Algorithms and Modelling for Large-Scale Bayesian Data Analysis

Rob Cornish

Balliol College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2019

Abstract

Bayesian statistics has emerged as a leading paradigm for the analysis of complicated datasets and for reasoning and making predictions under uncertainty. However, the framework faces significant difficulties when it is applied at scale. As datasets grow larger, simulation-based approaches to inference become unviably expensive. As models become more complex, the naïve application of traditional inference methods does not always produce valid predictions. And as Bayesian methods are applied to more complicated phenomena, it is often unclear how even to write down a suitable model on which to perform inference in the first place.

This thesis presents three pieces of work aimed at addressing these problems. We describe a Markov chain Monte Carlo method whose cost per iteration does not necessarily scale linearly with the size of the dataset when applied to Bayesian big-data posteriors. We provide a rigorous analysis of this method including precise conditions under which it yields a performance benefit over standard Metropolis–Hastings. We next provide an asymptotic analysis of nested Monte Carlo schemes that are required for certain complex Bayesian models such as probabilistic programs, along with prescriptions to ensure their consistency under well-specified conditions. Finally, we consider the task of learning models from data automatically using deep generative models. We identify a limitation of normalising flow models, which are defined to be homeomorphisms and so must preserve the topology of the prior. We propose a new family of deep generative models to address this, and demonstrate its benefits empirically across a variety of datasets.

Algorithms and Modelling for Large-Scale Bayesian Data Analysis

Rob Cornish

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2019

For Granny

Acknowledgements

Thank you to Rachael and my parents and the rest of my family. Your support has been a bedrock for me. Thank you to my supervisors Arnaud Doucet and George Deligiannidis. I deeply value your knowledge, wisdom, and guidance, and I have gained an incredible amount from working with you all. Thank you to my former supervisors Frank Wood, Hongseok Yang, and Pawan Kumar. I have learnt a great deal from being exposed to your distinct perspectives, which continue to shape my own. Thank you to my co-authors Anthony Caterini, Tom Rainforth, and Paul Vanetti. I could not ask for better people to share the highs and lows of the research process with. Thank you to the AIMS CDT and to NVIDIA for giving me the opportunity to undertake this degree at all. It is no exaggeration to say that the experience has been life-altering. Finally, thank you to Michael Robillard and Nikitas Rontsis for your excellent friendship, many laughs, and great conversations; to Anthony Caterini and Sam Davenport for bringing so much fun and happiness (and distraction) every day to the lab and elsewhere; to my friends back home for giving me a constant sense of connection to my life back in Australia; and to all the other fantastic people I have met during my time at Oxford (you know who you are). It would have been a bleak time without you all!

Abstract

Bayesian statistics has emerged as a leading paradigm for the analysis of complicated datasets and for reasoning and making predictions under uncertainty. However, the framework faces significant difficulties when it is applied at scale. As datasets grow larger, simulation-based approaches to inference become unviably expensive. As models become more complex, the naïve application of traditional inference methods does not always produce valid predictions. And as Bayesian methods are applied to more complicated phenomena, it is often unclear how even to write down a suitable model on which to perform inference in the first place.

This thesis presents three pieces of work aimed at addressing these problems. We describe a Markov chain Monte Carlo method whose cost per iteration does not necessarily scale linearly with the size of the dataset when applied to Bayesian big-data posteriors. We provide a rigorous analysis of this method including precise conditions under which it yields a performance benefit over standard Metropolis–Hastings. We next provide an asymptotic analysis of nested Monte Carlo schemes that are required for certain complex Bayesian models such as probabilistic programs, along with prescriptions to ensure their consistency under well-specified conditions. Finally, we consider the task of learning models from data automatically using deep generative models. We identify a limitation of normalising flow models, which are defined to be homeomorphisms and so must preserve the topology of the prior. We propose a new family of deep generative models to address this, and demonstrate its benefits empirically across a variety of datasets.

Contents

1

Introduction and Literature Review

This thesis is concerned with problem of extracting insights from data *at scale*. The motivation for this is straightforward. Inevitably, as computing power grows and datasets become larger, practitioners will seek to draw ever more difficult inferences using ever more complicated models. While this promises significant advances in diverse applications ranging from artificial intelligence [Ghahramani, 2015] to cosmology [Trotta, 2008], it also poses great challenges for classical statistical methodolgoies [Jordan et al., 2013, Welling et al., 2014]. In particular, a complicated tradeoff emerges between dataset size, model complexity, and quality of inference. As any of these factors changes, so too will the required computation time – often dramatically – so that the overall performance obtained on a fixed computational budget can often be difficult to reason about and optimise. Our aim here is to improve on this predicament.

1.1 Bayesian Statistics

Throughout this work, the methods we consider will either be explicitly Bayesian or will be heavily influenced by the Bayesian approach. We therefore begin by situating our goals within this context.

