exists $\alpha \in (0,1]$ where $k(p_t) \geq \alpha f_c^*(-p_t)$. Define*

$$\lambda(\alpha,\beta,\gamma) = \min\left(\frac{\alpha\gamma - \alpha\beta - \beta\gamma}{\alpha - \beta}, \frac{\beta(1-\gamma)}{1-\beta}\right). \tag{4.19}$$

If $\beta \in (0, \min(\alpha,\gamma)]$, then

$$\mathcal{V}_t' \leq -\lambda(\alpha,\beta,\gamma)\mathcal{V}_t.$$

Finally,

1. *The optimal $\beta \in (0, \min(\alpha,\gamma)]$, $\beta^\star = \arg\max_\beta \lambda(\alpha,\beta,\gamma)$ and $\lambda^\star = \lambda(\alpha,\beta^\star,\gamma)$ are given by,*

$$\beta^\star = \tfrac{1}{1+\alpha}\left(\alpha + \tfrac{\gamma}{2} - \sqrt{(1-\gamma)\alpha^2 + \tfrac{\gamma^2}{4}}\right), \tag{4.20}$$

$$\lambda^\star = \begin{cases} \tfrac{1}{1-\alpha}\left((1-\gamma)\alpha + \tfrac{\gamma}{2} - \sqrt{(1-\gamma)\alpha^2 + \tfrac{\gamma^2}{4}}\right) & \text{for } 0 < \alpha < 1, \\ \tfrac{\gamma(1-\gamma)}{2-\gamma} & \text{for } \alpha = 1, \end{cases} \tag{4.21}$$

2. *If $\beta \in (0, \alpha\gamma/2]$, then*

$$\lambda(\alpha,\beta,\gamma) = \frac{\beta(1-\gamma)}{1-\beta}, \text{ and} \tag{4.22}$$

$$\begin{aligned} &-\left(\gamma - \beta - \gamma^2(1-\gamma)/4\right) k(p_t) - \beta\gamma\left\langle x_t - x_{\min}, p_t\right\rangle - \beta\left\langle x_t - x_{\min}, \nabla f(x_t)\right\rangle \\ &\leq -\beta(1-\gamma)(k(p_t) + f(x_t) - f(x_{\min}) + \beta\left\langle x_t - x_{\min}, p_t\right\rangle). \end{aligned} \tag{4.23}$$

Proof.

$$\begin{aligned}\mathcal{V}_t{}' &= -\gamma \langle \nabla k(p_t), p_t \rangle + \beta \langle \nabla k(p_t), p_t \rangle - \beta\gamma \langle x_t - x_{\min}, p_t \rangle - \beta \langle x_t - x_{\min}, \nabla f(x_t) \rangle \\ &= -(\gamma - \beta) \langle \nabla k(p_t), p_t \rangle - \beta\gamma \langle x_t - x_{\min}, p_t \rangle - \beta \langle x_t - x_{\min}, \nabla f(x_t) \rangle \\ &\leq -(\gamma - \beta) k(p_t) - \beta\gamma \langle x_t - x_{\min}, p_t \rangle - \beta(f(x_t) - f(x_{\min}))\end{aligned}$$

by convexity and $\beta \leq \gamma$. Our goal is to show that $\mathcal{V}_t{}' \leq -\lambda \mathcal{V}_t$ for some $\lambda > 0$, which would hold if

$$\begin{aligned}-(\gamma - \beta) k(p_t) - \beta\gamma \langle x_t - x_{\min}, p_t \rangle - \beta(f(x_t) - f(x_{\min})) \leq \\ - \lambda(k(p_t) + f(x_t) - f(x_{\min}) + \beta \langle x_t - x_{\min}, p_t \rangle)\end{aligned}$$

which is equivalent by rearrangement to

$$-\beta(\gamma - \lambda) \langle x_t - x_{\min}, p_t \rangle \leq (\gamma - \beta - \lambda) k(p_t) + (\beta - \lambda)(f(x_t) - f(x_{\min})). \quad \text{(Ap.1)}$$

Assume that $\lambda \leq \gamma$. By assumption on f, k, and α we have (4.15), which implies by rearrangement that $k(p_t) \geq -\alpha \langle x_t - x_{\min}, p_t \rangle - \alpha(f(x_t) - f(x_{\min}))$, so

$$-\beta(\gamma - \lambda) \langle x_t - x_{\min}, p_t \rangle \leq \frac{\beta}{\alpha}(\gamma - \lambda)(k(p_t) + \alpha(f(x_t) - f(x_{\min}))), \quad \text{(Ap.2)}$$

and $k(p_t) \geq 0$ and $f(x_t) - f(x_{\min}) \geq 0$, hence it is enough to have $\frac{\beta}{\alpha}(\gamma - \lambda) \leq \gamma - \beta - \lambda$ and $\beta(\gamma - \lambda) \leq \beta - \lambda$ for showing (Ap.1). Thus we need $\lambda \leq \min(\gamma, \frac{\alpha\gamma - \alpha\beta - \beta\gamma}{\alpha - \beta}, \frac{\beta(1-\gamma)}{1-\beta})$. Here $\frac{\beta}{1-\beta}(1 - \gamma) \leq \gamma$ for $0 < \beta \leq \gamma < 1$, therefore $\mathcal{V}_t{}' \leq -\lambda(\alpha, \beta, \gamma)\mathcal{V}_t$ for

$$\lambda(\alpha, \beta, \gamma) = \min\left(\frac{\alpha\gamma - \alpha\beta - \beta\gamma}{\alpha - \beta}, \frac{\beta(1 - \gamma)}{1 - \beta}\right).$$

In order to obtain the optimal contraction rate, we need to maximize $\lambda(\alpha, \beta, \gamma)$ in β. Without loss of generality, we can assume that $0 < \beta < \frac{\alpha\gamma}{\alpha+\gamma}$, and it is easy to see that on this interval, $\frac{\alpha\gamma - \alpha\beta - \beta\gamma}{\alpha - \beta}$ is strictly monotone decreasing, while $\frac{\beta(1-\gamma)}{1-\beta}$ is strictly monotone increasing. Therefore, the maximum will be taken when the two terms are equal. This leads to a quadratic equation with two solutions

$$\beta_\pm = \frac{1}{1 + \alpha}\left(\alpha + \frac{\gamma}{2} \pm \sqrt{(1 - \gamma)\alpha^2 + \frac{\gamma^2}{4}}\right).$$

One can check that $\beta_+ > \frac{\alpha\gamma}{\alpha+\gamma}$, while $0 < \beta_- < \frac{\alpha\gamma}{\alpha+\gamma}$, hence

$$\max_{\beta \in [0,\alpha]} \lambda(\alpha, \beta, \gamma) = \lambda(\alpha, \beta_-, \gamma) = \frac{1}{1 - \alpha}\left((1 - \gamma)\alpha + \frac{\gamma}{2} - \sqrt{(1 - \gamma)\alpha^2 + \frac{\gamma^2}{4}}\right)$$

for $\alpha < 1$. For $\alpha = 1$, we obtain that $\beta^\star = \beta_- = \frac{\gamma}{2}$, and $\lambda^\star = \frac{\gamma(1-\gamma)}{2-\gamma}$.

Now assume $\beta \in (0, \alpha\gamma/2]$. Since we have shown that $\lambda(\alpha, \gamma, \beta) = \frac{\beta(1-\gamma)}{1-\beta}$ for $\beta < \beta_-$ it is enough to show that $\beta_- > \alpha\gamma/2$ to get our result. Notice that β_- as a function of γ, $\beta_-(\gamma)$, is strictly concave with $\beta_-(0) = 0$, and $\beta_-(1) = \frac{\alpha}{1+\alpha}$, thus

$$\beta_- = \beta_-(\gamma) > \gamma\beta_-(1) = \frac{\gamma\alpha}{1+\alpha} \geq \frac{\alpha\gamma}{2} \geq \beta.$$

Finally, the proof of (4.23) is equivalent by rearrangement to showing that for $\lambda = (1-\gamma)\beta$,

$$-\beta(\gamma-\lambda)\langle x_t - x_{\min}, p_t\rangle \leq (\gamma - \beta - \gamma^2(1-\gamma)/4 - \lambda)k(p_t) + (\beta-\lambda)(f(x_t) - f(x_{\min})), \tag{Ap.3}$$

hence by (Ap.2) it suffices to show that we have $\frac{\beta}{\alpha}(\gamma-\lambda) \leq \gamma - \gamma^2(1-\gamma)/4 - \beta - \lambda$ and $\beta(\gamma-\lambda) \leq \beta - \lambda$. The latter one was already verified in the previous section, and the first one is equivalent to

$$\gamma - \beta - \lambda - \frac{\beta}{\alpha}(\gamma-\lambda) \geq \gamma^2(1-\gamma)/4 \text{ for every } 0 < \gamma < 1,\ 0 < \alpha \leq 1,\ 0 < \beta \leq \alpha\gamma/2.$$

It is easy to see that we only need to check this for $\beta = \alpha\gamma/2$, and in this case by minimizing the left hand side for $0 \leq \alpha \leq 1$ and using the fact that $\lambda = (1-\gamma)\beta$, we obtain the claimed result. □

Ap.2 Proofs for partial lower bounds

In this section, we present the proofs of the lower bounds. First, we show the existence and uniqueness of solutions.

Lemma 4.3.9 (Existence and uniqueness of solutions of the ODE)**.** *Let $a, b, \gamma \in (0, \infty)$. For every $t_0 \in \mathbb{R}$ and $(x, p) \in \mathbb{R}^2$, there is a unique solution $(x_t, p_t)_{t\in\mathbb{R}}$ of the ODE (4.32) with $x_{t_0} = x$, $p_{t_0} = p$. Either $x_t = p_t = 0$ for every $t \in \mathbb{R}$, or $(x_t, p_t) \neq (0, 0)$ for every $t \in \mathbb{R}$.*

Proof. Let $\mathcal{H}_t := \frac{|x_t|^b}{b} + \frac{|p_t|^a}{a}$, then $\mathcal{H}_t \geq 0$ and

$$\mathcal{H}_t' = |x_t|^{b-1}\operatorname{sign}(x_t)x_t' + |p_t|^{a-1}\operatorname{sign}(p)p_t' = -\gamma|p_t|^a,$$

so $0 \geq \mathcal{H}_t' \geq -\gamma a\mathcal{H}_t$. By Grönwall's inequality, this implies that for any solution of (4.3),

$$\mathcal{H}_t \leq \mathcal{H}_0 \text{ for } t \geq 0, \text{ and} \tag{Ap.4}$$

$$\mathcal{H}_t \leq \mathcal{H}_0 \exp(-\gamma a t) \text{ for } t < 0. \tag{Ap.5}$$

The derivatives x', p' are continuous functions of (x, p), and these functions are locally Lipschitz if $x \neq 0$ and $p \neq 0$. So by the Picard-Lindelöf theorem, if $x_{t_0} \neq 0$, $p_{t_0} \neq 0$, then there exists a unique solution in the interval $(t_0 - \epsilon, t_0 + \epsilon)$ for some $\epsilon > 0$.

Now we will prove local existence and uniqueness for $(x_{t_0}, p_{t_0}) = (x_0, 0)$ with $x_0 \neq 0$, and for $(x_{t_0}, p_{t_0}) = (0, p_0)$ with $p_0 \neq 0$. Because of the central symmetry, we may assume that $x_0 > 0$ and $p_0 > 0$, and we may also assume that $t_0 = 0$.

First let $x_0 = 0$ and $p_0 > 0$. We take t close enough to 0 so that $p_t > 0$. Then $x_t' = p_t^{a-1} > 0$ and $p_t' = -|x_t|^{b-1} \operatorname{sign}(x_t) - \gamma p_t$. Then $p_t = \phi(x_t)$ for some function $\phi\colon (-\epsilon, \epsilon) \to \mathbb{R}_{>0}$, where t is close enough to 0. Here $\phi(0) = p_0$ and $p_t' = \phi'(x_t)x_t'$, so

$$\begin{aligned}\phi'(x_t) = \frac{p_t'}{x_t'} &= -(|x_t|^{b-1} \operatorname{sign}(x_t) + \gamma p_t) p_t^{1-a} \\ &= -(|x_t|^{b-1} \operatorname{sign}(x_t) + \gamma \phi(x_t)) \phi(x_t)^{1-a},\end{aligned}$$

and hence

$$\phi'(u) = -(|u|^{b-1} \operatorname{sign}(u) + \gamma \phi(u)) \phi(u)^{1-a} \quad \text{and} \quad \phi(0) = p_0.$$

This ODE satisfies the conditions of the Picard-Lindelöf theorem, so ϕ exists and is unique in a neighborhood of 0. Then $x_0 = 0$ and $x_t' = \phi(x_t)^{a-1}$, so for the Picard-Lindelöf theorem we just need to check that $u \mapsto \phi(u)^{a-1}$ is Lipschitz in a neighborhood of 0. This is true, because ϕ is C^1 in a neighborhood of 0. So x_t exists and is unique when t is near 0, hence $p_t = \phi(x_t)$ also exists and is unique there.

Now let $x_0 = x_0 > 0$ and $p_0 = 0$. We take t close enough to 0 so that $x_t > 0$. Then $x_t' = |p_t|^{a-1} \operatorname{sign}(p_t)$ and $p_t' = -x_t^{b-1} - \gamma p_t < 0$ for t close enough to 0. Then $x_t = \psi(p_t)$ for some function $\psi\colon (-\epsilon, \epsilon) \to \mathbb{R}_{>0}$, where t is close enough to 0. Here $\psi(0) = x_0$ and $x_t' = \phi'(p_t)p_t'$, so

$$\psi'(p_t) = \frac{x_t'}{p_t'} = -\frac{|p_t|^{a-1} \operatorname{sign}(p_t)}{x_t^{b-1} + \gamma p_t} = -\frac{|p_t|^{a-1} \operatorname{sign}(p_t)}{\psi(p_t)^{b-1} + \gamma p_t},$$

and thus

$$\psi'(u) = -\frac{|u|^{a-1} \operatorname{sign}(u)}{\psi(u)^{b-1} + \gamma u} \quad \text{and} \quad \psi(0) = x_0.$$

This ODE satisfies the conditions of the Picard-Lindelöf theorem, so ψ exists and is unique in a neighborhood of 0. Then $p_0 = 0$ and $p_t' = -\psi(p_t)^{b-1} - \gamma p_t$, so for the Picard-Lindelöf theorem we just need to check that $u \mapsto -\psi(u)^{b-1} - \gamma u$ is Lipschitz in a neighborhood of 0. This is true, because ψ is C^1 in a neighborhood of 0. So p_t exists and is unique when t is near 0, hence $x_t = \psi(p_t)$ also exists and is unique there.

Let $[0, t_{\max})$ and $(-t_{\min}, 0]$ be the longest intervals where the solution exists and unique. If $t_{\max} < \infty$ or $t_{\min} < \infty$, then by Theorem 3 of [202], page 91, the solution would have to be able to leave any compact set K in the interval $[0, t_{\max})$ or $(-t_{\max}, 0]$, respectively. However, due to the (Ap.4) and (Ap.5), the energy function $\mathcal{H}_t$ cannot converge to infinity in finite amount of time, so this is not possible. Hence, the existence and uniqueness for every $t \in \mathbb{R}$ follows. □

Before proving Theorem 4.3.10, we need to show a few preliminary results.

Lemma Ap.2.1. *If (x, p) is not constant zero, then $\lim_{t\to-\infty} \mathcal{H}_t = \infty$ and $\lim_{t\to\infty} \mathcal{H}_t = \lim_{t\to\infty} x_t = \lim_{t\to\infty} p_t = 0$.*

Proof. The limits of $\mathcal{H}$ exist, because $\mathcal{H}'_t = -\gamma|p_t|^a \leq 0$. First suppose that $\lim_{t\to-\infty} \mathcal{H}_t = M < \infty$. Then $\mathcal{H}_t \leq M$ for every t, so x and p are bounded functions, therefore x' and p' are also bounded by the differential equation. Then $\mathcal{H}''_t = -\gamma a|p_t|^{a-1}\operatorname{sign}(p_t)p'_t$ is also bounded, so $\mathcal{H}'$ is Lipschitz. This together with $\lim_{t\to-\infty} \mathcal{H}_t = M \in \mathbb{R}$ implies that $\lim_{t\to-\infty} \mathcal{H}'_t = 0$. So $\lim_{t\to-\infty} p_t = 0$. Then we must have $\lim_{t\to-\infty} x_t = x_0$ for some $x_0 \in \mathbb{R} \setminus \{0\}$. But then $\lim_{t\to-\infty} p'_t = -|x_0|^{b-1}\operatorname{sign}(x_0) \neq 0$, which contradicts $\lim_{t\to-\infty} p_t = 0$. So indeed $\lim_{t\to-\infty} \mathcal{H}_t = \infty$.

Now suppose that $\lim_{t\to\infty} \mathcal{H}_t > 0$. For $t \in [0, \infty)$ we have $\mathcal{H}_t \leq \mathcal{H}_0$, so for $t \geq 0$ the functions x and p are bounded, hence also x', p', $\mathcal{H}'$, $\mathcal{H}''$ are bounded there. So $\lim_{t\to\infty} \mathcal{H}_t \in \mathbb{R}$, and H' is Lipschitz for $t \geq 0$, therefore $\lim_{t\to\infty} \mathcal{H}'_t = 0$, thus $\lim_{t\to\infty} p_t = 0$. Then we must have $\lim_{t\to\infty} x_t = x_0$ for some $x_0 \in \mathbb{R} \setminus \{0\}$. But then $\lim_{t\to\infty} p'_t = -|x_0|^{b-1}\operatorname{sign}(x_0) \neq 0$, which contradicts $\lim_{t\to\infty} p_t = 0$. So indeed $\lim_{t\to\infty} \mathcal{H}_t = 0$, thus $\lim_{t\to\infty} x_t = \lim_{t\to\infty} p_t = 0$. □

From now on we assume that $\frac{1}{a} + \frac{1}{b} < 1$.

Lemma Ap.2.2. *If (x_t, p_t) is a solution, then for every $t_0 \in \mathbb{R}$ there is a $t \leq t_0$ such that $p_t = 0$.*

Proof. The statement is trivial for the constant zero solution, so assume that (x, p) is not constant zero. Then $\lim_{t\to-\infty} \mathcal{H}_t = \infty$. Suppose indirectly that $p_t < 0$ for every $t \leq t_0$. Then $x'_t = -|p_t|^{a-1} < 0$ for $t \leq t_0$. If $\lim_{t\to-\infty} x_t = x_{-\infty} < \infty$, then $\lim_{t\to-\infty} p_t = -\infty$, because $\lim_{t\to-\infty} \mathcal{H}_t = \infty$. Then $x \to x_{-\infty} \in \mathbb{R}$ and $x' = -|p|^{a-1} \to -\infty$ when $t \to -\infty$, which is impossible. So $\lim_{t\to-\infty} x_t = \infty$, hence there is a $t_1 \leq t_0$ such that $p_t < 0 < x_t$ for every $t \leq t_1$. Let $G_t := \frac{|p_t|^{a-1}}{a-1} - \gamma x_t$, then for $t \leq t_1$ we have

$$G'_t = -|p_t|^{a-2}p'_t - \gamma x'_t = |p_t|^{a-2}(x_t^{b-1} - \gamma|p_t|) + \gamma|p_t|^{a-1} = |p_t|^{a-2}x_t^{b-1} > 0.$$

So $G_t \le G(t_1)$ for every $t \le t_1$. Thus

$$|p_t| \le ((a-1)(\gamma x_t + G(t_1)))^{\frac{1}{a-1}} = (Ax_t + B)^{\frac{1}{a-1}},$$

for $t \le t_1$, where $A > 0$. For big enough x we have $(Ax+B)^{\frac{1}{a-1}} < \frac{1}{\gamma}x^{b-1}$, because $\frac{1}{a-1} < b-1$, since $ba > b+a$. So there is a $t_2 \le t_1$ such that $p_t < 0 < x_t$ and $|p_t| \le \frac{1}{\gamma}x_t^{b-1}$ for every $t \le t_2$. Then $p_t' = -x_t^{b-1} - \gamma p_t \le 0$, so $p_t < 0$ is monotone decreasing for $t \in (-\infty, t_2]$, hence $p_{-\infty} = \lim_{t\to-\infty} p_t \in \mathbb{R}$. But then $p_t' = -x_t^{b-1} - \gamma p_t \to -\infty$ when $t \to -\infty$, which together with $p_{-\infty} \in \mathbb{R}$ is impossible. This contradiction shows that indeed there is a $t \le t_0$ such that $p_t \ge 0$. Applying this for $(-x, -p)$, we get that there is a $t \le t_0$ such that $p_t \le 0$. So by continuity, there is a $t \le t_0$ such that $p_t = 0$. □

For $A > \frac{1}{\gamma}$ let $\xi(A) := (\frac{\gamma A - 1}{(b-1)A^a})^{\frac{1}{ba-b-a}}$ and

$$\mathcal{R}_A := \{(x,p) \in \mathbb{R}^2;\ 0 < x < \xi(A),\ -Ax^{b-1} < p < 0\}, \tag{Ap.6}$$

Lemma Ap.2.3. *Let $A > \frac{1}{\gamma}$. If (x,p) is a solution, $t_0 \in \mathbb{R}$ and $(x_{t_0}, p_{t_0}) \in \mathcal{R}_A$, then $(x_t, p_t) \in \mathcal{R}_A$ for every $t \ge t_0$.*

Proof. Suppose indirectly that there is a $t > t_0$ such that $(x_t, p_t) \notin \mathcal{R}_A$. Let T be the infimum of these t's. Then $T > t_0$, and $(x(T), p(T))$ is on the boundary of the region $\mathcal{R}_A$. We cannot have $(x(T), p(T)) = (0,0)$, because $(0,0)$ is unreachable in finite time. Since $x_t' = -|p_t|^{a-1} \le 0$ for $t \in [t_0, T)$, we have $x(T) \le x_{t_0} < \xi(A)$. So we have $0 < x(T) < \xi(A)$ and either $p(T) = 0$ or $p(T) = -Ax(T)^{b-1}$.

Suppose that $p(T) = 0$. Then $p'(T) = -x(T)^{b-1} < 0$, so if $t \in (t_0, T)$ is close enough to T, then $p_t > 0$, which contradicts $(x_t, p_t) \in \mathcal{R}_A$. So $p(T) = -Ax(T)^{b-1}$. Let $U_t := p_t + Ax_t^{b-1}$. Then $U(T) = 0$, and by the definition of T, we must have $U'(T) \le 0$. Using $|p(T)| = Ax(T)^{b-1}$ we get

$$\begin{aligned} 0 \ge U'(T) &= p'(T) + (b-1)Ax(T)^{b-2}x'(T) \\ &= \gamma|p(T)| - x(T)^{b-1} - (b-1)Ax(T)^{b-2}|p(T)|^{a-1} \\ &= x(T)^{b-1}(\gamma A - 1 - (b-1)A^a x(T)^{ba-b-a}), \end{aligned}$$

so $\xi(T) = (\frac{\gamma A-1}{(b-1)A^a})^{\frac{1}{ba-b-a}} \le x(T) < \xi(T)$. This contradiction proves that $(x_t, p_t) \in \mathcal{R}_A$ for every $t \ge t_0$. □

The following lemma characterises the paths of every solution of the ODE in terms of single parameter θ.

Lemma Ap.2.4. *There is a constant $\eta > 0$ such that every solution (x_t, p_t) which is not constant zero is of the form $(x^{(\theta)}_{t+\Delta}, p^{(\theta)}_{t+\Delta})$ for exactly one $\theta \in [-\eta, \eta] \setminus \{0\}$ and $\Delta \in \mathbb{R}$.*

Proof. For $u > 0$ let us take the solution (x_t, p_t) with $(x_{t_0}, p_{t_0}) = (-u, 0)$ for some $t_0 \in \mathbb{R}$. Then $p'_{t_0} = u^{b-1} > 0$, so $p_{t_0} < 0$ if $t < t_0$ is close enough to t_0. By Lemma Ap.2.2, there is a smallest $\mathcal{T}(u) \in \mathbb{R}_{>0}$ such that $p(t_0 - \mathcal{T}(u)) = 0$. We may take $t_0 = \mathcal{T}(u)$, and call this solution $(X^{(u)}(t), P^{(u)}(t))$. Then $X^{(u)}_{\mathcal{T}(u)} = -u$, $P^{(u)}_{\mathcal{T}(u)} = P^{(u)}_0 = 0$, and $P^{(u)}_t < 0$ for $t \in (0, \mathcal{T}(u))$. Let $g(u) = X^{(u)}_0$. Here $g(u) \neq 0$, because we cannot reach $(0, 0)$ in finite time due to (Ap.4)-(Ap.5). We cannot have $g(u) < 0$, because then $P'_u(0) = -|g(u)|^{b-1} \operatorname{sign}(g(u)) > 0$. So $g(u) > 0$, thus we have defined a function $g\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$. For $u > 0$ let us take the continuous path

$$\mathcal{C}_u := \{(X^{(u)}_t, P^{(u)}_t);\, t \in [0, \mathcal{T}(u)]\}.$$

Note that this path is below the x-axis except for the two endpoints, which are on the x-axis. If $0 < u < v$, then $\mathcal{C}_u \cap \mathcal{C}_v = \varnothing$, so we must have $g(u) < g(v)$ (otherwise the two paths would have to cross). So g is strictly increasing. Let

$$\eta := \lim_{u \to 0} g(u) \in \mathbb{R}_{\geq 0}.$$

If $0 < u < v$ and $z \in (g(u), g(v))$, then going forward in time after the point $(z, 0)$, the solution must intersect the x-axis first somewhere between the points $(-v, 0)$ and $(-u, 0)$, thus z is in the image of η. So $\eta\colon \mathbb{R}_{>0} \to (\eta, \infty)$ is a strictly increasing bijective function. We have $g(u) > u$ for every $u > 0$, because if $g(u) \leq u$, then for the solution $(X^{(u)}_t, P^{(u)}_t)$ we have $\mathcal{H}_0 \leq H(\mathcal{T}(u))$ and $\mathcal{H}'_t = -\gamma |P^{(u)}_t|^a < 0$ for $t \in (0, \mathcal{T}(u))$, which is impossible.

Let $A > \frac{1}{\gamma}$ and $0 < z < \xi(A)$, and take the solution (x_t, p_t) with $(x_0, p_0) = (z, 0)$. Then $p'_0 < 0$, so $(x_t, p_t) \in \mathcal{R}_A$ for $t > 0$ close enough to 0, and then by Lemma Ap.2.3, $(x_t, p_t) \in \mathcal{R}_A$ for every $t > 0$. So z is not in the image of η, hence $\eta \geq \xi(A) > 0$.

Let (x_t, p_t) be a solution, which is not constant zero. Let $\mathcal{S} = \{t \in \mathbb{R};\, p_t = 0\}$. This is a closed, nonempty subset of $\mathbb{R}$. Suppose that $\sup(\mathcal{S}) = \infty$. Since $\lim_{t \to \infty} x_t = 0$, this means that there are infinitely many $t \in \mathbb{R}$ such that $p_t = 0$ and $|x_t| \in (0, \eta)$. This is impossible, since there can be only one such t. So $\sup(\mathcal{S}) = \max(\mathcal{S}) = T \in \mathbb{R}$. We may translate time so that $T = 0$. Then $p_0 = 0$ and $p_t \neq 0$ for every $t > 0$. If $|x_0| > \eta$, then later we again intersect the x-axis, so we must have $0 < |x_0| \leq \eta$. So the not constant zero solutions can be described by their last intersection with the x-axis, and this intersection has its x-coordinate in $[-\eta, \eta] \setminus \{0\}$. □

By symmetry, we have $x_t^{(-\theta)} = -x_t^{(\theta)}$ and $p_t^{(-\theta)} = -p_t^{(\theta)}$. We now study the solutions $(x_t^{(\theta)}, p_t^{(\theta)})$ for $0 < \theta \leq \eta$ and $t \geq 0$. Then $p_t^{(\theta)} < 0 < x_t^{(\theta)}$ for $t \geq 0$ and $(x^{(\theta)})'_t < 0$ and $\lim_{t\to\infty} x_t^{(\theta)} = 0$, so $x_t^{(\theta)} \in (0, \theta)$ for $t > 0$. So we can write $p_t^{(\theta)} = -\phi^{(\theta)}(x_t^{(\theta)})$, where $\phi^{(\theta)} \colon (0, \theta) \to \mathbb{R}_{>0}$ and $\lim_{z\to 0} \phi^{(\theta)}(z) = \lim_{z\to\theta} \phi^{(\theta)}(z) = 0$. Let

$$\Theta := \left\{ \theta \in (0, \eta];\ (z, -\phi^{(\theta)}(z)) \in \mathcal{R}_A \text{ for some } A > \frac{1}{\gamma} \text{ and } z \in (0, \theta) \right\}.$$

The following lemmas characterize this set.

Lemma Ap.2.5. *If $\theta \in (0, \eta] \setminus \Theta$, then*

$$\lim_{z\to 0} \frac{\phi^{(\theta)}(z)}{(\gamma(a-1)z)^{\frac{1}{a-1}}} = 1.$$

Proof. By the definition of $\phi^{(\theta)}$,

$$-(x_t^{(\theta)})^{b-1} + \gamma\phi^{(\theta)}(x_t^{(\theta)}) = (p_t^{(\theta)})' = (\phi^{(\theta)})'(x_t^{(\theta)})\phi^{(\theta)}(x_t^{(\theta)})^{a-1}, \text{ so}$$
$$(\phi^{(\theta)})'(z) = \phi^{(\theta)}(z)^{1-a}(\gamma\phi^{(\theta)}(z) - z^{b-1}).$$

Because the orbits are disjoint for different θ's, we have $\phi^{(\theta_1)}(z) < \phi^{(\theta_2)}(z)$ for $0 < \theta_1 < \theta_2 \leq \eta$ and $z \in (0, \theta_1)$. If $(z_0, -\phi^{(\theta)}(z_0)) \in \mathcal{R}_A$ for some $z_0 \in (0, \theta)$, then by Lemma Ap.2.3, $(z, -\phi^{(\theta)}(z)) \in \mathcal{R}_A$ for every $z \in (0, z_0]$. So

$$\Theta = \left\{ \theta \in (0, \eta];\ \liminf_{z\to 0} z^{1-b}\phi^{(\theta)}(z) < \infty \right\}.$$

If $0 < \theta_1 < \theta_2$ and $\theta_2 \in \Theta$, then $\theta_1 \in \Theta$ too, since $\phi^{(\theta_1)}(z) < \phi^{(\theta_2)}(z)$ for $z \in (0, \theta_1)$. If $A > \frac{1}{\gamma}$ and $\theta < \xi(A)$, then $(x_t^{(\theta)}, p_t^{(\theta)}) \in \mathcal{R}_A$ for $t > 0$, so $(z, -\phi^{(\theta)}(z)) \in \mathcal{R}_A$ for $z \in (0, \theta)$. So $(0, \xi(A)) \subseteq \Theta$ for every $A > \frac{1}{\gamma}$. Let

$$F(z) := \gamma^{-1}(a-1)^{-1}\phi^{(\theta)}(z)^{a-1} - z,$$

then $\lim_{z\to 0} F(z) = 0$, and

$$F'(z) = \frac{(\phi^{(\theta)})'(z)}{\gamma\phi^{(\theta)}(z)^{2-a}} - 1 = -\gamma^{-1}(z^{1-b}\phi^{(\theta)}(z))^{-1},$$

so $\lim_{z\to 0} F'(z) = 0$, because $\lim_{z\to 0} z^{1-b}\phi^{(\theta)}(z) = \infty$, since $\theta \notin \Theta$. Then for every $\epsilon > 0$ there is a $\delta > 0$ such that F is ϵ-Lipschitz in $(0, \delta)$, and then $|F(z)| \leq \epsilon z$ for $z \in (0, \delta)$. So $\lim_{z\to 0} \frac{F(z)}{z} = 0$. □

Lemma Ap.2.6. $\Theta = (0, \eta)$.

Proof. Suppose indirectly that $\eta \in \Theta$. Then there is an $A > \frac{1}{\gamma}$ and a $z \in (0, \eta)$ such that $(z, -\phi_\eta(z)) \in \mathcal{R}_A$. Then for $\epsilon > 0$ small enough we have $(z, -\phi_\eta(z) - \epsilon) \in \mathcal{R}_A$ too. Let (x, p) be the solution with $(x_0, p_0) = (z, -\phi_\eta(z) - \epsilon)$. By Lemma Ap.2.2, there is a $T < 0$ such that $p(T) = 0$ and $p_t < 0$ for $t \in (T, 0]$. Then $x'_t < 0$ for $t \in (T, 0]$. Since this orbit cannot cross $\{(u, -\phi^{(\theta)}(u)); u \in (0, \eta)\}$, we must have $x(T) > \eta$. However $(x_0, p_0) \in \mathcal{R}_A$, so $(x_t, p_t) \in \mathcal{R}_A$ for every $t \geq 0$ by Lemma Ap.2.3. So $(x(T), 0)$ is the last intersection of the solution (x, p) with the x-axis, hence $x(T)$ is not in the image of η, so $\eta < x(T) \leq \eta$. This contradiction proves that $\eta \notin \Theta$.

Now suppose indirectly that there is a $\theta \in (0, \eta) \setminus \Theta$. Let us write $\phi = \phi_\eta$ and $\psi = \phi^{(\theta)}$ for simplicity. We have

$$\psi'(z) = \psi(z)^{1-a}(\gamma\psi(z) - z^{b-1}) > 0$$

for $z > 0$ close enough to 0, because $\lim_{z\to 0} \frac{\psi(z)}{z^{b-1}} = \infty$, since $\eta \notin \Theta$. So ψ has an inverse function ψ^{-1} near 0. So we can define a function $G(z) := \psi^{-1}(\phi(z))$ for $z \in (0, c)$, for some $c > 0$. We have $\psi(z) < \phi(z)$ for every $z \in (0, \theta)$, so $G(z) > z$ for $z \in (0, c)$. Then

$$\begin{aligned} G'(z) = \psi'(G(z))^{-1}\phi'(z) &= \frac{\phi(z)^{1-a}(\gamma\phi(z) - z^{b-1})}{\psi(G(z))^{1-a}(\gamma\psi(G(z)) - G(z)^{b-1})} \\ &= \frac{\gamma\phi(z) - z^{b-1}}{\gamma\phi(z) - G(z)^{b-1}}. \end{aligned}$$

Let $h(z) = G(z) - z$ for $z \in (0, c)$, then $h(z) > 0$, $\lim_{z\to 0} h(z) = 0$, and

$$h'(z) = \frac{(z + h(z))^{b-1} - z^{b-1}}{\gamma\phi(z) - G(z)^{b-1}} = z^{b-1}\frac{(1 + \frac{h(z)}{z})^{b-1} - 1}{\gamma\phi(z) - G(z)^{b-1}}.$$

If $z \to 0$, then $\phi(z)^{a-1} \sim \psi(z)^{a-1} \sim \gamma(a-1)z$ by Lemma Ap.2.5. Since $G(z) \to 0$, we also have $\gamma(a-1)z \sim \phi(z)^{a-1} = \psi(G(z))^{a-1} \sim \gamma(a-1)G(z)$. So $\lim_{z\to 0} \frac{G(z)}{z} = 1$ and $\lim_{z\to 0} \frac{h(z)}{z} = 0$. Then $\phi(z)/G(z)^{b-1} \sim (\gamma(a-1))^{\frac{1}{a-1}} z^{\frac{1}{a-1}-(b-1)}$, so $\lim_{z\to 0} \gamma\phi(z)/G(z)^{b-1} = \infty$, because $\frac{1}{a-1} < b - 1$. Therefore

$$h'(z) \sim z^{b-1}\frac{(1 + \frac{h(z)}{z})^{b-1} - 1}{\gamma\phi(z)} \sim z^{b-1}(b-1)\gamma^{-1}\frac{h(z)}{z}\frac{1}{(\gamma(a-1)z)^{\frac{1}{a-1}}} = Cz^{\lambda-1}h(z),$$

where $C = (b-1)\gamma^{-1}(\gamma(a-1))^{-\frac{1}{a-1}} > 0$ and $\lambda = b - 1 - \frac{1}{a-1} > 0$ are constants.

We know that $\phi(z) \sim (\gamma(a-1))^{\frac{1}{a-1}} z^{\frac{1}{a-1}} = 1$. Note that $\frac{1}{a-1} < b - 1$, so $\lim_{z\to 0} \gamma\phi(z)/G(z)^{b-1} = \infty$. We also have $\lim_{z\to 0} \frac{h(z)}{z} = 0$ and $(1 + \frac{h(z)}{z})^{b-1} - 1 \sim (b-1)\frac{h(z)}{z}$. So $h'(z) \sim \frac{\gamma^{-1}(b-1)}{(\gamma(a-1))^{\frac{1}{a-1}}} z^{b-2-\frac{1}{a-1}} h(z)$. Thus $h'(z) \sim Cz^{\lambda-1}h(z)$, where $C > 0$ and

$\lambda = b-2-\frac{1}{a-1}+1 > 0$, because $ba-b-a > 0$. So $\log(h(z))' = \frac{h'(z)}{h(z)} \sim Cz^{\lambda-1} = (\frac{C}{\lambda}z^{\lambda})'$. Applying L'Hôspital's rule, we get

$$1 = \lim_{z\to 0} \frac{\log(h(z))'}{(\frac{C}{\lambda}z^{\lambda})'} = \lim_{z\to 0} \frac{\log(h(z))}{\frac{C}{\lambda}z^{\lambda}} = -\infty.$$

This contradiction proves that $\Theta = (0, \eta)$. □

Now we are ready to prove our lower bound.

Proposition 4.3.10 (Lower bounds on the convergence rate in continuous time)**.** *Suppose that $\frac{1}{b} + \frac{1}{a} < 1$. For any $\theta \in \mathbb{R}$, we denote by $(x_t^{(\theta)}, p_t^{(\theta)})$ the unique solution of (4.32) with $x_0 = \theta, p_0 = 0$. Then there exists a constant $\eta \in (0, \infty)$ depending on a and b such that the path $(x_t^{(\eta)}, p_t^{(\eta)})$ and its mirrored version $(x_t^{(-\eta)}, p_t^{(-\eta)})$ satisfy that*

$$|x_t^{(-\eta)}| = |x_t^{(\eta)}| \leq \mathcal{O}(\exp(-\alpha t)) \text{ for every } \alpha < \gamma(a-1) \text{ as } t \to \infty.$$

For any path (x_t, p_t) that is not a time translation of $(x_t^{(\eta)}, p_t^{(\eta)})$ or $(x_t^{(-\eta)}, p_t^{(-\eta)})$, we have

$$\left|x_t^{-1}\right| = O(t^{\frac{1}{ba-b-a}}) \text{ as } t \to \infty,$$

so the speed of convergence is sub-linear and not linearly fast.

Proof. First let $\theta \in (0, \eta)$. Then $\theta \in \Theta$ by Lemma Ap.2.6, so there is an $A > \frac{1}{\gamma}$ and a $t_0 \in \mathbb{R}$ such that $(x_t^{(\theta)}, p_t^{(\theta)}) \in \mathcal{R}_A$ for $t \geq t_0$. Then $x_t^{(\theta)} > 0$ and $(x_t^{(\theta)})' = -|p_t^{(\theta)}|^{a-1} \geq -A^{a-1}(x_t^{(\theta)})^{(b-1)(a-1)}$ for $t \geq t_0$. Here $(b-1)(a-1) > 1$, so

$$((x_t^{(\theta)})^{-(ba-b-a)})' = -(ba-b-a)(x_t^{(\theta)})^{-(b-1)(a-1)}(x_t^{(\theta)})' \leq (ba-b-a)A^{a-1}$$

for $t \geq t_0$. Let $K := (ba-b-a)A^{a-1}$, then $(x_t^{(\theta)})^{-(ba-b-a)} \leq Kt + L$ for $t \geq t_0$ and some $L \in \mathbb{R}$. So $(x_t^{(\theta)})^{-1} = O(t^{\frac{1}{ba-b-a}})$ when $t \to \infty$, therefore the convergence is not linear.

By Lemma Ap.2.5, we have $p_t^{(\eta)} \sim -(\gamma(a-1)x_t^{(\eta)})^{\frac{1}{a-1}}$ when $t \to \infty$. So $(x^{(\eta)})_t' = -|p_t^{(\eta)}|^{a-1} \sim -\gamma(a-1)x_t^{(\eta)}$, hence $(\log(x_t^{(\eta)}))' \sim -\gamma(a-1) = (-\gamma(a-1)t)'$, when $t \to \infty$. So by L'Hôspital's rule, $\log(x_t^{(\eta)}) \sim -\gamma(a-1)t$, thus $|x_t^{(\eta)}| = O(e^{-\alpha t})$ when $t \to \infty$, for every $\alpha < \gamma(a-1)$. Then also $|p_t^{(\eta)}| = O(e^{-\beta t})$ when $t \to \infty$, for every $\beta < \gamma$. So the convergence is linear.

So up to time translation there are only two solutions, $(x^{(\eta)}, p^{(\eta)})$ and $(-x^{(\eta)}, -p^{(\eta)})$, where the convergence to $(0, 0)$ is linear. □

Ap.3 Proofs of convergence for discrete systems

Ap.3.1 Implicit Method

Firstly, we show the well-definedness of the implicit scheme.

Lemma 4.4.1 (Well-definedness of the implicit scheme)**.** *Suppose that f and k satisfy assumptions A.1 and A.2, and $\epsilon, \gamma \in (0, \infty)$. Then* (4.34) *has a unique solution for every $x_i, p_i \in \mathbb{R}^d$, and this solution also satisfies* (4.33)*.*

Proof. The proof is based on Theorem 26.3 of [222]. We start by introducing some concepts from [222] that are useful for dealing with convex functions on $\mathbb{R}^n$ taking values in $[-\infty, \infty]$. We say that $g : \mathbb{R}^n \to [-\infty, \infty]$ is *convex* if the epigraph of g, $\{(x, \mu) : \mu \geq g(x), x \in \mathbb{R}^n, \mu \in [-\infty, \infty]\}$ is convex. A convex function $g : \mathbb{R}^n \to [-\infty, \infty]$ is called *proper convex* if $g(x) \neq -\infty$ for every $x \in \mathbb{R}^n$, and there is at least one $x \in \mathbb{R}^n$ where $g(x) < \infty$. We say that $g : \mathbb{R}^n \to [-\infty, \infty]$ is *lower-semicontinuous* if $\lim_{x_i \to x} g(x_i) \geq g(x)$ for every sequence $x_i \to x$ such that $\lim_{x_i \to x} g(x_i)$ exists. The *relative interior* of a set $S \subset \mathbb{R}^n$, denoted by $\operatorname{ri} S$, is the interior of the set within its affine closure. We define the *essential domain* of a function $g : \mathbb{R}^n \to [-\infty, \infty]$, denoted by $\operatorname{dom} g$, as the set of points $x \in \mathbb{R}^n$ where $g(x)$ is finite. We call a proper convex function $g : \mathbb{R}^n \to [-\infty, \infty]$ *essentially smooth* if it satisfies the following 3 conditions for $C = \operatorname{int}(\operatorname{dom} g)$:

(a) C is non-empty

(b) g is differentiable throughout C

(c) $\lim_{i \to \infty} \|\nabla g(x_i)\| = +\infty$ whenever $x_1, x_2, \ldots$ is a sequence in C converging to a boundary point of C.

Let $\partial g(x)$ denote the subdifferential of g at x (which is the set of subgradients of g at x), and denote $\operatorname{dom} \partial g := \{x | \partial g(x) \neq \emptyset\}$. We say that a proper convex function $g : \mathbb{R}^n \to [-\infty, \infty]$ is *essentially strictly convex* if g is strictly convex for every convex subset of $\operatorname{dom} \partial g$.

By assumption A.2, k is differentiable everywhere, and it is strictly convex, hence it is both essentially smooth and essentially strictly convex (since its domain is $\operatorname{dom} k = \mathbb{R}^n$). Moreover, since k is a proper convex function, and it is lower semicontinuous everywhere (hence closed, see page 52 of [222]), it follows from Theorem 12.2 of [222] that $(k^*)^* = k$. Therefore, by Theorem 26.3 of [222], it follows that k^* is both essentially strictly convex and essentially smooth. Since f is convex and differentiable

everywhere in $\mathbb{R}^n$, based on the definitions and the assumption $\epsilon, \gamma \in (0, \infty)$, it is straightforward to show that

$$F(x) := \epsilon k^*(\tfrac{x-x_i}{\epsilon}) + \epsilon\delta f(x) - \delta \langle p_i, x\rangle$$

is also essentially strictly convex and essentially smooth. Now we are going to show that its infimum is reached at a unique point in $\mathbb{R}^n$. First, using the convexity of f, it follows that $f(x) \geq f(x_i) + \langle \nabla f(x_i), x - x_i\rangle$, hence

$$F(x) \geq \epsilon k^*(\tfrac{x-x_i}{\epsilon}) + \epsilon\delta \langle \nabla f(x_i), x - x_i\rangle - \delta \langle p_i, x - x_i\rangle - \delta \langle p_i, x_i\rangle + \epsilon\delta f(x_i)$$

using the definition (5.6) of the convex conjugate k^*

$$\begin{aligned} &\geq \langle p, x - x_i\rangle - k(p) + \epsilon\delta \langle \nabla f(x_i), x - x_i\rangle - \delta \langle p_i, x - x_i\rangle - \delta \langle p_i, x_i\rangle + \epsilon\delta f(x_i) \\ &= \langle p + \epsilon\delta\nabla f(x_i) - \delta p_i, x - x_i\rangle - k(p) - \delta \langle p_i, x_i\rangle + \epsilon\delta f(x_i), \end{aligned}$$

for any $p \in \mathbb{R}^n$. By setting $p = \frac{x-x_i}{\|x-x_i\|} - (\epsilon\delta\nabla f(x_i) - \delta p_i)$ for $\|x - x_i\| > 0$, and $p = -(\epsilon\delta\nabla f(x_i) - \delta p_i)$ for $\|x - x_i\| = 0$, using the continuity and finiteness of k, it follows that $F(x) \geq \|x - x_i\| - c$ for some $c < \infty$ depending only on ϵ, δ, x_i and p_i. Together with the lower semicontinuity of F, this implies that there exists at least one $y \in \mathbb{R}^n$ such that $F(y) = \inf_{x\in\mathbb{R}^n} F(x)$, and $-\infty < \inf_{x\in\mathbb{R}^n} F(x) < \infty$.

It remains to show that this y is unique. First, we are going to show that it falls within the interior of the domain of F. Let $C := \mathrm{int}(\mathrm{dom}\, F)$, then using the essential smoothness of F, it follows that $C \subseteq \mathrm{dom}\, F \subseteq \mathrm{cl}C$ is a non-empty convex set (cl refers to closure). If y would fall on the boundary of C, then by Lemma 26.2 of [222], $F(y)$ could not be equal to the infimum of F. Hence every such y falls within C. By Theorem 23.4 of [222], $\mathrm{ri}(\mathrm{dom}\, F) \subseteq \mathrm{dom}\, \partial F \subseteq \mathrm{dom}\, F$. Since C is a non-empty open convex set, $C = \mathrm{int}\,\mathrm{dom}\, F = \mathrm{ri}(\mathrm{dom}\, F)$, therefore from the definition of essential strict convexity, it follows that F is strictly convex on C. This means that there the infimum $\inf_{x\in\mathbb{R}^n} F(x)$ is achieved at a unique $y \in \mathbb{R}^n$, thus (4.34) is well-defined.

Finally, we show the equivalence with (4.33). First, note that using the fact that k is essentially smooth and essentially convex, it follows from Theorem 26.5 of [222] that $\nabla(k^*)(x) = (\nabla k)^{-1}(x)$ for every $x \in \mathrm{int}(\mathrm{dom}\, k^*)$. Since F is differentiable in the open set $C = \mathrm{int}(\mathrm{dom}\, F)$, and the infimum of F is taken at some $y \in C$, it follows that $\nabla F(y) = 0$. From the fact that $f(x)$ and $\langle p_i, x\rangle$ are differentiable for every $x \in \mathbb{R}^n$, it follows that for every point $z \in C$, $\frac{z-x_i}{\epsilon} \in \mathrm{int}(\mathrm{dom}\, k^*)$. Thus in particular, using the definition $x_{i+1} = y$, we have

$$\nabla(k^*)\left(\frac{x_{i+1} - x_i}{\epsilon}\right) + \epsilon\delta\nabla f(x_{i+1}) - \delta p_i = 0,$$

which can be rewritten equivalently using the second line of (4.34) as

$$\nabla(k^*)\left(\frac{x_{i+1}-x_i}{\epsilon}\right) = p_{i+1}.$$

Using the expression $\nabla(k^*)(x) = (\nabla k)^{-1}(x)$ for $x = \frac{x_{i+1}-x_i}{\epsilon} \in \operatorname{int}(\operatorname{dom} k^*)$, we obtain that $(\nabla k)^{-1}\left(\frac{x_{i+1}-x_i}{\epsilon}\right) = p_{i+1}$, and hence the first line of (4.33) follows by applying ∇k on both sides. The second line follows by rearrangement of the second line of (4.34). □

The following two lemmas are preliminary results that will be used in deriving convergence results for both this scheme and the two explicit schemes in the next sections.

Lemma Ap.3.1. *Given f, k, γ, α, $C_{\alpha,\gamma}$, and $C_{f,k}$ satisfying Assumptions A and B, and a sequence of points $x_i, p_i \in \mathbb{R}^d$ for $i \geq 0$, we define $\mathcal{H}_i := f(x_i) - f(x_{\min}) + k(p_i)$. Then the equation*

$$v = \mathcal{H}_i + \frac{C_{\alpha,\gamma}}{2}\alpha(2v)\left\langle x_i - x_{\min}, p_i\right\rangle. \tag{Ap.7}$$

has a unique solution in the interval $v \in [\mathcal{H}_i/2, 3\mathcal{H}_i/2]$, which we denote by $\mathcal{V}_i$. In addition, let

$$\beta_i := \frac{C_{\alpha,\gamma}}{2}\alpha(2\mathcal{V}_i), \tag{Ap.8}$$

then $\mathcal{V}_i = \mathcal{H}_i + \beta_i \left\langle x_i - x_{\min}, p_i\right\rangle$ and the differences $\mathcal{V}_{i+1} - \mathcal{V}_i$ can be expressed as

$$\begin{aligned}
&\mathcal{V}_{i+1} - \mathcal{V}_i \\
&= \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_{i+1}\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle - \beta_i\left\langle x_i - x_{\min}, p_i\right\rangle && \text{(Ap.9)}\\
&= \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_i(\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle - \left\langle x_i - x_{\min}, p_i\right\rangle) + (\beta_{i+1} - \beta_i)\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle && \text{(Ap.10)}\\
&= \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_{i+1}(\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle - \left\langle x_i - x_{\min}, p_i\right\rangle) + (\beta_{i+1} - \beta_i)\left\langle x_i - x_{\min}, p_i\right\rangle. && \text{(Ap.11)}
\end{aligned}$$

Proof. Similarly to (4.27), we have by Lemma 4.3.5,

$$|\left\langle x_i - x_{\min}, p_i\right\rangle| \leq k(p_i)/\alpha(k(p_i)) + f(x_i) - f(x_{\min}) \leq \frac{\mathcal{H}_i}{\alpha(k(p_i))}. \tag{Ap.12}$$

For every $i \geq 0$, we define $\mathcal{V}_i$ as the unique solution $v \in [\mathcal{H}_i/2, 3\mathcal{H}_i/2]$ of the equation

$$v = \mathcal{H}_i + \frac{C_{\alpha,\gamma}}{2}\alpha(2v)\left\langle x_i - x_{\min}, p_i\right\rangle. \tag{Ap.13}$$

The existence and uniqueness of this solution was shown in the proof of Theorem 4.3.8. The fact that $\mathcal{V}_i = \mathcal{H}_i + \beta_i \left\langle x_i - x_{\min}, p_i\right\rangle$ immediately follows from equation (Ap.13), and (Ap.10)-(Ap.11) follow by rearrangement. □

Lemma Ap.3.2. *Under the same assumptions and definitions as in Lemma Ap.3.1, if in addition we assume that for some constants $C_1, C_2 \geq 0$, for every $i \geq 0$,*

$$\begin{aligned}\mathcal{V}_{i+1} - \mathcal{V}_i \leq & -\epsilon(\gamma - \beta_{i+1} - C_1\epsilon)k(p_{i+1}) - \epsilon\gamma\beta_{i+1}\langle x_{i+1} - x_{\min}, p_{i+1}\rangle - \epsilon\beta_{i+1}(f(x_{i+1}) - f(x_{\min})) \\ & + C_2\epsilon^2\beta_{i+1}\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i)\langle x_i - x_{\min}, p_i\rangle, \text{ and} \qquad \text{(Ap.14)} \\ \mathcal{V}_{i+1} - \mathcal{V}_i \leq & -\epsilon(\gamma - \beta_i - C_1\epsilon)k(p_{i+1}) - \epsilon\gamma\beta_i\langle x_{i+1} - x_{\min}, p_{i+1}\rangle - \epsilon\beta_i(f(x_{i+1}) - f(x_{\min})) \\ & + C_2\epsilon^2\beta_i\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle, \qquad \text{(Ap.15)}\end{aligned}$$

then for every $0 < \epsilon \leq \min\left(\frac{1-\gamma}{C_2}, \frac{\gamma^2(1-\gamma)}{4C_1}\right)$, for every $i \geq 0$, we have

$$\mathcal{V}_{i+1} \leq \left[1 + \epsilon\beta_i(1 - \gamma - \epsilon C_2)/2\right]^{-1}\mathcal{V}_i. \qquad \text{(Ap.16)}$$

Similarly, if in addition to the assumptions of Lemma Ap.3.1, we assume that for some constants $C_1, C_2 \geq 0$, for every $i \geq 0$,

$$\begin{aligned}\mathcal{V}_{i+1} - \mathcal{V}_i \leq & -\epsilon(\gamma - \beta_{i+1} - C_1\epsilon)k(p_i) - \epsilon\gamma\beta_{i+1}\langle x_i - x_{\min}, p_i\rangle - \epsilon\beta_{i+1}(f(x_i) - f(x_{\min})) \\ & + C_2\epsilon^2\beta_{i+1}\mathcal{V}_i + (\beta_{i+1} - \beta_i)\langle x_i - x_{\min}, p_i\rangle, \text{ and} \qquad \text{(Ap.17)} \\ \mathcal{V}_{i+1} - \mathcal{V}_i \leq & -\epsilon(\gamma - \beta_i - C_1\epsilon)k(p_i) - \epsilon\gamma\beta_i\langle x_i - x_{\min}, p_i\rangle - \epsilon\beta_i(f(x_i) - f(x_{\min})) \\ & + C_2\epsilon^2\beta_i\mathcal{V}_i + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle, \qquad \text{(Ap.18)}\end{aligned}$$

then for every $0 < \epsilon \leq \min\left(\frac{1-\gamma}{C_2}, \frac{\gamma^2(1-\gamma)}{4C_1}\right)$, we have

$$\mathcal{V}_{i+1} \leq \left[1 - \epsilon\beta_i(1 - \gamma - \epsilon C_2)/2\right]\mathcal{V}_i. \qquad \text{(Ap.19)}$$

Proof. First suppose that assumptions (Ap.14) and (Ap.15) hold. Using (4.23) of Lemma 4.3.6 with $\alpha = \alpha(2\mathcal{V}_{i+1})$ and $\beta = \beta_{i+1}$, it follows that for $\epsilon \leq \frac{\gamma^2(1-\gamma)}{4C_1}$,

$$\begin{aligned}& -\epsilon(\gamma - \beta_{i+1} - C_1\epsilon)k(p_{i+1}) - \epsilon\beta_{i+1}(f(x_{i+1}) - f(x_{\min})) - \epsilon\beta_{i+1}\gamma\langle x_{i+1} - x_{\min}, p_{i+1}\rangle \\ & \leq -\epsilon\beta_{i+1}(1-\gamma)\mathcal{V}_{i+1},\end{aligned}$$

and by combining the terms in (Ap.14), we have

$$\mathcal{V}_{i+1} - \mathcal{V}_i \leq -\epsilon\beta_{i+1}[1 - \gamma - \epsilon C_2]\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i)\langle x_i - x_{\min}, p_i\rangle. \qquad \text{(Ap.20)}$$

Now we are going to prove that $\mathcal{V}_{i+1} \leq \mathcal{V}_i$ under the assumptions of the lemma. We argue by contradiction, suppose that $\mathcal{V}_{i+1} > \mathcal{V}_i$. Then by the non-increasing property of α, and the definition $\beta_i = \frac{C_{\alpha,\gamma}}{2}\alpha(2\mathcal{V}_i)$, we have $\beta_{i+1} \leq \beta_i$. Using the convexity of α, we have $\alpha(y) - \alpha(x) \leq \alpha'(y)(y - x)$ for any $x, y \geq 0$, hence we obtain that

$$|\beta_{i+1} - \beta_i| = \beta_i - \beta_{i+1} = \frac{C_{\alpha,\gamma}}{2}(\alpha(2\mathcal{V}_i) - \alpha(2\mathcal{V}_{i+1})) \leq C_{\alpha,\gamma}(\mathcal{V}_{i+1} - \mathcal{V}_i)(-\alpha'(2\mathcal{V}_i)),$$

and by (Ap.12) and assumption A.4 we have

$$|\beta_{i+1}-\beta_i|\,|\langle x_i - x_{\min}, p_i\rangle| \leq C_{\alpha,\gamma}(\mathcal{V}_{i+1}-\mathcal{V}_i)(-\alpha'(2\mathcal{V}_i))\frac{2\mathcal{V}_i}{\alpha(2\mathcal{V}_i)} < \mathcal{V}_{i+1}-\mathcal{V}_i.$$

Combining this with (Ap.20) we obtain that $\mathcal{V}_{i+1}-\mathcal{V}_i < \mathcal{V}_{i+1}-\mathcal{V}_i$, which is a contradiction. Hence we have shown that $\mathcal{V}_{i+1} \leq \mathcal{V}_i$, which implies that $\beta_{i+1} \geq \beta_i$.

Using (4.23) of Lemma 4.3.6 with $\alpha = \alpha(2\mathcal{V}_i)$ and $\beta = \beta_i$, it follows that for $0 < \epsilon \leq \frac{\gamma^2(1-\gamma)}{4C_1}$,

$$\begin{aligned}&-\epsilon(\gamma-\beta_i-C_1\epsilon)k(p_{i+1}) - \epsilon\beta_i(f(x_{i+1})-f(x_{\min})) - \epsilon\beta_i\gamma\,\langle x_{i+1}-x_{\min}, p_{i+1}\rangle \\ &\leq -\epsilon\beta_i(1-\gamma)\mathcal{V}_{i+1},\end{aligned}$$

and hence by substituting this to (Ap.15), it follows that

$$\mathcal{V}_{i+1}-\mathcal{V}_i \leq -\epsilon\beta_i[1-\gamma-\epsilon C_2]\mathcal{V}_{i+1} + (\beta_{i+1}-\beta_i)\,\langle x_{i+1}-x_{\min}, p_{i+1}\rangle\,. \tag{Ap.21}$$

Now using the convexity of α, and the fact that $\beta_{i+1} \geq \beta_i$, we have

$$|\beta_{i+1}-\beta_i| = \beta_{i+1}-\beta_i = \frac{C_{\alpha,\gamma}}{2}(\alpha(2\mathcal{V}_{i+1})-\alpha(2\mathcal{V}_i)) \leq C_{\alpha,\gamma}(\mathcal{V}_i-\mathcal{V}_{i+1})(-\alpha'(2\mathcal{V}_{i+1})),$$

and by (Ap.12) and assumption A.4 we have

$$|\beta_{i+1}-\beta_i|\,|\langle x_{i+1}-x_{\min}, p_{i+1}\rangle| \leq C_{\alpha,\gamma}(\mathcal{V}_i-\mathcal{V}_{i+1})(-\alpha'(2\mathcal{V}_{i+1}))\frac{2\mathcal{V}_{i+1}}{\alpha(2\mathcal{V}_{i+1})} < \mathcal{V}_i-\mathcal{V}_{i+1}.$$

By combining this with (Ap.21), we obtain that

$$\mathcal{V}_{i+1}-\mathcal{V}_i \leq -\frac{\epsilon\beta_i}{2}[1-\gamma-\epsilon C_2]\mathcal{V}_{i+1},$$

and the first claim of the lemma follows by rearrangement and monotonicity.

The proof of the second claim based on assumptions (Ap.17) and (Ap.18) is as follows. As previously, in the first step, we show that $\mathcal{V}_{i+1} \leq \mathcal{V}_i$ by contradiction. Suppose that $\mathcal{V}_{i+1} > \mathcal{V}_i$, then $\beta_{i+1} \leq \beta_i$. Using (4.23) of Lemma 4.3.6 with $\alpha = \alpha(2\mathcal{V}_i)$ and $\beta = \beta_{i+1} \leq \beta_i \leq \frac{\alpha\gamma}{2}$, it follows that for $\epsilon \leq \frac{\gamma^2(1-\gamma)}{4C_1}$,

$$\begin{aligned}&-\epsilon(\gamma-\beta_{i+1}-C_1\epsilon)k(p_{i+1}) - \epsilon\beta_{i+1}(f(x_{i+1})-f(x_{\min})) - \epsilon\beta_{i+1}\gamma\,\langle x_{i+1}-x_{\min}, p_{i+1}\rangle \\ &\leq -\epsilon\beta_{i+1}(1-\gamma)\mathcal{V}_{i+1},\end{aligned}$$

and by combining the terms in (Ap.17), we have

$$\mathcal{V}_{i+1}-\mathcal{V}_i \leq -\epsilon\beta_{i+1}[1-\gamma-\epsilon C_2]\mathcal{V}_i + (\beta_{i+1}-\beta_i)\,\langle x_i-x_{\min}, p_i\rangle\,. \tag{Ap.22}$$

The rest of the proof follows the same steps as for assumptions (Ap.14) and (Ap.15), hence it is omitted. □

Now we are ready to prove the main result of this section.

Proposition 4.4.3 (Convergence bound for the implicit scheme)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$, and $C_{f,k}$ satisfying assumptions A and B. Suppose that $\epsilon < \frac{1-\gamma}{2\max(C_{f,k},1)}$. Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, and let $\mathcal{W}_0 = f(x_0) - f(x_{\min})$ and for $i \geq 0$,*

$$\mathcal{W}_{i+1} = \mathcal{W}_i \left[1 + \epsilon C_{\alpha,\gamma}(1-\gamma-2C_{f,k}\epsilon)\alpha(2\mathcal{W}_i)/4\right]^{-1}.$$

Then for any (x_0, p_0) with $p_0 = 0$, the iterates of (4.33) satisfy for every $i \geq 0$,

$$f(x_i) - f(x_{\min}) \leq 2\mathcal{W}_i \leq 2\mathcal{W}_0[1 + \epsilon C_{\alpha,\gamma}(1-\gamma-2C_{f,k}\epsilon)\alpha_\star/4]^{-i}.$$

Proof. We follow the notations of Lemma Ap.3.1, and the proof is based on Lemma Ap.3.2. By rearrangement of the (4.33), we have

$$\begin{aligned} x_{i+1} - x_i &= \epsilon \nabla k(p_{i+1}) \\ p_{i+1} - p_i &= -\gamma\epsilon p_{i+1} - \epsilon \nabla f(x_{i+1}) \end{aligned} \tag{Ap.23}$$

For the Hamiltonian terms, by the convexity of f and k, we have

$$\begin{aligned} \mathcal{H}_{i+1} - \mathcal{H}_i &\leq \langle \nabla k(p_{i+1}), p_{i+1} - p_i \rangle + \langle \nabla f(x_{i+1}), x_{i+1} - x_i \rangle \\ &= \langle \nabla k(p_{i+1}), -\gamma\epsilon p_{i+1} - \epsilon \nabla f(x_{i+1}) \rangle + \epsilon \langle \nabla f(x_{i+1}), \nabla k(p_{i+1}) \rangle \qquad \text{(Ap.24)} \\ &= -\gamma\epsilon \langle \nabla k(p_{i+1}), p_{i+1} \rangle \qquad \text{(Ap.25)} \end{aligned}$$

For the inner product terms, we have

$$\begin{aligned} &\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle \\ &= \langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_{i+1} - x_{\min} - (x_{i+1} - x_i), p_{i+1} - (p_{i+1} - p_i) \rangle \\ &= \langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_{i+1} - x_{\min} - \epsilon \nabla k(p_{i+1}), p_{i+1} + \epsilon\gamma p_{i+1} + \epsilon \nabla f(x_{i+1}) \rangle \\ &= (\epsilon + \gamma\epsilon^2) \langle p_{i+1}, \nabla k(p_{i+1}) \rangle - \epsilon \langle x_{i+1} - x_{\min}, \nabla f(x_{i+1}) \rangle \\ &\quad - \epsilon\gamma \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + \epsilon^2 \langle \nabla k(p_{i+1}), \nabla f(x_{i+1}) \rangle, \end{aligned}$$

and by assumption B.1 we have

$$\langle \nabla k(p_{i+1}), \nabla f(x_{i+1}) \rangle \leq C_{f,k} \mathcal{H}_{i+1} \leq 2C_{f,k} \mathcal{V}_{i+1}, \tag{Ap.26}$$

and hence

$$\begin{aligned} \langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle \leq (\epsilon + \gamma\epsilon^2) \langle p_{i+1}, \nabla k(p_{i+1}) \rangle - \epsilon \langle x_{i+1} - x_{\min}, \nabla f(x_{i+1}) \rangle \\ - \epsilon\gamma \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + 2\epsilon^2 C_{f,k} \mathcal{V}_{i+1}. \end{aligned} \tag{Ap.27}$$

By assumption A.4 on $C_{\alpha,\gamma}$ we have $\beta_{i+1} \leq \frac{\gamma}{2}$, and using the condition $\epsilon < \frac{1-\gamma}{2(C_{f,k}+\gamma)}$ of the lemma, we have

$$\gamma - \beta_{i+1} - \epsilon\gamma\beta_{i+1} \geq \gamma - \frac{\gamma}{2} - \frac{(1-\gamma)}{2\gamma}\gamma\frac{\gamma}{2} > 0. \tag{Ap.28}$$

By (Ap.11), (Ap.25), (Ap.27), we have

$$\begin{aligned}
&\mathcal{V}_{i+1} - \mathcal{V}_i \\
&\leq -\epsilon(\gamma - \beta_{i+1} - \epsilon\gamma\beta_{i+1}) \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon\beta_{i+1} \langle x_{i+1} - x_{\min}, \nabla f(x_{i+1}) \rangle \\
&- \epsilon\gamma\beta_{i+1} \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + 2\epsilon^2 C_{f,k}\beta_{i+1}\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle
\end{aligned}$$

using the convexity of f and k, and inequality (Ap.28)

$$\begin{aligned}
&\leq -\epsilon(\gamma - \beta_{i+1} - \epsilon\gamma\beta_{i+1})k(p_{i+1}) - \epsilon\beta_{i+1}(f(x_{i+1}) - f(x_{\min})) - \epsilon\gamma\beta_{i+1} \langle x_{i+1} - x_{\min}, p_{i+1} \rangle \\
&+ 2\epsilon^2 C_{f,k}\beta_{i+1}\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle .
\end{aligned}$$

Using the fact that $\beta_{i+1} \leq \frac{C_{\alpha,\gamma}}{2} \leq \frac{\gamma}{2}$, it follows that (Ap.14) holds with $C_1 = \frac{\gamma^2}{2}$ and $C_2 = 2C_{f,k}$.

By (Ap.10), (Ap.25), (Ap.27), it follows that

$$\begin{aligned}
&\mathcal{V}_{i+1} - \mathcal{V}_i \leq \\
&\mathcal{H}_{i+1} - \mathcal{H}_i + \beta_i(\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle) + (\beta_{i+1} - \beta_i) \langle x_{i+1} - x_{\min}, p_{i+1} \rangle \\
&\leq -\gamma\epsilon \langle \nabla k(p_{i+1}), p_{i+1} \rangle + \beta_i(\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle) + (\beta_{i+1} - \beta_i) \langle x_{i+1} - x_{\min}, p_{i+1} \rangle \\
&\leq -\gamma\epsilon \langle \nabla k(p_{i+1}), p_{i+1} \rangle + (\beta_{i+1} - \beta_i) \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + (\beta_i\epsilon + \beta_i\gamma\epsilon^2) \langle p_{i+1}, \nabla k(p_{i+1}) \rangle \\
&- \beta_i\epsilon \langle x_{i+1} - x_{\min}, \nabla f(x_{i+1}) \rangle - \beta_i\epsilon\gamma \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + 2\beta_i\epsilon^2 C_{f,k}\mathcal{V}_{i+1}
\end{aligned}$$

using the convexity of f and k, and inequality (Ap.28)

$$\begin{aligned}
&\leq -\epsilon(\gamma - \beta_i - \epsilon\gamma\beta_i)k(p_{i+1}) - \epsilon\beta_i\gamma \langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \epsilon\beta_i(f(x_{i+1}) - f(x_{\min})) \\
&+ 2\epsilon^2 C_{f,k}\beta_i\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i) \langle x_{i+1} - x_{\min}, p_{i+1} \rangle ,
\end{aligned}$$

implying that (Ap.15) holds with $C_1 = \frac{\gamma^2}{2}$ and $C_2 = 2C_{f,k}$. The claim of the Lemma now follows from Lemma Ap.3.2. □

Ap.3.2 First Explicit Method

The following lemma is a preliminary result that will be useful for proving our convergence bounds for this discretization.

Lemma Ap.3.3. *Given f, k, γ, α, $C_{\alpha,\gamma}$, $C_{f,k}$, C_k, $D_{f,k}$ satisfying assumptions A, B, and C, and $0 < \epsilon \leq \frac{C_{\alpha,\gamma}}{10C_{f,k}+5\gamma C_k}$, the iterates (4.36) satisfy that for every $i \geq 0$,*

$$\langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle \leq 3\epsilon^2 D_{f,k} \min(\alpha(3\mathcal{H}_i), \alpha(3\mathcal{H}_{i+1}))\mathcal{H}_{i+1}. \quad \text{(Ap.29)}$$

Proof. Let $x_{i+1}^{(t)} := x_{i+1} - t\epsilon\nabla k(p_{i+1})$ and $\mathcal{H}_{i+1}^{(t)} := \mathcal{H}(x_{i+1}^{(t)}, p_{i+1})$. Using the assumptions that f is 2 times continuously differentiable, and assumption C.3, we have

$$\langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle = \int_{t=0}^{1} \left\langle x_{i+1} - x_i, \nabla^2 f(x_{i+1} - t(x_{i+1} - x_i))(x_{i+1} - x_i) \right\rangle dt$$

$$= \epsilon^2 \int_{t=0}^{1} \left\langle \nabla k(p_{i+1}), \nabla^2 f(x_{i+1}^{(t)}) \nabla k(p_{i+1}) \right\rangle dt \leq \epsilon^2 D_{f,k} \int_{t=0}^{1} \alpha(3\mathcal{H}_{i+1}^{(t)})\mathcal{H}_{i+1}^{(t)} dt, \quad \text{(Ap.30)}$$

where we have used the fundamental theorem of calculus, which is applicable since $\left\langle \nabla k(p_{i+1}), \nabla^2 f(x_{i+1}^{(t)}) \nabla k(p_{i+1}) \right\rangle$ is piecewise continuous by assumption C.2. We are going to show the following inequalities based on the assumptions of the Lemma,

$$\mathcal{H}_{i+1}^{(t)} \leq \frac{1}{1 - \epsilon C_{f,k}} \mathcal{H}_{i+1}, \quad \text{(Ap.31)}$$

$$\alpha(3\mathcal{H}_{i+1}^{(t)}) \leq \alpha(3\mathcal{H}_{i+1}) \cdot \frac{1 - \epsilon C_{f,k}}{1 - \epsilon C_{f,k}(1 + 1/C_{\alpha,\gamma})}, \quad \text{(Ap.32)}$$

$$\alpha(3\mathcal{H}_{i+1}^{(t)}) \leq \alpha(3\mathcal{H}_i) \cdot \frac{1 - \epsilon(2C_{f,k} + \gamma C_k)}{1 - \epsilon[C_{f,k}(2 + 3/C_{\alpha,\gamma}) + \gamma C_k(1 + 1/C_{\alpha,\gamma})]}. \quad \text{(Ap.33)}$$

The claim of the lemma follows directly by combining these 3 inequalities with (Ap.30) and using the assumptions on ϵ.

First, by convexity and assumption B.1, we have

$$\mathcal{H}_{i+1}^{(t)} - \mathcal{H}_{i+1} = f(x_{i+1}^{(t)}) - f(x_{i+1}) \leq -\left\langle \nabla f(x_{i+1}^{(t)}), t\epsilon\nabla k(p_{i+1}) \right\rangle \leq t\epsilon C_{f,k}\mathcal{H}_{i+1}^{(t)},$$

and (Ap.31) follows by rearrangement. In the other direction, by convexity and assumption B.1, we have

$$\mathcal{H}_{i+1} - \mathcal{H}_{i+1}^{(t)} = f(x_{i+1}) - f(x_{i+1}^{(t)}) \leq \langle \nabla f(x_{i+1}), t\epsilon\nabla k(p_{i+1}) \rangle \leq t\epsilon C_{f,k}\mathcal{H}_{i+1},$$

so by rearrangement, it follows that

$$\mathcal{H}_{i+1} - \mathcal{H}_{i+1}^{(t)} \leq \frac{t\epsilon C_{f,k}}{1 - t\epsilon C_{f,k}} \mathcal{H}_{i+1}^{(t)}.$$

Using this, and the convexity of α, and Assumption A.4, we have

$$\alpha(3\mathcal{H}_{i+1}^{(t)}) - \alpha(3\mathcal{H}_{i+1}) \leq -3\alpha'(3\mathcal{H}_{i+1}^{(t)})(\mathcal{H}_{i+1} - \mathcal{H}_{i+1}^{(t)}) \leq -\alpha'(3\mathcal{H}_{i+1}^{(t)})3\mathcal{H}_{i+1}^{(t)} \frac{t\epsilon C_{f,k}}{1 - t\epsilon C_{f,k}}$$

$$\leq \frac{1}{C_{\alpha,\gamma}} \frac{t\epsilon C_{f,k}}{1 - t\epsilon C_{f,k}} \alpha(3\mathcal{H}_{i+1}^{(t)}),$$

and (Ap.32) follows by rearrangement. Finally, using the convexity of f and k, we have

$$\begin{aligned}\mathcal{H}_i - \mathcal{H}_{i+1}^{(t)} &= k(p_i) - k(p_{i+1}) + f(x_i) - f(x_{i+1}^{(t)}) \\ &\leq \left\langle \nabla k(p_i), \frac{\gamma\epsilon}{1+\gamma\epsilon} p_i + \frac{\epsilon}{1+\gamma\epsilon}\nabla f(x_i) \right\rangle + \langle \nabla f(x_i), -\epsilon(1-t)\nabla k(p_{i+1})\rangle\end{aligned}$$

using Assumptions B.1 and C.1

$$\begin{aligned}&\leq \gamma\epsilon C_k k(p_i) + \epsilon C_{f,k}\mathcal{H}_i + \epsilon C_{f,k}(k(p_{i+1}) + f(x_i) - f(x_{\min})) \\ &\leq \epsilon[(2C_{f,k} + \gamma C_k)\mathcal{H}_i + C_{f,k}\mathcal{H}_{i+1}^{(t)}].\end{aligned}$$

By rearrangement, this implies that

$$\mathcal{H}_i - \mathcal{H}_{i+1}^{(t)} \leq \frac{\epsilon(3C_{f,k} + \gamma C_k)}{1 - (2C_{f,k} + \gamma C_k)\epsilon} \cdot \mathcal{H}_{i+1}^{(t)}.$$

Using this, the convexity of α, and Assumption A.4, we have

$$\begin{aligned}&\alpha(3\mathcal{H}_{i+1}^{(t)}) - \alpha(3\mathcal{H}_i) \leq -3\alpha'(3\mathcal{H}_{i+1}^{(t)})(\mathcal{H}_i - \mathcal{H}_{i+1}^{(t)}) \leq -\alpha'(3\mathcal{H}_{i+1}^{(t)})3\mathcal{H}_{i+1}^{(t)} \cdot \frac{\epsilon(3C_{f,k} + \gamma C_k)}{1 - (2C_{f,k} + \gamma C_k)\epsilon} \\ &\leq \frac{1}{C_{\alpha,\gamma}} \cdot \frac{\epsilon(3C_{f,k} + \gamma C_k)}{1 - (2C_{f,k} + \gamma C_k)\epsilon} \cdot \alpha(3\mathcal{H}_{i+1}^{(t)}),\end{aligned}$$

and (Ap.33) follows by rearrangement. □

Now we are ready to prove our convergence bound for this discretization.

Proposition 4.4.6 (Convergence bound for the first explicit scheme)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$, $C_{f,k}$, C_k, $D_{f,k}$ satisfying assumptions A, B, and C, and that $0 < \epsilon < \min\left(\frac{1-\gamma}{2\max(C_{f,k}+6D_{f,k}/C_{\alpha,\gamma},1)}, \frac{C_{\alpha,\gamma}}{10C_{f,k}+5\gamma C_k}\right)$. Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, $\mathcal{W}_0 := f(x_0) - f(x_{\min})$, and for $i \geq 0$, let*

$$\mathcal{W}_{i+1} = \mathcal{W}_i \left(1 + \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1 - \gamma - 2\epsilon(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\right]\alpha(2\mathcal{W}_i)\right)^{-1}.$$

Then for any (x_0, p_0) with $p_0 = 0$, the iterates (4.36) satisfy for every $i \geq 0$,

$$f(x_i) - f(x_{\min}) \leq 2\mathcal{W}_i \leq 2\mathcal{W}_0 \left(1 + \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1 - \gamma - 2\epsilon(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\right]\alpha_\star\right)^{-i}.$$

Proof. We follow the notations of Lemma Ap.3.1, and the proof is based on Lemma Ap.3.2. For the Hamiltonian terms, by the convexity of f and k, we have

$$\begin{aligned}
&\mathcal{H}_{i+1} - \mathcal{H}_i \\
&= f(x_{i+1}) - f(x_i) + k(p_{i+1}) - k(p_i) \qquad \text{(Ap.34)} \\
&\le \langle \nabla f(x_{i+1}), x_{i+1} - x_i \rangle + \langle \nabla k(p_{i+1}), p_{i+1} - p_i \rangle \\
&= \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \langle \nabla k(p_{i+1}), p_{i+1} - p_i \rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle \\
&= \epsilon \langle \nabla f(x_i), \nabla k(p_{i+1}) \rangle - \epsilon \langle \nabla k(p_{i+1}), \nabla f(x_i) + \gamma p_{i+1} \rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle \\
&= -\gamma\epsilon \langle \nabla k(p_{i+1}), p_{i+1} \rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle \qquad \text{(Ap.35)}
\end{aligned}$$

for any $\epsilon > 0$. Note that by convexity and assumption B.1, we have

$$\begin{aligned}
-f(x_i) &= -f(x_{i+1}) + f(x_{i+1}) - f(x_i) \le -f(x_{i+1}) + \epsilon \langle \nabla f(x_{i+1}), \nabla k(p_{i+1}) \rangle \\
&\le -f(x_{i+1}) + \epsilon C_{f,k} \mathcal{H}_{i+1} \le -f(x_{i+1}) + 2\epsilon C_{f,k} \mathcal{V}_{i+1}.
\end{aligned}$$

For the inner product terms, using the above inequality and convexity, we have

$$\begin{aligned}
&\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle \\
&= \langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_{i+1} \rangle + \langle x_i - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle \\
&= \epsilon \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon \langle x_i - x_{\min}, \nabla f(x_i) \rangle - \gamma\epsilon \langle x_i - x_{\min}, p_{i+1} \rangle \\
&= (\epsilon + \gamma\epsilon^2) \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon \langle x_i - x_{\min}, \nabla f(x_i) \rangle - \gamma\epsilon \langle x_{i+1} - x_{\min}, p_{i+1} \rangle \\
&\le (\epsilon + \gamma\epsilon^2) \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon(f(x_i) - f(x_{\min})) - \gamma\epsilon \langle x_{i+1} - x_{\min}, p_{i+1} \rangle \\
&\le (\epsilon + \gamma\epsilon^2) \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon(f(x_{i+1}) - f(x_{\min})) - \gamma\epsilon \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + 2\epsilon^2 C_{f,k} \mathcal{V}_{i+1}. \\
&\qquad \text{(Ap.36)}
\end{aligned}$$

Since $C_{\alpha,\gamma} \le \gamma$, it follows that $\beta_{i+1} = \frac{C_{\alpha,\gamma}}{2} \alpha(2\mathcal{V}_{i+1}) \le \frac{\gamma}{2}$, and using the assumption on ϵ, we have

$$\gamma - \beta_{i+1} - \epsilon\gamma\beta_{i+1} \ge \gamma - \frac{\gamma}{2} - \frac{1-\gamma}{2\gamma} \gamma \frac{\gamma}{2} > 0. \qquad \text{(Ap.37)}$$

By (Ap.11), (Ap.35), and (Ap.36), it follows that

$$\begin{aligned}
&\mathcal{V}_{i+1} - \mathcal{V}_i \\
&= \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_{i+1}(\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle) + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle \\
&\le -\gamma\epsilon \langle \nabla k(p_{i+1}), p_{i+1} \rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle \\
&+ \beta_{i+1} \left((\epsilon + \gamma\epsilon^2) \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon(f(x_{i+1}) - f(x_{\min})) - \gamma\epsilon \langle x_{i+1} - x_{\min}, p_{i+1} \rangle + 2\epsilon^2 C_{f,k} \mathcal{V}_{i+1} \right) \\
&\le -\epsilon(\gamma - \beta_{i+1} - \epsilon\gamma\beta_{i+1}) \langle \nabla k(p_{i+1}), p_{i+1} \rangle - \epsilon\beta_{i+1} f(x_{i+1}) - \epsilon\gamma\beta_{i+1} \langle x_{i+1} - x_{\min}, p_{i+1} \rangle \\
&+ 2\epsilon^2 \beta_{i+1} C_{f,k} \mathcal{V}_{i+1} + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i \rangle + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle
\end{aligned}$$

which can be further bounded using (Ap.37), the convexity of k, and Lemma Ap.3.3 as

$$\leq -\epsilon(\gamma - \beta_{i+1} - \epsilon\frac{\gamma^2}{2})k(p_{i+1}) - \epsilon\beta_{i+1}\gamma\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle - \epsilon\beta_{i+1}(f(x_{i+1}) - f(x_{\min}))$$
$$+ 2\epsilon^2(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\beta_{i+1}\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i)\left\langle x_i - x_{\min}, p_i\right\rangle,$$

implying that (Ap.14) holds with $C_1 = \frac{\gamma^2}{2}$ and $C_2 = 2(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})$.

Since $\mathcal{H}_i \leq 2\mathcal{V}_i \leq 3\mathcal{H}_i$, and by applying Lemma Ap.3.3 it follows that

$$\langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i\rangle \leq 6\epsilon^2 D_{f,k}\frac{\beta_i}{C_{\alpha,\gamma}}\mathcal{H}_{i+1} \leq 12\epsilon^2\frac{D_{f,k}}{C_{\alpha,\gamma}}\beta_i\mathcal{V}_{i+1}. \quad \text{(Ap.38)}$$

By (Ap.10), (Ap.35), (Ap.36), and assumption B.1, we have

$$\mathcal{V}_{i+1} - \mathcal{V}_i$$
$$= \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_i(\langle x_{i+1} - x_{\min}, p_{i+1}\rangle - \langle x_i - x_{\min}, p_i\rangle) + (\beta_{i+1} - \beta_i)\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle$$
$$\leq -\gamma\epsilon\left\langle \nabla k(p_{i+1}), p_{i+1}\right\rangle + \langle \nabla f(x_{i+1}) - \nabla f(x_i), x_{i+1} - x_i\rangle + (\beta_{i+1} - \beta_i)\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle$$
$$+ (\beta_i\epsilon + \gamma\beta_i\epsilon^2)\left\langle \nabla k(p_{i+1}), p_{i+1}\right\rangle - \beta_i\epsilon(f(x_{i+1}) - f(x_{\min})) - \gamma\beta_i\epsilon\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle$$
$$+ 2\epsilon^2\beta_i C_{f,k}\mathcal{V}_{i+1}$$

using (Ap.38) and the convexity of f and k

$$\leq -\epsilon\left(\gamma - \beta_i - \epsilon\frac{\gamma^2}{2}\right)k(p_{i+1}) - \epsilon\beta_i\gamma\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle - \epsilon\beta_i(f(x_{i+1}) - f(x_{\min}))$$
$$+ 2\epsilon^2(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})\beta_i\mathcal{V}_{i+1} + (\beta_{i+1} - \beta_i)\left\langle x_{i+1} - x_{\min}, p_{i+1}\right\rangle,$$

implying that (Ap.15) holds with $C_1 = \frac{\gamma^2}{2}$ and $C_2 = 2(C_{f,k} + 6D_{f,k}/C_{\alpha,\gamma})$. The claim of the lemma now follows by Lemma Ap.3.2. □

Ap.3.3 Second Explicit Method

The following preliminary result will be used in the proof of the convergence bound.

Lemma Ap.3.4. *Given f, k, γ, α, $C_{\alpha,\gamma}$, $C_{f,k}$, C_k, $D_{f,k}$ satisfying assumptions A, B, and D, and $0 < \epsilon \leq \min\left(\frac{C_{\alpha,\gamma}}{6(5C_{f,k}+2\gamma C_k)+12\gamma C_{\alpha,\gamma}}, \sqrt{\frac{1}{6\gamma^2 D_k F_k}}\right)$, the iterates* (4.39) *satisfy that for every $i \geq 0$,*

$$\langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i\rangle \leq \epsilon^2 C\min(\alpha(3\mathcal{H}_i), \alpha(3\mathcal{H}_{i+1}))\mathcal{H}_i + \epsilon^2 Dk(p_i), \quad \text{(Ap.39)}$$

where

$$C = 3D_{f,k}, \qquad D = 2\gamma^2 D_k(1 + E_k). \quad \text{(Ap.40)}$$

Proof. For $0 \leq t \leq 1$, let

$$p_i^{(t)} := p_i + t(p_{i+1} - p_i) = (1 - \epsilon\gamma t)p_i - \epsilon t \nabla f(x_{i+1}), \tag{Ap.41}$$

$$\mathcal{H}_i^{(t)} := \mathcal{H}\left(x_{i+1}, p_i^{(t)}\right) = f(x_{i+1}) - f(x_{\min}) + k\left(p_i^{(t)}\right), \tag{Ap.42}$$

$$\mathcal{P}_{i,i+1} := \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i \rangle . \tag{Ap.43}$$

Note that by rearrangement we have $p_i = \left(p_i^{(t)} + \epsilon t \nabla f(x_{i+1})\right)/(1 - \epsilon\gamma t)$, and hence

$$p_{i+1} - p_i = \frac{p_i^{(t)} - p_i}{t} = \frac{-\epsilon\gamma p_i^{(t)} - \epsilon \nabla f(x_{i+1})}{1 - \epsilon\gamma t}. \tag{Ap.44}$$

Using assumption D.1, it follows that $\left\langle p_{i+1} - p_i, \nabla^2 k\left(p_i^{(t)}\right)(p_{i+1} - p_i)\right\rangle$ is piecewise continuous, hence by the fundamental theorem of calculus, we have

$$\begin{aligned}
\mathcal{P}_{i,i+1} &= \int_{t=0}^{1} \left\langle p_{i+1} - p_i, \nabla^2 k\left(p_i^{(t)}\right)(p_{i+1} - p_i)\right\rangle dt \\
&= \frac{1}{(1-\epsilon\gamma t)^2} \int_{t=0}^{1} \left\langle \epsilon\gamma p_i^{(t)} + \epsilon \nabla f(x_{i+1}), \nabla^2 k\left(p_i^{(t)}\right)(\epsilon\gamma p_i^{(t)} + \epsilon \nabla f(x_{i+1})\right\rangle dt \\
&\leq \frac{2\epsilon^2\gamma^2}{(1-\epsilon\gamma)^2} \int_{t=0}^{1} \left\langle p_i^{(t)}, \nabla^2 k\left(p_i^{(t)}\right) p_i^{(t)}\right\rangle + \frac{2\epsilon^2}{(1-\epsilon\gamma)^2} \int_{t=0}^{1} \left\langle \nabla f(x_{i+1}), \nabla^2 k\left(p_i^{(t)}\right) \nabla f(x_{i+1})\right\rangle dt
\end{aligned} \tag{Ap.45}$$

For the first integral, using Assumptions D.3, the convexity of k, and then D.4, we have

$$\begin{aligned}
\int_{t=0}^{1} \left\langle p_i^{(t)}, \nabla^2 k\left(p_i^{(t)}\right) p_i^{(t)}\right\rangle dt &\leq D_k \int_{t=0}^{1} k\left(p_i^{(t)}\right) dt \leq \frac{D_k}{2}\left(k(p_i) + k(p_{i+1})\right) \\
&\leq \frac{D_k}{2}((1 + E_k)k(p_i) + F_k \mathcal{P}_{i,i+1})
\end{aligned} \tag{Ap.46}$$

For the second integral, using Assumption D.5, we have

$$\int_{t=0}^{1} \left\langle \nabla f(x_{i+1}), \nabla^2 k\left(p_i^{(t)}\right) \nabla f(x_{i+1})\right\rangle dt \leq D_{f,k} \int_{t=0}^{1} \mathcal{H}_i^{(t)} \alpha(3\mathcal{H}_i^{(t)}) dt. \tag{Ap.47}$$

We are going to show the following 3 inequalities based on the assumptions of the Lemma.

$$\mathcal{H}_i^{(t)} \leq \frac{1 - \epsilon\gamma}{1 - \epsilon(\gamma + 2C_{f,k})} \cdot \mathcal{H}_i, \tag{Ap.48}$$

$$\alpha(3\mathcal{H}_i^{(t)}) \leq \alpha(3\mathcal{H}_{i+1}) \cdot \frac{1 - (C_{f,k} + \gamma)\epsilon}{1 - (C_{f,k} + \gamma + C_{f,k}/C_{\alpha,\gamma})\epsilon}, \tag{Ap.49}$$

$$\alpha(3\mathcal{H}_i^{(t)}) \leq \alpha(3\mathcal{H}_i) \cdot \frac{1 - \epsilon(2C_{f,k} + \gamma C_k)}{1 - \epsilon[C_{f,k}(2 + 3/C_{\alpha,\gamma}) + \gamma C_k(1 + 1/C_{\alpha,\gamma})]}. \tag{Ap.50}$$

The claim of the lemma follows from substituting these bounds into (Ap.47), and then substituting the bounds (Ap.46) and (Ap.47) into (Ap.45) and rearranging.

First, by the convexity of f and Assumption B.1, we have

$$\begin{aligned} f(x_{i+1}) - f(x_i) &\leq \epsilon \left\langle \nabla f(x_{i+1}), \nabla k(p_i) \right\rangle \leq \epsilon C_{f,k}(f(x_{i+1}) - f(x_{\min}) + k(p_i)) \\ &= \epsilon C_{f,k}((f(x_{i+1}) - f(x_i)) + \mathcal{H}_i) \end{aligned}$$

so by rearrangement it follows that

$$f(x_{i+1}) - f(x_i) \leq \frac{\epsilon C_{f,k}}{1 - \epsilon C_{f,k}} \cdot \mathcal{H}_i, \tag{Ap.51}$$

and similiarly

$$f(x_i) - f(x_{i+1}) \leq -\epsilon \left\langle \nabla f(x_i), \nabla k(p_i) \right\rangle \leq \epsilon C_{f,k} \mathcal{H}_i. \tag{Ap.52}$$

Using (Ap.51), and the convexity of k, we have

$$\begin{aligned} \mathcal{H}_i^{(t)} - \mathcal{H}_i &= f(x_{i+1}) - f(x_i) + k\left(p_i^{(t)}\right) - k(p_i) \\ &\leq \frac{\epsilon C_{f,k}}{1 - \epsilon C_{f,k}} \cdot \mathcal{H}_i + \left\langle \nabla k\left(p_i^{(t)}\right), t(p_{i+1} - p_i) \right\rangle \end{aligned}$$

now using (Ap.44), and then Assumption B.1,

$$\begin{aligned} &\leq \frac{\epsilon C_{f,k}}{1 - \epsilon C_{f,k}} \cdot \mathcal{H}_i - \epsilon t \left\langle \nabla k\left(p_i^{(t)}\right), \frac{\gamma p_i^{(t)} + \nabla f(x_{i+1})}{1 - \epsilon\gamma t} \right\rangle \\ &\leq \frac{\epsilon C_{f,k}}{1 - \epsilon C_{f,k}} \cdot \mathcal{H}_i + \frac{\epsilon C_{f,k}}{1 - \epsilon\gamma} \mathcal{H}_i^{(t)}, \end{aligned}$$

and inequality (Ap.48) follows by rearrangement.

By the convexity of k, and using (Ap.44) for $t = 1$, we have

$$\begin{aligned} &\mathcal{H}_{i+1} - \mathcal{H}_i^{(t)} = k(p_{i+1}) - k\left(p_i^{(t)}\right) \leq \left\langle \nabla k(p_{i+1}), p_{i+1} - p_i^{(t)} \right\rangle \\ &= \left\langle \nabla k(p_{i+1}), (1-t)(p_{i+1} - p_i) \right\rangle = -(1-t) \left\langle \nabla k(p_{i+1}), \frac{\epsilon\gamma}{1 - \epsilon\gamma} p_{i+1} + \frac{\epsilon}{1 - \epsilon\gamma} \nabla f(x_{i+1}) \right\rangle \end{aligned}$$

using Assumption B.1,

$$\leq \frac{\epsilon C_{f,k}}{1 - \gamma\epsilon} \cdot \mathcal{H}_{i+1},$$

so by rearrangement,

$$\mathcal{H}_{i+1} - \mathcal{H}_i^{(t)} \leq \frac{\epsilon C_{f,k}}{1 - (C_{f,k} + \gamma)\epsilon} \mathcal{H}_i^{(t)}. \tag{Ap.53}$$

Using this, the convexity of α, and Assumption A.4, we have

$$\alpha(3\mathcal{H}_i^{(t)}) - \alpha(3\mathcal{H}_{i+1}) \leq -3\alpha'(3\mathcal{H}_i^{(t)})(\mathcal{H}_{i+1} - \mathcal{H}_i^{(t)}) \leq -\alpha'(3\mathcal{H}_i^{(t)})3\mathcal{H}_i^{(t)} \cdot \frac{\epsilon C_{f,k}}{1-(C_{f,k}+\gamma)\epsilon}$$
$$\leq \frac{1}{C_{\alpha,\gamma}} \cdot \frac{\epsilon C_{f,k}}{1-(C_{f,k}+\gamma)\epsilon} \cdot \alpha(3\mathcal{H}_i^{(t)}),$$

and (Ap.49) follows by rearrangement. Finally, using inequality (Ap.52), we have

$$\begin{aligned}\mathcal{H}_i - \mathcal{H}_i^{(t)} &= f(x_i) - f(x_{i+1}) + k(p_i) - k\left(p_i^{(t)}\right) \\ &\leq \epsilon C_{f,k}\mathcal{H}_i + \langle \nabla k(p_i), -t(p_{i+1} - p_i)\rangle \\ &\leq \epsilon C_{f,k}\mathcal{H}_i + \epsilon t \langle \nabla k(p_i), \gamma p_i + \nabla f(x_{i+1})\rangle\end{aligned}$$

now using Assumptions B.1 and D.2,

$$\begin{aligned}&\leq \epsilon(C_{f,k}\mathcal{H}_i + \gamma C_k k(p_i) + C_{f,k}k(p_i) + C_{f,k}(f(x_{i+1}) - f(x_{\min})) \\ &\leq \epsilon((2C_{f,k}+\gamma C_k)\mathcal{H}_i + C_{f,k}\mathcal{H}_i^{(t)}),\end{aligned}$$

and by rearrangement this implies that

$$\mathcal{H}_i - \mathcal{H}_i^{(t)} \leq \frac{(3C_{f,k}+\gamma C_k)\epsilon}{1-(2C_{f,k}+\gamma C_k)\epsilon} \cdot \mathcal{H}_i^{(t)}. \tag{Ap.54}$$

Using this, the convexity of α, and Assumption A.4, we have

$$\alpha(3\mathcal{H}_i^{(t)}) - \alpha(3\mathcal{H}_i) \leq -3\alpha'(3\mathcal{H}_i^{(t)})(\mathcal{H}_i - \mathcal{H}_i^{(t)}) \leq -\alpha'(3\mathcal{H}_i^{(t)})3\mathcal{H}_i^{(t)} \cdot \frac{\epsilon(3C_{f,k}+\gamma C_k)}{1-(2C_{f,k}+\gamma C_k)\epsilon}$$
$$\leq \frac{1}{C_{\alpha,\gamma}} \cdot \frac{\epsilon(3C_{f,k}+\gamma C_k)}{1-(2C_{f,k}+\gamma C_k)\epsilon} \cdot \alpha(3\mathcal{H}_i^{(t)}),$$

and (Ap.50) follows by rearrangement. □

Now we are ready to prove the convergence bound.

Proposition 4.4.9 (Convergence bound for the second explicit scheme)**.** *Given f, k, γ, α, $C_{\alpha,\gamma}$, $C_{f,k}$, C_k, D_k, $D_{f,k}$, E_k, F_k satisfying assumptions A, B, and D, and that*

$$0 < \epsilon < \min\left(\frac{1-\gamma}{2(C_{f,k}+6D_{f,k}/C_{\alpha,\gamma})}, \frac{1-\gamma}{8D_k(1+E_k)}, \frac{C_{\alpha,\gamma}}{6(5C_{f,k}+2\gamma C_k)+12\gamma C_{\alpha,\gamma}}, \sqrt{\frac{1}{6\gamma^2 D_k F_k}}\right).$$

Let $\alpha_\star = \alpha(3\mathcal{H}_0)$, $\mathcal{W}_0 := f(x_0) - f(x_{\min})$, and for $i \geq 0$, let

$$\mathcal{W}_{i+1} = \mathcal{W}_i\left(1 - \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1-\gamma-2\epsilon(C_{f,k}+6D_{f,k}/C_{\alpha,\gamma})\right]\alpha(2\mathcal{W}_i)\right).$$

Then for any (x_0, p_0) with $p_0 = 0$, the iterates (4.39) satisfy for every $i \geq 0$,

$$f(x_i) - f(x_{\min}) \leq 2\mathcal{W}_i \leq 2\mathcal{W}_0 \cdot \left(1 - \frac{\epsilon C_{\alpha,\gamma}}{4}\left[1-\gamma-2\epsilon(C_{f,k}+6D_{f,k}/C_{\alpha,\gamma})\right]\alpha_\star\right)^i.$$

Proof. We follow the notations of Lemma Ap.3.1, and the proof is based on Lemma Ap.3.2. For the Hamiltonian terms, by the convexity of f and k, we have

$$\begin{aligned}\mathcal{H}_{i+1} - \mathcal{H}_i &= f(x_{i+1}) - f(x_i) + k(p_{i+1}) - k(p_i) \\ &\leq \langle \nabla f(x_{i+1}), x_{i+1} - x_i \rangle + \langle \nabla k(p_{i+1}), p_{i+1} - p_i \rangle \\ &= \langle \nabla f(x_{i+1}), x_{i+1} - x_i \rangle + \langle \nabla k(p_i), p_{i+1} - p_i \rangle + \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i \rangle \\ &= \epsilon \langle \nabla f(x_{i+1}), \nabla k(p_i) \rangle - \epsilon \langle \nabla k(p_i), \nabla f(x_{i+1}) + \gamma p_i \rangle + \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i \rangle \\ &= -\gamma\epsilon \langle \nabla k(p_i), p_i \rangle + \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i \rangle \qquad \text{(Ap.55)}\end{aligned}$$

for any $\epsilon > 0$. For the inner product terms, we have

$$\begin{aligned}&\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle \\ &= \langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_{i+1} - x_{\min}, p_i \rangle + \langle x_{i+1} - x_{\min}, p_i \rangle - \langle x_i - x_{\min}, p_i \rangle \\ &= \langle x_{i+1} - x_{\min}, -\epsilon\gamma p_i - \epsilon \nabla f(x_{i+1}) \rangle + \langle x_{i+1} - x_i, p_i \rangle \\ &= -\epsilon \langle \nabla f(x_{i+1}), x_{i+1} - x_{\min} \rangle - \epsilon\gamma \langle x_{i+1} - x_{\min} - (x_{i+1} - x_i), p_i \rangle + (1 - \epsilon\gamma) \langle x_{i+1} - x_i, p_i \rangle \\ &= -\epsilon \langle \nabla f(x_{i+1}), x_{i+1} - x_{\min} \rangle + \epsilon(1 - \epsilon\gamma) \langle \nabla k(p_i), p_i \rangle - \epsilon\gamma \langle x_i - x_{\min}, p_i \rangle . \qquad \text{(Ap.56)}\end{aligned}$$

Note that from assumption B.1 and the convexity of f it follows that

$$\begin{aligned}&-(f(x_{i+1}) - f(x_{\min})) \leq -(f(x_i) - f(x_{\min})) + f(x_i) - f(x_{i+1}) \\ &\leq -(f(x_i) - f(x_{\min})) + \langle \nabla f(x_i), -\epsilon \nabla k(p_i) \rangle \leq -(f(x_i) - f(x_{\min})) + \epsilon C_{f,K} \mathcal{H}_i. \qquad \text{(Ap.57)}\end{aligned}$$

By combining (Ap.11), (Ap.55), and (Ap.56), it follows that

$$\begin{aligned}\mathcal{V}_{i+1} - \mathcal{V}_i &= \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_{i+1}(\langle x_{i+1} - x_{\min}, p_{i+1} \rangle - \langle x_i - x_{\min}, p_i \rangle) + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle \\ &\leq -\gamma\epsilon \langle \nabla k(p_i), p_i \rangle + \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i \rangle + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle \\ &+ \epsilon\beta_{i+1} (- \langle \nabla f(x_{i+1}), x_{i+1} - x_{\min} \rangle + (1 - \epsilon\gamma) \langle \nabla k(p_i), p_i \rangle - \gamma \langle x_i - x_{\min}, p_i \rangle) \\ &\leq -\epsilon(\gamma - \beta_{i+1}) \langle \nabla k(p_i), p_i \rangle - \epsilon\beta_{i+1} \langle \nabla f(x_{i+1}), x_{i+1} - x_{\min} \rangle - \epsilon\gamma\beta_{i+1} \langle x_i - x_{\min}, p_i \rangle \\ &+ \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i \rangle + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle\end{aligned}$$

which can be further bounded using $\beta_{i+1} \leq \frac{\gamma}{2}$, the convexity of k and f, and Lemma Ap.3.4 as

$$\begin{aligned}&\leq -\epsilon(\gamma - \beta_{i+1} - \epsilon D) k(p_i) - \epsilon\beta_{i+1}(f(x_{i+1}) - f(x_{\min})) - \epsilon\gamma\beta_{i+1} \langle x_i - x_{\min}, p_i \rangle \\ &+ 2\epsilon^2 \beta_{i+1} \cdot C/C_{\alpha,\gamma} \cdot \mathcal{H}_i + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle\end{aligned}$$

and now using (Ap.57) and $\mathcal{H}_i \leq 2\mathcal{V}_i$ leads to

$$\begin{aligned}&\leq -\epsilon(\gamma - \beta_{i+1} - \epsilon D) k(p_i) - \epsilon\beta_{i+1}(f(x_i) - f(x_{\min})) - \epsilon\gamma\beta_{i+1} \langle x_i - x_{\min}, p_i \rangle \\ &+ \epsilon^2 \beta_{i+1} \cdot (4C/C_{\alpha,\gamma} + 2C_{f,k}) \cdot \mathcal{V}_i + (\beta_{i+1} - \beta_i) \langle x_i - x_{\min}, p_i \rangle ,\end{aligned}$$

implying that (Ap.17) holds with $C_1 = D$ and $C_2 = 4C/C_{\alpha,\gamma} + 2C_{f,k}$.

By combining (Ap.10), (Ap.55), and (Ap.56), it follows that

$$\begin{aligned}
&\mathcal{V}_{i+1} - \mathcal{V}_i \\
&\quad = \mathcal{H}_{i+1} - \mathcal{H}_i + \beta_i(\langle x_{i+1} - x_{\min}, p_{i+1}\rangle - \langle x_i - x_{\min}, p_i\rangle) + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle \\
&\quad \leq -\gamma\epsilon\langle \nabla k(p_i), p_i\rangle + \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i\rangle + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle \\
&\quad + \epsilon\beta_i\left(-\langle \nabla f(x_{i+1}), x_{i+1} - x_{\min}\rangle + (1 - \epsilon\gamma)\langle \nabla k(p_i), p_i\rangle - \gamma\langle x_i - x_{\min}, p_i\rangle\right) \\
&\quad \leq -\epsilon(\gamma - \beta_i)\langle \nabla k(p_i), p_i\rangle - \epsilon\beta_i\langle \nabla f(x_{i+1}), x_{i+1} - x_{\min}\rangle - \epsilon\gamma\beta_i\langle x_i - x_{\min}, p_i\rangle \\
&\quad + \langle \nabla k(p_{i+1}) - \nabla k(p_i), p_{i+1} - p_i\rangle + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle
\end{aligned}$$

which can be further bounded using $\beta_i \leq \frac{\gamma}{2}$, the convexity of k and f, and Lemma Ap.3.4 as

$$\begin{aligned}
&\leq -\epsilon(\gamma - \beta_i - \epsilon D)k(p_i) - \epsilon\beta_i(f(x_{i+1}) - f(x_{\min})) - \epsilon\gamma\beta_i\langle x_i - x_{\min}, p_i\rangle \\
&+ 2\epsilon^2\beta_i \cdot C/C_{\alpha,\gamma} \cdot \mathcal{H}_i + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle
\end{aligned}$$

and now using (Ap.57) and $\mathcal{H}_i \leq 2\mathcal{V}_i$ leads to

$$\begin{aligned}
&\leq -\epsilon(\gamma - \beta_i - \epsilon D)k(p_i) - \epsilon\beta_i(f(x_i) - f(x_{\min})) - \epsilon\gamma\beta_i\langle x_i - x_{\min}, p_i\rangle \\
&+ \epsilon^2\beta_i \cdot (4C/C_{\alpha,\gamma} + 2C_{f,k}) \cdot \mathcal{V}_i + (\beta_{i+1} - \beta_i)\langle x_{i+1} - x_{\min}, p_{i+1}\rangle
\end{aligned}$$

implying that (Ap.18) holds with $C_1 = D$ and $C_2 = 4C/C_{\alpha,\gamma} + 2C_{f,k}$ (see (Ap.40) for the definition of C and D). The claim of the lemma now follows by Lemma Ap.3.2. □

Ap.3.4 Explicit Method on Non-Convex f

Lemma 4.4.11 (Convergence of the first explicit scheme without convexity)**.** *Given $\|\cdot\|$, f, k, γ, b, D_k, D_f, σ satisfying assumptions E and A.2. If $\epsilon \in (0, \sqrt[b-1]{\gamma/D_f D_k}]$, then the iterates* (4.36) *of the first explicit method satisfy*

$$\mathcal{H}_{i+1} - \mathcal{H}_i \leq (\epsilon^b D_f D_k - \epsilon\gamma)k(p_{i+1}) \leq 0, \tag{4.47}$$

and $\|\nabla f(x_i)\|_2 \to 0$.

Proof. By assumption E.3

$$\mathcal{H}_{i+1} - \mathcal{H}_i \leq k(p_{i+1}) - k(p_i) + \epsilon\langle \nabla f(x_i), \nabla k(p_{i+1})\rangle + D_f\sigma(\epsilon\|\nabla k(p_{i+1})\|)$$

now with convexity of k

$$\begin{aligned}&\leq -\epsilon\gamma \left\langle \nabla k(p_{i+1}), p_{i+1}\right\rangle + D_f\sigma(\epsilon \left\|\nabla k(p_{i+1})\right\|)\\&\leq -\epsilon\gamma k(p_{i+1}) + D_f\sigma(\epsilon \left\|\nabla k(p_{i+1})\right\|)\end{aligned}$$

we have $\sigma(\epsilon t) \leq \epsilon^b \sigma(t)$ and $\sigma(\|\nabla k(p)\|) \leq D_k k(p)$ by assumption E.3

$$\leq (\epsilon^b D_f D_k - \epsilon\gamma)k(p_{i+1})$$

If $\epsilon \leq \sqrt[b-1]{\gamma/D_f D_k}$, then $\mathcal{H}_{i+1} \leq \mathcal{H}_i$. If $\mathcal{H}$ is bounded below we get that x_i, p_i is such that $k(p_i) \to 0$ and thus $p_i \to 0$. Since $\left\|p_{i+1}\right\|_2 + \delta \left\|p_i\right\|_2 \geq \epsilon\delta \left\|\nabla f(x_i)\right\|_2$, we get $\|\nabla f(x_i)\| \to 0$. □

Ap.4 Proofs for power kinetic energies

The proofs of the results in this section will be based on the following three preliminary lemmas.

Lemma Ap.4.1. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and $x \in \mathbb{R}^d \setminus \{0\}$. If $\|x\|$ is differentiable, then*

$$\left\|\nabla \|x\|\right\|_* = 1 \qquad \left\langle \nabla \|x\|, x\right\rangle = \|x\|, \tag{Ap.58}$$

and if $\|x\|$ is twice differentiable, then

$$(\nabla^2 \|x\|)x = 0. \tag{Ap.59}$$

Proof. Let $x \in \mathbb{R}^d \setminus \{0\}$. By the convexity of the norm we have

$$\left\langle \nabla \|x\|, y\right\rangle - \|y\| \leq \left\langle \nabla \|x\|, x\right\rangle - \|x\|$$

for all $y \in \mathbb{R}^d$. Thus

$$\sup_{y\in\mathbb{R}^d}\{\left\langle \nabla \|x\|, y\right\rangle - \|y\|\} \leq \left\langle \nabla \|x\|, x\right\rangle - \|x\|$$

Because the right hand side is finite, we must have $\left\|\nabla \|x\|\right\|_* \leq 1$ and the left hand side equal to 0.

$$0 \leq \left\langle \nabla \|x\|, x\right\rangle - \|x\| \leq \left\|\nabla \|x\|\right\|_* \|x\| - \|x\| \leq 0$$

forces $\left\|\nabla \|x\|\right\|_* = 1$ and $\left\langle \nabla \|x\|, x\right\rangle = \|x\|$. In fact, this argument goes through for non-differentiable $\|\cdot\|$, by definition of the subderivative. For twice differentiable norms, take the derivative of $\left\langle \nabla \|x\|, x\right\rangle = \|x\|$ to get

$$(\nabla^2 \|x\|)x + \nabla \|x\| = \nabla \|x\|$$

and our result follows. □

Lemma Ap.4.2. *Given $a \in [1,\infty)$, $A \in [1,\infty)$, and φ_a^A in (4.48). Define $B = A/(A-1)$, $b = a/(a-1)$. For convenience, define*

$$\varphi(t) = \varphi_a^A(t) \qquad \phi(t) = \varphi_b^B(t). \tag{Ap.60}$$

The following hold.

1. *Monotonicity. For $t \in (0,\infty)$, $\varphi'(t) > 0$. If $a = A = 1$, then for $t \in (0,\infty)$, $\varphi''(t) = 0$, otherwise $\varphi''(t) > 0$. This implies that φ is strictly increasing on $[0,\infty)$.*

2. *Subhomogeneity. For all $t, \epsilon \in [0,\infty)$,*

$$\varphi(\epsilon t) \le \max\{\epsilon^a, \epsilon^A\}\varphi(t) \tag{Ap.61}$$

   with equality iff $a = A$ or $t = 0$ or $\epsilon = 0$ or $\epsilon = 1$.

3. *Strict Convexity. If $a > 1$ or $A > 1$, then $\varphi(t)$ is strictly convex on $[0,\infty)$ with a unique minimum at 0.*

4. *Derivatives. For all $t \in (0,\infty)$,*

$$\min\{a, A\}\varphi(t) \le t\varphi'(t) \le \max\{a, A\}\varphi(t), \tag{Ap.62}$$

$$(\min\{a, A\} - 1)\varphi'(t) \le t\varphi''(t) \le (\max\{a, A\} - 1)\varphi'(t). \tag{Ap.63}$$

   If $a, A \ge 2$, then for all $t, s \in (0,\infty)$,

$$\varphi(t) \le s\varphi'(s) + (\varphi'(t) - \varphi'(s))(t - s) \tag{Ap.64}$$

Proof. First, for $t \in (0,\infty)$, the following identities can be easily verified.

$$t\varphi'(t) = (t^a + 1)^{\frac{A-a}{a}} t^a \tag{Ap.65}$$

$$t\varphi''(t) = \varphi'(t)\left(a - 1 + (A - a)\tfrac{t^a}{t^a+1}\right) \tag{Ap.66}$$

1. *Monotonicity.* First, for $t > 0$ we have,

$$\varphi'(t) = (t^a + 1)^{\frac{A-a}{a}} t^{a-1} > 0 \tag{Ap.67}$$

   For $\varphi'(t)$ for $t > 0$, we have the following with equality iff $a = A = 1$

$$\varphi''(t) = t^{-1}\varphi'(t)\left(a - 1 + (A - a)\tfrac{t^a}{t^a+1}\right) \ge 0. \tag{Ap.68}$$

   Finally, $\varphi(t)$ is continuous at 0, which gives our result.

2. *Subhomogeneity.* If $a = A$ or $t = 0$ or $\epsilon = 0$ or $\epsilon = 1$, then equality clearly holds. Assume $a \neq A$, $t, \epsilon > 0$, and $\epsilon \neq 1$. Assuming $A > a$, $t^{\frac{A-a}{a}}$ is strictly increasing. If $\epsilon < 1$, then

$$\epsilon\varphi'(\epsilon t) = \epsilon^a(\epsilon^a t^a + 1)^{\frac{A-a}{a}} t^{a-1} < \epsilon^a(t^a + 1)^{\frac{A-a}{a}} t^{a-1}.$$

If $\epsilon > 1$, then

$$\epsilon\varphi'(\epsilon t) = \epsilon^A(t^a + \epsilon^{-a})^{\frac{A-a}{a}} t^{a-1} < \epsilon^A(t^a + 1)^{\frac{A-a}{a}} t^{a-1}.$$

Integrating both sides gives $\varphi(\epsilon t) < \max\{\epsilon^a, \epsilon^A\}\varphi(t)$. The case $A < a$ follows similarly, using the fact that $t^{\frac{A-a}{a}}$ is strictly decreasing.

3. *Strict convexity.* First, since φ is strictly increasing we get $\varphi(t) > \varphi(0) = 0$, which proves that 0 is the unique minimizer. Our goal is to prove that for $t, s \in [0, \infty)$ and $\epsilon \in (0, 1)$ such that $t \neq s$,

$$\varphi(\epsilon t + (1 - \epsilon)s) < \epsilon\varphi(t) + (1 - \epsilon)\varphi(s)$$

First, for $t = 0$ or $s = 0$, this reduces to a condition of the form $\varphi(\epsilon t) < \epsilon\varphi(t)$ for all $t \in [0, \infty)$ and $\epsilon \in (0, 1)$. Considering separately the cases $A = 1, a > 1$ and $a = 1, A > 1$ and $a, A > 1$, it is easy to see that this follows from the subhomogeneity result (Ap.61). For $s, t > 0$, our result follows from the positivity of φ'', (Ap.68).

4. *Derivatives.* Since,

$$\min\{a, A\} - 1 \leq a - 1 + (A - a)\frac{t^a}{t^a + 1} \leq \max\{a, A\} - 1,$$

we get the second derivative bound (Ap.63) from identity (Ap.82). The first derivative bound (Ap.62) follows from (Ap.63), since

$$t\varphi'(t) = \int_0^t \varphi'(t) + t\varphi''(t)\, dt.$$

Our goal is now to prove the uniform gradient bound (Ap.64) for $a, A \geq 2$. In the case that $0 < t < s$, the bound reduces to $(\varphi'(t) - \varphi'(s))(t - s) \geq 0$, which follows from convexity. The remaining case is $0 < s \leq t$. Notice that for the case $0 < s < t$, convexity implies

$$\varphi'(s) \leq \frac{\varphi(t) - \varphi(s)}{t - s} \leq \varphi'(t) \tag{Ap.69}$$

Notice that in the case $s = 0$ for (Ap.69) we get the inequality $\varphi(t) \leq t\varphi'(t)$, again a condition of convexity. On the other hand, we have just shown that for our φ the stronger inequality $\min\{a, A\}\varphi(t) \leq t\varphi'(t)$ holds. This motivates a strategy of searching for a stronger bound of form (Ap.69), and using this to derive the uniform gradient bound (Ap.64). Indeed, let $\lambda = t/s > 1$, then we will show

$$\sigma(\lambda)\varphi'(s) \leq \frac{\varphi(\lambda s) - \varphi(s)}{\lambda s - s} \leq \varphi'(\lambda s)\tau(\lambda) \tag{Ap.70}$$

where

$$\sigma(\lambda) = \begin{cases} \frac{\lambda^a - 1}{a(\lambda - 1)} & A \geq a \\ \frac{\lambda^A - 1}{A(\lambda - 1)} & A \leq a \end{cases} \qquad \tau(\lambda) = \begin{cases} \frac{\lambda(1 - \lambda^{-a})}{a(\lambda - 1)} & A \geq a \\ \frac{\lambda(1 - \lambda^{-A})}{A(\lambda - 1)} & A \leq a \end{cases} \tag{Ap.71}$$

First, assume $A \geq a$. We need to prove

$$\frac{\lambda^a - 1}{a} s\varphi'(s) \leq \varphi(\lambda s) - \varphi(s) \leq \frac{1 - \lambda^{-a}}{a} \lambda s \varphi'(\lambda s)$$

We fix $s > 0$, and take $F_1(\lambda) := \varphi(\lambda s) - \varphi(s) - \frac{\lambda^a - 1}{a} s\varphi'(s)$ and $F_2(\lambda) := \frac{1 - \lambda^{-a}}{a} \lambda s \varphi'(\lambda s) - \varphi(\lambda s) + \varphi(s)$. We need to prove that $F_1(\lambda) \geq 0$ and $F_2(\lambda) \geq 0$ for $\lambda \geq 1$. We have $F_1(1) = F_2(1) = 0$,

$$F_1'(\lambda) = \frac{(\lambda s)^a}{\lambda} (((\lambda s)^a + 1)^{\frac{A-a}{a}} - (s^a + 1)^{\frac{A-a}{a}}) \geq 0$$

and

$$\begin{aligned} F_2'(\lambda) &= \frac{(1 - \lambda^{-a})s}{a} \left(\varphi'(\lambda s) + \lambda s \varphi''(\lambda s) - a\varphi'(\lambda s)\right) \\ &= \frac{(1 - \lambda^{-a})s}{a} \varphi'(\lambda s)(A - a) \frac{(\lambda s)^a}{(\lambda s)^a + 1} \geq 0, \end{aligned}$$

so indeed $F_1(\lambda) \geq 0$ and $F_2(\lambda) \geq 0$ for every $\lambda > 1$.

Now let $A \leq a$. Then we need to prove that

$$\frac{\lambda^A - 1}{A} s\varphi'(s) \leq \varphi(\lambda s) - \varphi(s) \leq \frac{1 - \lambda^{-A}}{A} \lambda s \varphi'(\lambda s).$$

We fix $s > 0$, and take $F_3(\lambda) := \varphi(\lambda s) - \varphi(s) - \frac{\lambda^A - 1}{A} s\varphi'(s)$ and $F_4(\lambda) := \frac{1 - \lambda^{-A}}{A} \lambda s \varphi'(\lambda s) - \varphi(\lambda s) + \varphi(s)$. We need to prove that $F_3(\lambda) \geq 0$ and $F_4(\lambda) \geq 0$ for $\lambda \geq 1$. We have $F_3(1) = F_4(1) = 0$,

$$F_3'(\lambda) = \frac{(\lambda s)^a}{\lambda} (((\lambda s)^a + 1)^{-\frac{a-A}{a}} - ((\lambda s)^a + \lambda^a)^{-\frac{a-A}{a}}) \geq 0$$

and

$$F_4'(\lambda) = \frac{(1-\lambda^{-A})s}{A}\left(\varphi'(\lambda s) + \lambda s\varphi''(\lambda s) - A\varphi'(\lambda s)\right)$$
$$= \frac{(1-\lambda^{-A})s}{A}\varphi'(\lambda s)\left(a - A + (A-a)\frac{(\lambda s)^a}{(\lambda s)^a + 1}\right) \geq 0,$$

so indeed $F_3(\lambda) \geq 0$ and $F_4(\lambda) \geq 0$ for every $\lambda > 1$.

Now, we can prove (Ap.64). We have so far proven the following inequalities in $\varphi(t), \varphi(s), \varphi'(t), \varphi'(s)$:

$$\varphi(s) \geq 0, \qquad s\varphi'(s) - \min(a, A)\varphi(s) \geq 0,$$
$$\varphi(t) - \varphi(s) - (\lambda-1)\phi(\lambda)s\varphi'(s) \geq 0, \quad (\lambda-1)\tau(\lambda)s\varphi'(t) - \varphi(t) + \varphi(s) \geq 0.$$

We try to express the inequality $s\varphi'(s) + (t-s)(\varphi'(t) - \varphi'(s)) - \varphi(t) \geq 0$ as a linear combination of the above four inequalities with non-negative coefficients:

$$s\varphi'(s) + (t-s)(\varphi'(t) - \varphi'(s)) - \varphi(t) = c_1\varphi(s) + c_2(s\varphi'(s) - \min(a, A)\varphi(s))$$
$$+ c_3(\varphi(t) - \varphi(s) - (\lambda-1)\sigma(\lambda)s\varphi'(s)) + c_4((\lambda-1)\tau(\lambda)s\varphi'(t) - \varphi(t) + \varphi(s)).$$

Comparing the coefficients of $\varphi(s)$, $\varphi(t)$, $\varphi'(s)$, $\varphi'(t)$, we get the following equations:

$$c_1 - \min(a, A)c_2 - c_3 + c_4 = 0,$$
$$c_3 - c_4 = -1,$$
$$c_2 - (\lambda-1)\sigma(\lambda)c_3 = 2 - \lambda,$$
$$(\lambda-1)\tau(\lambda)c_4 = \lambda - 1.$$

This system of equations has a unique solution: $c_4 = \frac{1}{\tau(\lambda)}$, $c_3 = \frac{1}{\tau(\lambda)} - 1$, $c_2 = 2 - \lambda + (\lambda-1)\sigma(\lambda)(\frac{1}{\tau(\lambda)} - 1)$ and $c_1 = \min(a, A)c_2 - 1$. We will prove that $c_1, c_2, c_3, c_4 \geq 0$. Clearly $\tau(\lambda) > 0$. We claim that $\tau(\lambda) \leq 1$. For this it is enough to check that $\lambda(1-\lambda^{-\alpha}) \leq \alpha(\lambda-1)$ for every $\lambda > 1$ and $\alpha \geq 2$. After reordering the terms, we get $1 + (1-\alpha)(\lambda-1) \leq \lambda^{1-\alpha} = (1 + (\lambda-1))^{1-\alpha}$, which follows from the generalized Bernoulli inequality. So $0 < \sigma(\lambda) \leq 1$, therefore $c_3, c_4 \geq 0$. We just need to prove now that $\min(a, A)c_2 \geq 1$, because then $c_1, c_2 \geq 0$. If $a \leq A$, then $c_1 = \min(a, A)c_2 - 1 = a(2 - \lambda + \lambda^a - \lambda^{a-1} - \frac{\lambda^a}{a})$. If $a \geq A$, then $c_1 = A(2 - \lambda + \lambda^A - \lambda^{A-1} - \frac{\lambda^A}{A})$. So the only remaining thing to show is

$$2\alpha - \alpha\lambda + \alpha\lambda^\alpha - \alpha\lambda^{\alpha-1} - \lambda^\alpha \geq 0$$

for every $\lambda > 1$ and $\alpha \geq 2$. Letting $\epsilon = 1/\lambda$, this is equivalent to showing

$$2\alpha\epsilon^{\alpha} - \alpha\epsilon^{\alpha-1} + \alpha - \alpha\epsilon - 1 \geq 0$$

for $\epsilon \in (0, 1)$. Let $\pi(\epsilon) = 2\alpha\epsilon^{\alpha} - \alpha\epsilon^{\alpha-1} + \alpha - \alpha\epsilon - 1$. To see that $\pi(\epsilon) \geq 0$, note

$$\pi'(\epsilon) = \alpha\epsilon^{\alpha-2}(2\alpha\epsilon - \alpha + 1) - \alpha$$

from which $\pi'(\epsilon) < \pi'(1/2) \leq 0$ for $\epsilon < 1/2$ and $\pi'(\epsilon) > \pi'((\alpha-1)/\alpha) \geq 0$ for $\epsilon > \alpha/(\alpha-1)$. This implies that π is minimized on $[1/2, (\alpha-1)/\alpha]$. Our result follows then from the fact that for $\epsilon \in [1/2, (\alpha-1)/\alpha]$,

$$\pi(\epsilon) \geq \alpha - \alpha\epsilon - 1 \geq 0$$

□

Lemma Ap.4.3. *Given $a \in [1, \infty)$, $A \in [1, \infty)$, and φ_a^A in (4.48). Define $B = A/(A-1)$, $b = a/(a-1)$. For convenience, define*

$$\varphi(t) = \varphi_a^A(t) \qquad \phi(t) = \varphi_b^B(t). \tag{Ap.72}$$

For $t \in [0, \infty)$ define the function

$$\rho(t) = \left(\frac{t^a}{t^a + 1} + (t^a + 1)^{-\frac{A-1}{a-1}} \right)^{\frac{a-A}{a(A-1)}}, \tag{Ap.73}$$

and for $a \neq A$ define the constant

$$C_{a,A} = \left(1 - \left(\tfrac{a-1}{A-1}\right)^{\frac{a-1}{A-a}} + \left(\tfrac{a-1}{A-1}\right)^{\frac{A-1}{A-a}} \right)^{\frac{B-b}{b}}. \tag{Ap.74}$$

We have the following results. For all $t \in (0, \infty)$,

$$\phi'(\varphi'(t)) = \rho(t)t, \tag{Ap.75}$$

which means that ρ captures the relative error between $(\varphi^*)'$ and ϕ', because $(\varphi^*)'(t) = (\varphi')^{-1}(t)$. Finally, ρ is bounded for all $t \in (0, \infty)$ between the constants,

$$1 \leq \rho(t) \leq C_{a,A} \tag{Ap.76}$$

Proof. We will show the results backwards, starting with (Ap.76). By rearrangement,

$$\rho(t) = (\tfrac{t^a}{t^a+1} + (t^a+1)^{-1}(t^a+1)^{\frac{a-A}{a-1}})^{\frac{a-A}{a(A-1)}}.$$

If $a \geq A$ we have $(t^a+1)^{\frac{a-A}{a-1}} \geq 1$ and $t^{\frac{a-A}{a(A-1)}}$ is increasing, so $\rho(t) \geq 1$. If $a < A$ we have $(t^a+1)^{\frac{a-A}{a-1}} \leq 1$ and $t^{\frac{a-A}{a(A-1)}}$ is decreasing, so again $\rho(t) \geq 1$. This proves the left hand inequality of (Ap.76). Now, assume that $A \neq a$. Looking at $\frac{t^a}{t^a+1} + (t^a+1)^{-\frac{A-1}{a-1}}$ we have

$$\left[\frac{t^a}{t^a+1} + (t^a+1)^{-\frac{A-1}{a-1}}\right]' = \frac{at^{a-1}}{(t^a+1)^2} - \frac{A-1}{a-1}(t^a+1)^{-\frac{A-1}{a-1}-1}at^{a-1}$$

Since $t \neq 0$ we see that it has a stationary point at

$$(t^a+1)^{-1} - \frac{A-1}{a-1}(t^a+1)^{-\frac{A-1}{a-1}} = 0,$$

which is equivalent to $(t^a+1) = \left(\frac{A-1}{a-1}\right)^{\frac{a-1}{A-a}}$. This is also a stationary point of $\rho(t)$. Since $\rho(0) = \rho(\infty) = 1$ and $\rho(t) \geq 1$ this stationary point must be a maximum. Thus

$$\begin{aligned}\rho(t) &= \left(1 - (t^a+1)^{-1} + (t^a+1)^{-\frac{A-1}{a-1}}\right)^{\frac{B-b}{b}} \\ &\leq \left(1 - \left(\tfrac{a-1}{A-1}\right)^{\frac{a-1}{A-a}} + \left(\tfrac{a-1}{A-1}\right)^{\frac{A-1}{A-a}}\right)^{\frac{B-b}{b}}\end{aligned}$$

This proves the right hand inequality of (Ap.76). For (Ap.75), since $(b-1)(a-1) = ab - a - b + 1 = 1$, we have,

$$\begin{aligned}\phi'(\varphi'(t)) &= ([(t^a+1)^{\frac{A-a}{a}}t^{a-1}]^b + 1)^{\frac{B-b}{b}}[(t^a+1)^{\frac{A-a}{a}}t^{a-1}]^{b-1} \\ &= ((t^a+1)^{\frac{A-a}{a-1}}t^a + 1)^{\frac{B-b}{b}}(t^a+1)^{\frac{A-a}{a(a-1)}}t\end{aligned}$$

we have $\frac{B-b}{b} = \frac{A(a-1)-a(A-1)}{(a-1)(A-1)b} = \frac{a-A}{a(A-1)}$ and thus

$$\begin{aligned}&= ((t^a+1)^{\frac{A-a}{a-1}}t^a + 1)^{\frac{a-A}{a(A-1)}}(t^a+1)^{\frac{A-a}{a(a-1)}}t \\ &= \rho(t)t\end{aligned}$$

□

Now we are ready to prove the key results in this section.

Lemma 4.5.1 (Verifying assumptions on k)**.** *Given a norm $\|p\|_*$ on $p \in \mathbb{R}^d$, $a, A \in [1,\infty)$, and φ_a^A in (4.48). Define the constant,*

$$C_{a,A} = \left(1 - \left(\tfrac{a-1}{A-1}\right)^{\frac{a-1}{A-a}} + \left(\tfrac{a-1}{A-1}\right)^{\frac{A-1}{A-a}}\right)^{\frac{B-b}{b}}. \tag{4.50}$$

$k(p) = \varphi_a^A(\|p\|_*)$ satisfies the following.

1. *Convexity. If $a > 1$ or $A > 1$, then k is strictly convex with a unique minimum at $0 \in \mathbb{R}^d$.*

2. *Conjugate. For all $x \in \mathbb{R}^d$, $k^*(x) = (\varphi_a^A)^*(\|x\|)$.*

3. *Gradient. If $\|p\|_*$ is differentiable at $p \in \mathbb{R}^d \setminus \{0\}$ and $a > 1$, then k is differentiable for all $p \in \mathbb{R}^d$, and for all $p \in \mathbb{R}^d$,*

$$\langle \nabla k(p), p \rangle \leq \max\{a, A\} k(p), \tag{4.51}$$

$$(\varphi_a^A)^*(\|\nabla k(p)\|) \leq (\max\{a, A\} - 1) k(p). \tag{4.52}$$

Additionally, if $a, A > 1$, define $B = A/(A-1)$, $b = a/(a-1)$, and then

$$\varphi_b^B(\|\nabla k(p)\|) \leq C_{a,A}(\max\{a, A\} - 1) k(p). \tag{4.53}$$

Additionally, if $a, A \geq 2$, then for all $p, q \in \mathbb{R}^d$,

$$k(p) \leq \langle \nabla k(q), q \rangle + \langle \nabla k(p) - \nabla k(q), p - q \rangle . \tag{4.54}$$

4. *Hessian. If $\|p\|_*$ is twice continuously differentiable at $p \in \mathbb{R}^d \setminus \{0\}$, then k is twice continuously differentiable for all $p \in \mathbb{R}^d \setminus \{0\}$, and for all $p \in \mathbb{R}^d \setminus \{0\}$,*

$$\left\langle p, \nabla^2 k(p) p \right\rangle \leq \max\{a, A\}(\max\{a, A\} - 1) k(p). \tag{4.55}$$

Additionally, if $a, A \geq 2$ and there exists $N \in [0, \infty)$ such that $\|p\|_* \lambda_{\max}^{\|\cdot\|_*}(\nabla^2 \|p\|_*) \leq N$ for $p \in \mathbb{R}^d \setminus \{0\}$, then for all $p \in \mathbb{R}^d \setminus \{0\}$

$$(\varphi_{a/2}^{A/2})^* \left(\frac{\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 k(p))}{\max\{a, A\} - 1 + N} \right) \leq (\max\{a, A\} - 2) k(p). \tag{4.56}$$

Proof. Again, for the purposes of this proof, let $\varphi(t) = \varphi_a^A(t)$ and $\phi(t) = \varphi_b^B(t)$ for $t \in [0, \infty)$.

1. *Convexity.* First, since norms are positive definite and φ uniquely minimized at $0 \in \mathbb{R}$ by Lemma Ap.4.2, this proves that $0 \in \mathbb{R}^d$ is a unique minimizer of k. Let $\epsilon \in (0, 1)$ and $p, q \in \mathbb{R}^d$ such that $p \neq q$. By the monotonicity proved in Lemma Ap.4.2 and the triangle inequality

$$\begin{aligned} k(\epsilon p + (1-\epsilon) q) &= \varphi(\|\epsilon p + (1-\epsilon) q\|_*) \\ &\leq \varphi(\epsilon \|p\|_* + (1-\epsilon) \|q\|_*) \end{aligned}$$

and finally, by the strict convexity proved in Lemma Ap.4.2

$$< \epsilon k(p) + (1-\epsilon) k(q).$$

2. *Conjugate.* Let $x \in \mathbb{R}^d$. First, by definition of the convex conjugate and the dual norm,

$$\begin{aligned} k^*(x) &= \sup_{p\in\mathbb{R}^d}\{\langle x,p\rangle - k(p)\} = \sup_{p\in\mathbb{R}^d}\{\langle x,p\rangle - \varphi(\|p\|_*)\} \\ &= \sup_{t\geq 0}\sup_{\|p\|_*=t}\{\langle x,p\rangle - \varphi(t)\} = \sup_{t\geq 0}\{t\,\|x\| - \varphi(t)\} = \varphi^*(\|x\|) \end{aligned}$$

3. *Gradient.* First we argue for differentiability. For $p = 0$ (or $q = 0$ in the case of (4.54)), we have by the equivalence of the norms that there exists $c > 0$ such that $\|p\|_* < c\,\|p\|_2$. Thus, $\lim_{\|p\|_2\to 0} k(p)\,\|p\|_2^{-1} \leq \lim_{\|p\|_2\to 0}\varphi(c\,\|p\|_2)\,\|p\|^{-1} = c\lim_{t\to 0}\varphi(t)t^{-1} = 0$, and thus we have $\nabla k(0) = 0$. Now for $p \neq 0$, we have $\|p\|_* \neq 0$. Since $\varphi(t)$ is differentiable for $t > 0$ and $\|p\|_*$ at $p \neq 0$, we have by the chain rule $\nabla k(p) = \varphi'(\|p\|_*)\nabla\,\|p\|_*$.

   All four results follow trivially when $p = 0$. In particular, (4.54) reduces to $k(p) \leq \langle \nabla k(p), p\rangle$ for $q = 0$ and $0 \leq \langle \nabla k(q), q\rangle$ for $p = 0$; both follow from convexity.

   Now, assume $p \neq 0$. For (4.51), (4.52), and (4.53) we have, by Lemma Ap.4.1, $\langle \nabla k(p), p\rangle = \|p\|_*\,\varphi'(\|p\|_*)$ and $\varphi^*(\|\nabla k(p)\|) = \varphi^*(\varphi'(\|p\|_*))$ and $\phi(\|\nabla k(p)\|) = \phi(\varphi'(\|p\|_*))$. Letting $t = \|p\|_* > 0$, (4.51) follows directly from (Ap.62) of Lemma Ap.4.2. For (4.52), we have from convex analysis (5.7) that

$$\varphi^*(\varphi'(t)) = t\varphi'(t) - \varphi(t). \tag{Ap.77}$$

   This implies that $\varphi^*(\varphi'(t)) \leq (\max\{a, A\} - 1)\varphi(t)$, again by (Ap.62) of Lemma Ap.4.2. For (4.53) assume $a, A > 1$ and consider that by (Ap.63) of Lemma Ap.4.2 and (Ap.76) of Lemma Ap.4.3,

$$[\phi(\varphi'(t))]' = \phi'(\varphi'(t))\varphi''(t) = \rho(t)t\varphi''(t) \leq \varphi'(t)C_{a,A}(\max\{a, A\} - 1)$$

   Integrating both sides of this inequality gives $\phi(\varphi'(t)) \leq C_{a,A}(\max\{a, A\} - 1)\varphi(t)$.

   Finally, for the uniform gradient bound (4.54), assume $p \neq 0$ and $q \neq 0$. Lemma Ap.4.1 implies by Cauchy-Schwartz that $-\langle \nabla\,\|p\|_*, q\rangle \geq -\|q\|$ for any $p, q \in \mathbb{R}^d \setminus \{0\}$. Thus by Lemma Ap.4.1,

$$\begin{aligned} &\langle \nabla k(q), q\rangle + \langle \nabla k(p) - \nabla k(q), p - q\rangle \geq \\ &\quad \varphi'(\|q\|_*)\,\|q\|_* + (\varphi'(\|p\|_*) - \varphi'(\|q\|_*))(\|p\|_* - \|q\|_*) \end{aligned}$$

   and our result is implied by the one dimensional result (Ap.64) of Lemma Ap.4.2.

4. *Hessian.* Throughout, assume $p \in \mathbb{R}^d \backslash \{0\}$. First we argue for twice differentiability. We have $\nabla k(p) = \varphi'(\|p\|_*) \nabla \|p\|_*$, which for $p \neq 0$ is a product of a differentiable function and a composition of differentiable functions. Thus, we have differentiability, and by the chain rule,

$$\nabla^2 k(p) = \varphi''(\|p\|_*) \nabla \|p\|_* \nabla \|p\|_*^T + \varphi'(\|p\|_*) \nabla^2 \|p\|_* \tag{Ap.78}$$

All of these terms are continuous at $p \neq 0$ by assumption or inspection of (Ap.82).

We study (Ap.78). For (4.55), we have by Lemma Ap.4.1 and (Ap.62),(Ap.63) of Lemma Ap.4.2,

$$\begin{aligned}
\langle p, \nabla^2 k(p) p \rangle &= \varphi''(\|p\|_*) \left\langle p, \nabla \|p\|_* \nabla \|p\|_*^T p \right\rangle + \varphi'(\|p\|_*) \langle p, \nabla^2 \|p\|_* p \rangle \\
&= \varphi''(\|p\|_*) \|p\|_*^2 \\
&\leq \max\{a, A\}(\max\{a, A\} - 1)\varphi(\|p\|_*).
\end{aligned}$$

For (4.56) first note, by Lemma Ap.4.1

$$\left\langle v, \nabla \|p\|_* \nabla \|p\|_*^T v \right\rangle = (\langle v, \nabla \|p\|_* \rangle)^2 \leq \|v\|_*^2$$

and further $\left\langle p, \nabla \|p\|_* \nabla \|p\|_*^T p \right\rangle = \|p\|_*^2$. Thus $\lambda_{\max}^{\|\cdot\|_*}\left(\nabla \|p\|_* \nabla \|p\|_*^T\right) = 1$. Together, along with our assumption on the Hessian of $\|p\|_*$, we have

$$\begin{aligned}
\lambda_{\max}^{\|\cdot\|_*}\left(\nabla^2 k(p)\right) &\leq \varphi''(\|p\|_*) \lambda_{\max}^{\|\cdot\|_*}\left(\nabla \|p\|_* \nabla \|p\|_*^T\right) + \varphi'(\|p\|_*) \lambda_{\max}^{\|\cdot\|_*}\left(\nabla^2 \|p\|_*\right) \\
&\leq \varphi''(\|p\|_*) + N\varphi'(\|p\|_*) \|p\|_*^{-1}
\end{aligned}$$

and by (Ap.63) of Lemma Ap.4.2

$$\leq \varphi'(\|p\|_*) \|p\|_*^{-1} (\max\{a, A\} - 1 + N)$$

On the other hand, by Lemma Ap.4.1 and the monotonicity of Lemma Ap.4.2,

$$\begin{aligned}
&\lambda_{\max}^{\|\cdot\|_*}\left(\nabla^2 k(p)\right) \\
&\geq \varphi''(\|p\|_*) \left\langle \frac{p}{\|p\|_*}, \nabla \|p\|_* \nabla \|p\|_*^T \frac{p}{\|p\|_*} \right\rangle + \varphi'(\|p\|_*) \left\langle \frac{p}{\|p\|_*}, \nabla^2 \|p\|_* \frac{p}{\|p\|_*} \right\rangle \\
&= \varphi''(\|p\|_*) > 0
\end{aligned}$$

Taken together, we have

$$0 < \frac{\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 k(p))}{\max\{a, A\} - 1 + N} \leq \varphi'(\|p\|_*) \|p\|_*^{-1}$$

Now, assume $a, A \geq 2$ and let $t = \|p\|_* > 0$ and $\chi(t) = \varphi_{a/2}^{A/2}(t)$. Our goal is to show that

$$\chi^*\left(\frac{\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 k(p))}{\max\{a, A\} - 1 + N}\right) \leq (\max\{a, A\} - 2)\varphi(\|p\|_*)$$

To do this we argue that $\chi^*(t)$ is an non-decreasing function on $(0, \infty)$ and $\chi^*(\varphi'(t)t^{-1}) \leq (\max\{a, A\} - 2)\varphi(t)$, from which our result would follow. First, we have for $r \in [0, \infty)$ and $s \in (0, \infty)$ such that for $t \geq s$,

$$\chi^*(t) \geq tr - \chi(r) \geq sr - \chi(r)$$

Taking the supremum in r returns the result that χ^* is non-decreasing. Otherwise it can be verified directly from (4.57) – (4.61). Thus, what remains to show is $\chi^*(\varphi'(t)t^{-1}) \leq (\max\{a, A\}-2)\varphi(t)$. Note that $\chi(t^2) = 2\varphi(t)$ and $\chi'(t^2)t = \varphi'(t)$. Using (5.7) of convex analysis and (Ap.62) of Lemma Ap.4.2,

$$\begin{aligned}\chi^*(\varphi(t)t^{-1}) &= \chi^*(\chi'(t^2)) = t^2\chi'(t^2) - \chi(t^2) \\ &\leq (\max\{\tfrac{a}{2}, \tfrac{A}{2}\} - 1)\chi(t^2) = (\max\{a, A\} - 2)\varphi(t)\end{aligned}$$

from which our result follows.

□

Lemma 4.5.3 (Bounds on $\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 \|p\|_*)$ for b-norms)**.** *Given $b \in [2, \infty)$, let $\|x\|_b = \left(\sum_{n=1}^d |x^{(n)}|^b\right)^{1/b}$ for $x \in \mathbb{R}^d$. Then for $x \in \mathbb{R}^d \setminus \{0\}$,*

$$\|x\|_b \, \lambda_{\max}^{\|\cdot\|_b}\big(\nabla^2 \|x\|_b\big) \leq (b - 1).$$

Proof. A short calculation reveals that

$$\nabla^2 \|x\|_b = \frac{(b-1)}{\|x\|_b}\left(\operatorname{diag}\left(\frac{|x^{(n)}|^{b-2}}{\|x\|_b^{b-2}}\right) - \nabla \|x\|_b \nabla \|x\|_b^T\right) \tag{Ap.79}$$

Thus, since $\langle b, aa^T b\rangle = \langle a, b\rangle^2 \geq 0$ for any $a, b \in \mathbb{R}^d$, we have $\lambda_{\max}^{\|\cdot\|_b}\Big((1-b)\nabla \|x\|_b \nabla \|x\|_b^T\Big) \leq 0$ and

$$\|x\|_b \, \lambda_{\max}^{\|\cdot\|_b}\big(\nabla^2 \|x\|_b\big) \leq (b-1)\lambda_{\max}^{\|\cdot\|_b}\Big(\operatorname{diag}\left(|x^{(n)}|^{b-2} \|x\|_b^{2-b}\right)\Big).$$

First, consider the case $b > 2$. Given $v \in \mathbb{R}^d$ such that $\|v\|_b = 1$, we have by the Hölder's inequality along with the conjugacy of $b/2$ and $b/(b-2)$,

$$\left\langle v, \operatorname{diag}\left(|x^{(n)}|^{b-2} \|x\|_b^{2-b}\right) v\right\rangle \leq \left(\sum_{n=1}^d \frac{|x^{(n)}|^b}{\|x\|_b^b}\right)^{\frac{b-2}{b}} \left(\sum_{n=1}^d |v^{(n)}|^b\right)^{\frac{2}{b}} = 1$$

Now, for the case $b = 2$ we get $\operatorname{diag}\left(|x^{(n)}|^{b-2}\|x\|_2^{2-b}\right) = I$ and $\lambda_{\max}^{\|\cdot\|_2}(I) = 1$. Our result follows. □

Lemma 4.5.4 (Convex conjugates of φ_a^A)**.** *Given $a, A \in (1,\infty)$ and φ_a^A in (4.48). Define $B = A/(A-1)$, $b = a/(a-1)$. The following hold.*

1. *Near Conjugate. φ_b^B upper bounds the conjugate $(\varphi_a^A)^*$ for all $t \in [0,\infty)$,*

$$(\varphi_a^A)^*(t) \le \varphi_b^B(t). \tag{4.57}$$

2. *Conjugate. For all $t \in [0,\infty)$,*

$$(\varphi_a^a)^*(t) = \varphi_b^b(t). \tag{4.58}$$

$$(\varphi_1^A)^*(t) = \begin{cases} 0 & t \in [0,1] \\ \frac{1}{B}t^B - t + \frac{1}{A} & t \in (1,\infty) \end{cases}. \tag{4.59}$$

$$(\varphi_a^1)^*(t) = \begin{cases} 1-(1-t^b)^{\frac{1}{b}} & t \in [0,1] \\ \infty & t \in (1,\infty) \end{cases}. \tag{4.60}$$

$$(\varphi_1^1)^*(t) = \begin{cases} 0 & t \in [0,1] \\ \infty & t \in (1,\infty) \end{cases}. \tag{4.61}$$

Proof. For convenience, define

$$\varphi(t) = \varphi_a^A(t) \qquad \phi(t) = \varphi_b^B(t). \tag{Ap.80}$$

As a reminder, for $t \in (0,\infty)$, the following identities can be easily verified.

$$\varphi'(t) = (t^a+1)^{\frac{A-a}{a}} t^{a-1} \tag{Ap.81}$$

$$\varphi''(t) = t^{-1}\varphi'(t)\left(a - 1 + (A-a)\tfrac{t^a}{t^a+1}\right) \tag{Ap.82}$$

1. *Near Conjugate.* As a reminder $\varphi^*(t) = \sup_{s\ge 0}\{ts - \varphi(s)\}$. First, since $\varphi^*(0) = -\inf_{s\ge 0}\{\varphi(s)\} = 0$, this result holds for $t = 0$. Assume $t > 0$. Our strategy will be to show that for $s \in [0,\infty)$ we have $st - \varphi(s) \le \phi(t)$. This is true for $s = 0$ by the monotonicity of ϕ in Lemma Ap.4.2, so assume $s > 0$. Now, consider $s = \phi'(\varphi'(r))$ for some $r \in (0,\infty)$. To see that this is a valid parametrization for s, notice that $\lim_{r\to 0}\phi'(\varphi'(r)) = 0$ and

$$[\phi'(\varphi'(r))]' = \phi''(\varphi'(r))\varphi''(r) > 0$$

Thus $s(r) = \phi'(\varphi'(r))$ is one-to-one and onto $(0,\infty)$. Further we have by Lemma Ap.4.3 that

$$t \le \rho(t)t = \phi'(\varphi'(t)) \tag{Ap.83}$$

and thus $(\phi')^{-1}(t) \leq \varphi'(t)$, since $\phi''(t) > 0$. All together, using convexity we have

$$\phi(t) \geq \phi(\varphi'(r)) + \phi'(\varphi'(r))(t - \varphi'(r)) = st + \phi((\phi')^{-1}(s)) - s(\phi')^{-1}(s)$$

taking the derivative of $\phi((\phi')^{-1}(s)) - s(\phi')^{-1}(s)$ we get $-(\phi')^{-1}(s)$. Since $-(\phi')^{-1}(s) \geq -\varphi'(s)$, we finally get

$$\geq st - \varphi(s).$$

Taking the supremum in s gives us our result.

2. *Conjugate.* As a reminder $\varphi^*(t) = \sup_{s\geq 0}\{ts - \varphi(s)\}$. Since $\varphi^*(0) = -\inf_{s\geq 0}\{\varphi(s)\} = 0$, these results all hold when $t = 0$. (4.58) is a standard result, since $\varphi_a^a(t) = \frac{1}{a}t^a$. (4.61) is a standard result, since $\varphi_1^1(s) = s$. Thus, we assume $a, A > 1$ and $t > 0$ for the remainder. For (4.60), assume just $A = 1$. The stationary condition of the supremum of $ts - \varphi(s)$ in s is

$$t = (s^a + 1)^{-\frac{1}{b}} s^{a-1}$$

Raising both sides to b we get $t^b = \frac{s^a}{s^a+1}$, whose solution for $t \in [0, 1]$ is $s = \left(\frac{t^b}{1-t^b}\right)^{\frac{1}{a}}$. Thus,

$$\begin{aligned}
\varphi^*(t) &= t\left(\frac{t^b}{1-t^b}\right)^{\frac{1}{a}} - \left(\frac{1}{1-t^b}\right)^{\frac{1}{a}} + 1 \\
&= \frac{t^b}{(1-t^b)^{\frac{1}{a}}} - \frac{1}{(1-t^b)^{\frac{1}{a}}} + 1 \\
&= 1 - (1-t^b)^{1-\frac{1}{a}}
\end{aligned}$$

When $t > 1$, ts dominates $\varphi(s)$ and the supremum is infinite. Now for (4.59), assume just $a = 1$. We have the stationary condition of the conjugate equal to $t = (s + 1)^{A-1}$, which corresponds to

$$s = \max\{t^{\frac{1}{A-1}} - 1, 0\}$$

Thus, when $t > 1$ we have

$$\begin{aligned}
\varphi^*(t) &= t(t^{\frac{1}{A-1}} - 1) - \tfrac{1}{A}t^B + \tfrac{1}{A} \\
&= \tfrac{1}{B}t^B - t + \tfrac{1}{A}
\end{aligned}$$

otherwise $\varphi^*(s) = 0$.

□

Proposition 4.5.8 (Verifying assumptions for f with known power behavior and appropriate k)**.** *Given a norm $\|\cdot\|_*$ satisfying F.2 and $a, A \in (1, \infty)$, take*

$$k(p) = \varphi_a^A(\|p\|_*).$$

with φ_a^A defined in (4.48)*. The following cases hold with this choice of k on $f : \mathbb{R}^d \to \mathbb{R}$ convex.*

1. *For the implicit method* (4.33)*, assumptions A, B hold with constants*

$$\alpha = \min\{\mu^{a-1}, \mu^{A-1}, 1\} \qquad C_{\alpha,\gamma} = \gamma \qquad C_{f,k} = \max\{a-1, A-1, L\}, \quad (4.64)$$

   if $f, a, A, \mu, L, \|\cdot\|_*$ satisfy assumptions F.1, F.2, F.3, F.4.

2. *For the first explicit method* (4.36)*, assumptions A, B, and C hold with constants* (4.64) *and*

$$C_k = \max\{a, A\} \qquad D_{f,k} = L_f \alpha^{-1} \max\left\{D_f,\, 2C_{a,A}(\max\{a, A\} - 1)\right\}, \quad (4.65)$$

   if $f, a, A, \mu, L, L_f, D_f, \|\cdot\|_*$ satisfy assumptions F.1, F.2, F.3, F.4, and F.5.

3. *For the second explicit method* (4.39)*, assumptions A, B, and D hold with constants* (4.64) *and*

$$\begin{aligned} C_k &= \max\{a, A\} \qquad D_k = \max\{a, A\}(\max\{a, A\} - 1) \\ E_k &= \max\{a, A\} - 1 \qquad F_k = 1 \\ D_{f,k} &= \alpha^{-1}(\max\{a, A\} - 1 + N) \max\left\{2L, a-2, A-2\right\}, \end{aligned} \quad (4.66)$$

   if $f, a, A, \mu, L, N, \|\cdot\|_*$ satisfy assumptions F.1, F.2, F.3, F.4, and F.6.

Proof. First, by Lemma 4.5.1, this choice of k satisfies assumptions A.2 and C.1 / D.2 with constant $C_k = \max\{a, A\}$. We consider the remaining assumptions of A, B, C, and D..

1. Our first goal is to derive α. By assumption F.4, we have $\mu\varphi_b^B(\|x\|) \le f_c(x)$. Since $(\mu\varphi_b^B(\|\cdot\|))^* = \mu(\varphi_b^B)^*(\mu^{-1}\|\cdot\|_*)$ by Lemma 4.5.1 and the results discussed in the review of convex analysis, we have by assumption F.3,

$$f_c^*(p) \le \mu(\varphi_b^B)^*\left(\mu^{-1}\|p\|_*\right) \le \max\{\mu^{1-a}, \mu^{1-A}\}k(p)$$

Thus, we have $\alpha = \min\{\mu^{a-1}, \mu^{A-1}, 1\}$ constant. Moreover, we can take $C_{\alpha,\gamma} = \gamma$. This along with F.1 implies that f and k satisfy assumptions A

By Lemma 4.5.1 and assumption F.4 we have

$$\varphi_a^A(\|\nabla f(x)\|_*) \leq L(f(x) - f(x_{\min})),$$
$$(\varphi_a^A)^*(\|\nabla k(p)\|) = (\max\{a, A\} - 1)k(p),$$

By Fenchel-Young and the symmetry of norms, we have $|\langle \nabla f(x), \nabla k(p)\rangle| \leq \max\{\max\{a, A\} - 1, L\}\mathcal{H}(x, p)$, from which we derive $D_{f,k}$ and the fact that f, k satisfy assumptions B.

2. The analysis of 1. holds for assumptions A and B. Now, to derive the conditions for assumptions C consider

$$\|\nabla k(p)\|^2 = [(\varphi_a^A)'(\|p\|_*)]^2$$

Note that,

$$\varphi_{b/2}^{B/2}([(\varphi_a^A)'(t)]^2) = 2\varphi_b^B((\varphi_a^A)'(t))$$

Thus $\varphi_{b/2}^{B/2}(\|\nabla k(p)\|^2) \leq 2C_{a,A}(\max\{a, A\} - 1)k(p)$ for all $p \in \mathbb{R}^d$, by Lemma 4.5.1. Now all together, by the Fenchel-Young inequality and assumption F.5, we have for $p \in \mathbb{R}^d$ and $x \in \mathbb{R}^d \setminus \{x_{\min}\}$

$$\begin{aligned}
\left\langle \nabla k(p), \nabla^2 f(x) \nabla k(p) \right\rangle &\leq \|\nabla k(p)\|^2 \lambda_{\max}^{\|\cdot\|}\left(\nabla^2 f(x)\right) \\
&\leq L_f \varphi_{b/2}^{B/2}(\|\nabla k(p)\|^2) + L_f (\varphi_{b/2}^{B/2})^* \left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f} \right) \\
&\leq L_f 2C_{a,A}(\max\{a, A\} - 1)k(p) + L_f (\varphi_{b/2}^{B/2})^* \left(\frac{\lambda_{\max}^{\|\cdot\|}(\nabla^2 f(x))}{L_f} \right) \\
&\leq D_{f,k} \alpha \mathcal{H}(x, p).
\end{aligned}$$

This gives us assumptions C.

3. The analysis of 1. holds again for assumptions A and B. Now, for assumptions D, we first note that Lemma 4.5.1 gives us constants $D_k = \max\{a, A\}(\max\{a, A\} - 1), E_k = \max\{a, A\} - 1, F_k = 1$.

For the remaining constant $D_{f,k}$ we follow a similar path as 2. First, note that since $b, B \leq 2$ we have that $a, A \geq 2$. This, along with assumption F.6, let's us use (4.56) of Lemma 4.5.1 for $p \in \mathbb{R}^d \setminus \{0\}$. Now, letting $M = (\max\{a, A\} - 1 +$

N) and applying (4.56) of Lemma 4.5.1 along with the Fenchel-Young inequality, we have for $p \in \mathbb{R}^d \setminus \{0\}$ and $x \in \mathbb{R}^d$

$$\begin{aligned}
\left\langle \nabla f(x), \nabla^2 k(p) \nabla f(x) \right\rangle &\leq \|\nabla f(x)\|_*^2 \, \lambda_{\max}^{\|\cdot\|_*}\big(\nabla^2 k(p)\big) \\
&\leq M \varphi_{a/2}^{A/2}(\|\nabla f(x)\|_*^2) + M(\varphi_{a/2}^{A/2})^* \left(\frac{\lambda_{\max}^{\|\cdot\|_*}(\nabla^2 k(p))}{M} \right) \\
&\leq 2LM(f(x) - f(x_{\min})) + M(\max\{a, A\} - 2)k(p) \\
&\leq D_{f,k} \alpha \mathcal{H}(x, p).
\end{aligned}$$

This gives us assumptions D.

□

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv e-prints*, page arXiv:1603.04467, March 2016.

[2] Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Iterative refinement for ℓ_p-norm regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1405–1424. SIAM, 2019.

[3] Deeksha Adil, Richard Peng, and Sushant Sachdeva. Fast, provably convergent irls algorithm for p-norm linear regression. In *Advances in Neural Information Processing Systems*, 2019.

[4] Deeksha Adil and Sushant Sachdeva. Faster p-norm minimizing flows, via smoothed q-norm problems. *arXiv e-prints*, Oct 2019.

[5] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, 2009.

[6] J Aitchison. A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 136–146, 1985.

[7] Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, 2016.

[8] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010.

[9] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[10] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

[11] J Atchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[12] Alfred Auslender, Roberto Cominetti, and Mounir Haddou. Asymptotic analysis for penalty and barrier methods in convex and linear programming. *Mathematics of Operations Research*, 22(1):43–62, 1997.

[13] Dominique Azé and Jean-Paul Penot. Uniformly convex and uniformly smooth convex functions. *Annales de la Faculté des Sciences de Toulouse : Mathématiques, Série 6*, 4(4):705–730, 1995.

[14] Matej Balog, Nilesh Tripuraneni, Zoubin Ghahramani, and Adrian Weller. Lost relatives of the gumbel trick. In *International Conference on Machine Learning*, 2017.

[15] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[16] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[17] Matthew J. Beal. *Variational algorithms for approximate Bayesian inference*. 2003.

[18] Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017.

[19] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[20] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE transactions on image processing*, 18(11):2419–2434, 2009.

[21] Amir Beck and Marc Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1):1–6, 2014.

[22] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[23] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, 2013.

[24] Yoshua Bengio, Li Yao, and Kyunghyun Cho. Bounding the Test Log-Likelihood of Generative Models. *arXiv e-prints*, page arXiv:1311.6184, November 2013.

[25] Jean Bérard, Pierre Del Moral, and Arnaud Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.*, 19(94):1–28, 2014.

[26] Espen Bernton. Langevin monte carlo and jko splitting. In *Conference On Learning Theory*, pages 1777–1798, 2018.

[27] Dimitri P. Bertsekas. Proximal Algorithms and Temporal Differences for Large Linear Systems: Extrapolation, Approximation, and Simulation. *arXiv e-prints*, page arXiv:1610.05427, October 2016.

[28] Dimitri P Bertsekas, Angelia Nedi, and Asuman E Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[29] Michael Betancourt, Michael I. Jordan, and Ashia C. Wilson. On Symplectic Optimization. *arXiv e-prints*, page arXiv:1802.03653, February 2018.

[30] Ashish Bhatt, Dwayne Floyd, and Brian E Moore. Second order conformal symplectic schemes for damped Hamiltonian systems. *Journal of Scientific Computing*, 66(3):1234–1259, 2016.

[31] Benjamin Birnbaum, Nikhil R. Devanur, and Lin Xiao. New convex programs and distributed algorithms for fisher markets with linear and spending constraint utilities. Technical Report MSR-TR-2010-112, August 2010.

[32] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

[33] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[34] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[35] Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. In *International Conference on Learning Representations*, 2015.

[36] Jonathan Borwein and Adrian S Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer Science & Business Media, 2010.

[37] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[38] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[39] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[40] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *International Conference on Machine Learning*, 2012.

[41] Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

[42] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

[43] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018.

[44] Lev Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[45] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[46] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[47] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[48] Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, and Yuanzhi Li. An homotopy method for lp regression provably beyond self-concordance and in input-sparsity time. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1130–1137. ACM, 2018.

[49] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Accurate and conservative estimates of MRF log-likelihood using reverse annealing. In *International Conference on Artificial Intelligence and Statistics*, 2015.

[50] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

[51] Richard Caron and Tim Traynor. The zero set of a polynomial. Technical report, 2005. Available at `http://www1.uwindsor.ca/math/sites/uwindsor.ca.`

math/files/05-03.pdf and https://www.researchgate.net/publication/281285245_The_Zero_Set_of_a_Polynomial.

[52] Christos G Cassandras, Yorai Wardi, Christos G Panayiotou, and Chen Yao. Perturbation analysis and optimization of stochastic hybrid systems. *European Journal of Control*, 16(6):642–661, 2010.

[53] Frédéric Cérou, Pierre Del Moral, and Arnaud Guyader. A nonasymptotic theorem for unnormalized Feynman–Kac particle models. *Ann. Inst. H. Poincaré B*, 47(3):629–649, 2011.

[54] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[55] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2017.

[56] Y. Chen and Z. Ghahramani. Scalable Discrete Sampling as a Multi-Armed Bandit Problem. *ArXiv e-prints*, June 2015.

[57] Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[58] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, 2015.

[59] Roberto Cominetti and Jean-Pierre Dussault. Stable exponential-penalty algorithm with superlinear convergence. *Journal of Optimization Theory and Applications*, 83(2):285–309, 1994.

[60] Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, 1994.

[61] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[62] Ivo Danihelka, Balaji Lakshminarayanan, Benigno Uria, Daan Wierstra, and Peter Dayan. Comparison of Maximum Likelihood and GAN-based training of Real NVPs. *arXiv e-prints*, page arXiv:1705.05263, May 2017.

[63] Pierre Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer Verlag, 2004.

[64] Pierre Del Moral. *Mean field simulation for Monte Carlo integration*. CRC Press, 2013.

[65] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):411–436, 2006.

[66] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[67] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

[68] Arnaud Doucet and Adam M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2011.

[69] Radu-Alexandru Dragomir, Jérôme Bolte, and Alexandre d'Aspremont. Fast Gradient Methods for Symmetric Nonnegative Matrix Factorization. *arXiv e-prints*, page arXiv:1901.10791, January 2019.

[70] Radu-Alexandru Dragomir, Alexandre d'Aspremont, and Jérôme Bolte. Quartic First-Order Methods for Low Rank Minimization. *arXiv e-prints*, page arXiv:1901.10791, January 2019.

[71] Radu-Alexandru Dragomir, Adrien Taylor, Alexand re d'Aspremont, and Jérôme Bolte. Optimal Complexity and Certification of Bregman First-Order Methods. *arXiv e-prints*, page arXiv:1911.08510, November 2019.

[72] Dmitriy Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

[73] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 2018.

[74] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

[75] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[76] Víctor Elvira, Luca Martino, David Luengo, Mónica F Bugallo, et al. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

[77] Stefano Favaro, Georgia Hadjicharalambous, and Igor Prünster. On a class of distributions on the simplex. *Journal of Statistical Planning and Inference*, 141(9):2987 – 3004, 2011.

[78] Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

[79] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368, 1922.

[80] Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

[81] Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. In *Conference on Learning Theory*, 2017.

[82] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, 2016.

[83] Guilherme Franca, Daniel P Robinson, and René Vidal. ADMM and accelerated ADMM as continuous dynamical systems. *International Conference on Machine Learning*, 2018.

[84] Guilherme Franca, Daniel P Robinson, and René Vidal. Relax, and accelerate: A continuous perspective on ADMM. In *International Conference on Machine Learning*, 2018.

[85] Brendan Frey. Continuous sigmoidal belief networks trained using slice sampling. In *Advances in Neural Information Processing Systems*, 1997.

[86] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[87] Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.

[88] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[89] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[90] Geoffrey Hinton. Neural Networks for Machine Learning. URL: `http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`, 2014. Slides 26-31 of Lecture 6.

[91] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *Control Conference (ECC), 2015 European*, pages 310–315. IEEE, 2015.

[92] Zoubin Ghahramani and Michael I Jordan. Factorial hidden Markov models. In *Advances in Neural Information Processing Systems*, 1996.

[93] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[94] Pontus Giselsson. Improved fast dual gradient methods for embedded model predictive control. *IFAC Proceedings Volumes*, 47(3):2303–2309, 2014.

[95] Pontus Giselsson and Stephen Boyd. Preconditioning in fast dual gradient methods. In *53rd IEEE Conference on Decision and Control*, pages 5040–5045. IEEE, 2014.

[96] Paul Glasserman and Yu-Chi Ho. *Gradient estimation via perturbation analysis*, volume 116. Springer Science & Business Media, 1991.

[97] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[98] Peter W Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

[99] Herbert Goldstein, Charles P. Poole, and John Safko. *Classical Mechanics*. Pearson Education, 2011.

[100] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[101] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

[102] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[103] Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

[104] Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. Learning to transduce with unbounded memory. In *Advances in Neural Information Processing Systems*, pages 1828–1836, 2015.

[105] Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. In *Advances in Neural Information Processing Systems*, 2016.

[106] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, 2015.

[107] Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv e-prints*, page arXiv:1511.02543, November 2015.

[108] Shixiang Gu, Zoubin Ghahramani, and Richard E Turner. Neural adaptive sequential Monte Carlo. In *Advances in Neural Information Processing Systems*, 2015.

[109] Shixiang Gu, Sergey Levine, Ilya Sutskever, and Andriy Mnih. MuProp: Unbiased backpropagation for stochastic neural networks. In *International Conference on Learning Representations*, 2016.

[110] Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*. Number 33. US Govt. Print. Office, 1954.

[111] Mert Gurbuzbalaban, Asuman Ozdaglar, and Pablo A Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *SIAM Journal on Optimization*, 27(2):1035–1048, 2017.

[112] David H. Gutman and Javier F. Peña. A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated gradient schemes. *arXiv e-prints*, Dec 2018.

[113] Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated Bregman Proximal Gradient Methods for Relatively Smooth Convex Optimization. *arXiv e-prints*, Aug 2018.

[114] P Hartman. *Ordinary Differential Equations*. Society for Industrial and Applied Math, 2002.

[115] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[116] Tamir Hazan and Tommi Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *International Conference on Machine Learning*, 2012.

[117] Tamir Hazan, Subhransu Maji, and Tommi Jaakkola. On Sampling from the Gibbs Distribution with Random Maximum A-Posteriori Perturbations. In *Advances in Neural Information Processing Systems*, 2013.

[118] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[119] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[120] Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[121] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[122] Martin Hutzenthaler, Arnulf Jentzen, Peter E Kloeden, et al. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4):1611–1641, 2012.

[123] Tommi S Jaakkola and Michael I Jordan. Computing upper and lower bounds on likelihoods in intractable networks. In *Proceedings of the twelfth international conference on Uncertainty in Artificial Intelligence*, pages 340–348. Morgan Kaufmann Publishers Inc., 1996.

[124] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2016.

[125] Shanyu Ji, János Kollár, and Bernard Shiffman. A global lojasiewicz inequality for algebraic varieties. *Transactions of the American Mathematical Society*, 329(2):813–818, 1992.

[126] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on Learning Theory*, 2018.

[127] Matthew J Johnson, David Duvenaud, Alexander B Wiltschko, Sandeep R Datta, and Ryan P Adams. Composing graphical models with neural networks for structured representations and fast inference. In *Neural Information Processing Systems*, 2016.

[128] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[129] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[130] Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[131] Anatoli Juditsky and Yurii Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

[132] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. 2009.

[133] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[134] Carolyn Kim, Ashish Sabharwal, and Stefano Ermon. Exact Sampling with Integer Linear Programs and Random Perturbations. In *AAAI*, 2016.

[135] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[136] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, 2016.

[137] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[138] Tomáš Kočiský, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. Semantic parsing with semi-supervised sequential autoencoders. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

[139] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, pages 2845–2853, 2015.

[140] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[141] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[142] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A flow-based generative model for video. In *International Conference on Learning Representations*, 2014.

[143] Rasmus Kyng, Anup Rao, Sushant Sachdeva, and Daniel A Spielman. Algorithms for lipschitz learning on graphs. In *Conference on Learning Theory*, pages 1190–1223, 2015.

[144] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[145] Joseph LaSalle. Some extensions of Liapunov's second method. *IRE Transactions on Circuit Theory*, 7(4):520–527, 1960.

[146] Tuan Anh Le, Maximilian Igl, Tom Jin, Tom Rainforth, and Frank Wood. Auto-encoding sequential Monte Carlo. In *International Conference on Learning Representations*, 2018.

[147] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[148] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[149] Samuel Livingstone, Michael F. Faulkner, and Gareth O. Roberts. Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. *arXiv e-prints*, page arXiv:1706.02649, June 2017.

[150] Stanislas Lojasiewicz. Sur la géométrie semi-et sous-analytique. *Ann. Inst. Fourier*, 43(5):1575–1595, 1993.

[151] Stanislaw Lojasiewicz. Sur le probleme de la division. 1961.

[152] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[153] Haihao Lu. "Relative Continuity" for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent. *INFORMS Journal on Optimization*, 2019.

[154] Haihao Lu. "relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *Informs Journal on Optimization*, 1(4):288–303, 2019.

[155] Haihao Lu and Robert M Freund. Generalized stochastic frank–wolfe algorithm with stochastic "substitute" gradient for structured convex optimization. *Mathematical Programming*, pages 1–33, 2020.

[156] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[157] Xiaoyu Lu, Valerio Perrone, Leonard Hasenclever, Yee Whye Teh, and Sebastian Vollmer. Relativistic Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[158] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[159] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[160] Chris J. Maddison. A Poisson process model for Monte Carlo. In Tamir Hazan, George Papandreou, and Daniel Tarlow, editors, *Perturbation, Optimization, and Statistics*. MIT Press, 2016.

[161] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[162] Chris J. Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian Descent Methods. *arXiv e-prints*, page arXiv:1809.05042, September 2018.

[163] Chris J. Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arXiv e-prints*, page arXiv:1809.05042, September 2018.

[164] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* Sampling. In *Advances in Neural Information Processing Systems*, 2014.

[165] Robert McLachlan and Matthew Perlmutter. Conformal Hamiltonian systems. *Journal of Geometry and Physics*, 39(4):276–300, 2001.

[166] Wenjun Mei and Francesco Bullo. LaSalle Invariance Principle for Discrete-time Dynamical Systems: A Concise and Self-contained Tutorial. *arXiv e-prints*, page arXiv:1710.03710, October 2017.

[167] Ron Meir and Gunnar Rätsch. An introduction to boosting and leveraging. In *Advanced lectures on machine learning*, pages 118–183. Springer, 2003.

[168] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

[169] Kerrie L Mengersen and Richard L Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

[170] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[171] Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv e-prints*, Sep 2019.

[172] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, 2014.

[173] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, 2014.

[174] Andriy Mnih and Danilo J Rezende. Variational inference for Monte Carlo objectives. In *International Conference on Machine Learning*, 2016.

[175] Andriy Mnih and Danilo Jimenez Rezende. Variational inference for monte carlo objectives. In *International Conference on Machine Learning*, 2016.

[176] Shakir Mohamed and Balaji Lakshminarayanan. Learning in Implicit Generative Models. *arXiv e-prints*, page arXiv:1610.03483, October 2016.

[177] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo Gradient Estimation in Machine Learning. *arXiv e-prints*, page arXiv:1906.10652, June 2019.

[178] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[179] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[180] Christian A Naesseth, Scott W Linderman, Rajesh Ranganath, and David M Blei. Variational sequential Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[181] Christian Andersson Naesseth, Fredrik Lindsten, and Thomas B Schön. Sequential Monte Carlo for graphical models. In *Advances in Neural Information Processing Systems*, 2014.

[182] Radford M Neal. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.

[183] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

[184] Radford M. Neal. Estimating Ratios of Normalizing Constants Using Linked Importance Sampling. *arXiv Mathematics e-prints*, page math/0511216, November 2005.

[185] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011.

[186] Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[187] Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

[188] Arkadi Nemirovski and Yurii Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

[189] Arkadi Nemirovski and David Yudin. Effective methods for the solution of convex programming problems of large dimensions. *Ekonom. i Mat. Metody*, 15(1):135–152, 1979.

[190] Arkadi Nemirovski and David B Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

[191] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[192] Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[193] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[194] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[195] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[196] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 2016.

[197] Brooks Paige and Frank Wood. Inference networks for sequential Monte Carlo in graphical models. In *International Conference on Machine Learning*, 2016.

[198] John William Paisley, David M. Blei, and Michael I. Jordan. Variational bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.

[199] G. Papandreou and A. Yuille. Perturb-and-MAP Random Fields: Using Discrete Optimization to Learn and Sample from Energy Models. In *International Conference on Computer Vision*, 2011.

[200] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[201] Giuseppe Peano. Démonstration de l'intégrabilité des équations différentielles ordinaires. In *Arbeiten zur Analysis und zur mathematischen Logik*, pages 76–126. Springer, 1990.

[202] Lawrence Perko. *Differential Equations and Dynamical Systems*, volume 7. Springer Science & Business Media, 2013.

[203] Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.

[204] Michael K Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics*, 171(2):134–151, 2012.

[205] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[206] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[207] Boris T Polyak. *Introduction to Optimization*. Optimization Software, 1987.

[208] BT Polyak and PS Shcherbakov. Optimisation and asymptotic stability. *International Journal of Control*, 91(11):2404–2410, 2018.

[209] Roman Polyak and Marc Teboulle. Nonlinear rescaling and proximal-like methods in convex optimization. *Mathematical programming*, 76(2):265–284, 1997.

[210] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[211] Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

[212] Tapani Raiko, Mathias Berglund, Guillaume Alain, and Laurent Dinh. Techniques for Learning Binary Stochastic Feedforward Neural Networks. *arXiv e-prints*, page arXiv:1406.2989, June 2014.

[213] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2014.

[214] Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Advances in neural information processing systems. In *NIPS*, 2016.

[215] William S Rayens and Cidambi Srinivasan. Dependence properties of generalized liouville distributions on the simplex. *Journal of the American Statistical Association*, 89(428):1465–1470, 1994.

[216] Benjamin Recht. CS726 - Lyapunov analysis and the heavy ball method. *Lecture notes*, 2012.

[217] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.

[218] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[219] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[220] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[221] Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[222] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[223] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.

[224] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[225] Francisco JR Ruiz, Michalis K Titsias, and David M Blei. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, 2016.

[226] Sotirios Sabanis. A note on tamed Euler approximations. *Electronic Communications in Probability*, 18, 2013.

[227] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

[228] Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*, volume 25, 2008.

[229] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2015.

[230] Lawrence K Saul, Tommi Jaakkola, and Michael I Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4(1):61–76, 1996.

[231] John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, pages 3528–3536, 2015.

[232] Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications. 2007.

[233] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the Acceleration Phenomenon via High-Resolution Differential Equations. *arXiv e-prints*, page arXiv:1810.08907, October 2018.

[234] Bin Shi, Simon S. Du, Weijie J. Su, and Michael I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *arXiv e-prints*, Feb 2019.

[235] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[236] John Skilling. Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

[237] Gabriel Stoltz and Zofia Trstanova. Langevin dynamics with general kinetic energies. *Multiscale Modeling & Simulation*, 16(2):777–806, 2018.

[238] Weijie Su, Stephen Boyd, and Emmanuel Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[239] Weijie Su, Stephen Boyd, and Emmanuel J Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[240] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

[241] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[242] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[243] Daniel Tarlow, Ryan Prescott Adams, and Richard S Zemel. Randomized Optimum Models for Structured Prediction. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[244] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[245] The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, Yoshua Bengio, Arnaud Bergeron, James Bergstra, Valentin Bisson, Josh Bleecher Snyder, Nicolas Bouchard, Nicolas Boulanger-Lewandowski, Xavier Bouthillier, Alexandre de Brébisson, Olivier Breuleux, Pierre-Luc Carrier, Kyunghyun Cho, Jan Chorowski, Paul Christiano, Tim Cooijmans, Marc-Alexandre Côté, Myriam Côté, Aaron Courville, Yann N. Dauphin, Olivier Delalleau, Julien Demouth, Guillaume Desjardins, Sander Dieleman, Laurent Dinh, Mélanie Ducoffe, Vincent Dumoulin, Samira Ebrahimi Kahou, Dumitru Erhan, Ziye Fan, Orhan Firat, Mathieu Germain, Xavier Glorot, Ian Goodfellow, Matt Graham, Caglar Gulcehre, Philippe Hamel, Iban Harlouchet, Jean-Philippe Heng, Balázs Hidasi, Sina Honari, Arjun Jain, Sébastien Jean, Kai Jia, Mikhail Korobov, Vivek Kulkarni, Alex Lamb, Pascal Lamblin, Eric Larsen, César Laurent, Sean Lee, Simon Lefrancois, Simon Lemieux, Nicholas Léonard, Zhouhan Lin, Jesse A. Livezey, Cory Lorenz, Jeremiah Lowin, Qianli Ma, Pierre-Antoine Manzagol, Olivier Mastropietro, Robert T. McGibbon, Roland Memisevic, Bart van Merriënboer, Vincent Michalski, Mehdi Mirza, Alberto Orlandi, Christopher Pal, Razvan Pascanu, Mohammad Pezeshki, Colin Raffel, Daniel Renshaw, Matthew Rocklin, Adriana Romero, Markus Roth, Peter Sadowski, John Salvatier, Francois Savard, Jan Schlüter, John Schulman, Gabriel Schwartz, Iulian Vlad Serban, Dmitriy Serdyuk, Samira Shabanian, Étienne Simon, Sigurd Spieckermann, S. Ramana Subramanyam, Jakub Sygnowski, Jérémie Tanguay, Gijs van Tulder, Joseph Turian, Sebastian Urban, Pascal Vincent, Francesco Visin, Harm de Vries, David Warde-Farley, Dustin J. Webb, Matthew Willson, Kelvin Xu, Lijun Xue, Li Yao, Saizheng Zhang, and Ying Zhang. Theano: A Python framework for fast computation of

mathematical expressions. *arXiv e-prints*, abs/1605.02688:arXiv:1605.02688, May 2016.

[246] Michalis Titsias and Miguel Lázaro-gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.

[247] Michalis Titsias and Miguel Lázaro-Gredilla. Local expectation gradients for black box variational inference. In *Advances in Neural Information Processing Systems*, pages 2620–2628, 2015.

[248] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

[249] Paul Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

[250] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, MIT, 2008.

[251] Paul Tseng and Dimitri P Bertsekas. On the convergence of the exponential multiplier method for convex programming. *Mathematical Programming*, 60(1-3):1–19, 1993.

[252] Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In *International Conference on Machine Learning*, 2014.

[253] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv e-prints*, page arXiv:1609.03499, September 2016.

[254] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. *arXiv e-prints*, page arXiv:1601.06759, January 2016.

[255] Henk A Van der Vorst. *Iterative Krylov methods for large linear systems*, volume 13. Cambridge University Press, 2003.

[256] Quang Van Nguyen. Forward-backward splitting with bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.

[257] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

[258] Eric Veach and Leonidas J Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 419–428. ACM, 1995.

[259] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005.

[260] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[261] Nick Whiteley and Anthony Lee. Twisted particle filters. *Ann. Statist.*, 42(1):115–141, 2014.

[262] Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference On Learning Theory*, pages 2093–3027, 2018.

[263] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[264] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[265] Ashia Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

[266] Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A Lyapunov Analysis of Momentum Methods in Optimization. *arXiv e-prints*, page arXiv:1611.02635, November 2016.

[267] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

[268] CF Jeff Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, pages 95–103, 1983.

[269] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv e-prints*, page arXiv:1609.08144, September 2016.

[270] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. In *International Conference on Learning Representations*, 2017.

[271] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[272] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(1):1–33, 2018.

[273] John I Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.

[274] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv e-prints*, page arXiv:1212.5701, December 2012.

[275] Yizhe Zhang, Xiangyu Wang, Changyou Chen, Ricardo Henao, Kai Fan, and Lawrence Carin. Towards unifying hamiltonian monte carlo and slice sampling. In *Advances in Neural Information Processing Systems*, pages 1741–1749, 2016.

[276] Xingyu Zhou. On the Fenchel duality between strong convexity and Lipschitz continuous gradient. *arXiv e-prints*, page arXiv:1803.06573, March 2018.

[277] Constantin Zălinescu. On uniformly convex functions. *Journal of Mathematical Analysis and Applications*, 95(2):344–374, 1983.

[278] Constantin Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.