Bayesian statistics has emerged as a leading paradigm for the analysis of complicated datasets and for reasoning and making predictions under uncertainty [Robert, 2007, Gelman et al., 2013]. Many factors contribute to its appeal. Philosophically, Bayes' rule can be justified as the only rational way to update one's beliefs in the light of new information via so-called Dutch Book arguments [De Finetti, 1937, Terenin and Draper, 2015]. Similarly, the framework is highly congruent with the interpretation of probabilities as subjective degrees of belief. In contrast to frequentism, which interprets probabilities as long-running relative frequencies obtained from repeated experimentation, this approach is arguably more natural when reasoning about events that will occur only once (such as a political election), or for which only incomplete information is available (such as the hand of an opponent in a game of poker).

However, Bayesian statistics may also be justified on a more pragmatic basis [Jaynes, 2002, Gelman et al., 2011, 2013]. By defining a *prior* probability distribution over our parameters of interest, the Bayesian approach allows our uncertainty about those parameters to be expressed as a probability distribution also, which is arguably a more intuitive object for subsequent reasoning than a frequentist confidence interval. This prior distribution also allows a principled mechanism for injecting domain-specific knowledge into the model by assigning prior mass according to how strongly we believe each outcome will occur. At the same time, despite allowing subjectivity to enter in this way, a Bayesian model will in many cases converge naturally to an objective, "true" prediction when provided with sufficient data [Van der Vaart, 2000].

Gelman et al. [2013] provide a useful breakdown of a typical Bayesian workflow into three steps. The first step is *modelling*: the practitioner must write down a mathematical description of the phenomena they wish to investigate. In a Bayesian context, this is specified *generatively*; that is, as a joint probability distribution over all observed and unobserved parameters of interest.[1] Having obtained a model, the practitioner then must perform *inference*: given some observed data, they must

[1] This stands in contrast with the *discriminiative* modelling approach, wherein the model is specified as a conditional distribution given some observed quantity.

approximate the *posterior distribution* over unobserved target quantities. Usually this involves integrating some function with respect to the posterior distribution, and is typically the most computationally demanding step of the three. Finally, with their estimate in hand, the practitioner must assess the sensibility of their model to ensure that their predictions are reliable. Depending on their conclusions at this stage, the practitioner may need to repeat these three steps in order to obtain more accurate results.

Our focus in this work is on the first two steps: modelling and inference. We begin with the latter, which is considered a major challenge in the way of truly scaling the Bayesian framework [Welling et al., 2014]. Two major families of inference methods exist at present, which differ most fundamentally by the form of approximations to the posterior distribution that they consider. The first class consists of *Monte Carlo* methods [Robert and Casella, 2013], whose approximations are obtained as (sometimes weighted) sequences of random points that are chosen to converge asymptotically to the true posterior in an appropriate sense. In contrast, *variational* methods [Jordan et al., 1999, Blei et al., 2017] define a family (often but not necessarily parametric) of tractable distributions and then choose a member of this family that is closest to the posterior in some sense.

We make use of both Monte Carlo and variational methods in this work. Our first contribution is concerned with the former, and we begin with this now.

1.2 Subsampling MCMC

Monte Carlo methods are highly appealing for their asymptotic exactness. Under many circumstances these methods yield a direct correspondence between computational effort and quality of inference, so that by increasing our computational budget we are assured of obtaining more accurate estimates. Variational methods on the other hand usually do not have such guarantees. However, this asymptotic exactness comes at a cost: Monte Carlo has proven difficult to scale on many large problems, and variational methods are often preferred in contexts such as

machine learning where asymptotic exactness is less of a concern than computation time [Blei et al., 2017].

One straightforward reason for the relative inefficiency of many Monte Carlo methods is their need to process at each iteration the entire dataset on which the posterior is conditioned. This is especially the case for *Markov chain Monte Carlo* (MCMC) methods, which have seen widespread use within Bayesian statistics in part due to their black-box compatibility with a very broad class of possible models [Gelfand and Smith, 1990]. The celebrated Metropolis–Hastings (MH) algorithm [Metropolis et al., 1953, Hastings, 1970] provides the foundation for many such methods. MH requires only the ability to compute ratios of posterior density values to produce a Markov chain that converges asymptotically to the true posterior under weak conditions. At least in principle, posterior ratios are often straightforward to obtain since they correspond to ratios of the joint distribution over target parameters and observed data that is typically specified in the modelling stage of the Bayesian workflow. However, in practice, these ratios become increasingly expensive to compute as datasets grow larger since they typically include one likelihood term per observed datapoint, so that their cost to compute grows linearly as more data is received [Bardenet et al., 2017]. To make matters worse, this posterior ratio must be computed at *every iteration* of the algorithm, which can amount to a significant cost overall when run for many steps.

This situation contrasts markedly with the variational setting. Since variational inference is posed as an optimisation problem, it is possible to take advantage of stochastic optimisation techniques [Robbins and Monro, 1951] to use only a subset or *minibatch* of data points at each iteration [Hoffman et al., 2013, Kingma and Welling, 2013, Rezende et al., 2014]. The most straightforward such method is *stochastic gradient descent* (SGD), which simply requires an unbiased estimate of the gradient in order to converge to a local optimum under general conditions. In the big data context, we can obtain an unbiased estimate by sampling a subset of our dataset uniformly at random at each step. This subset can be as small as we like, and need not grow with the size of the dataset in order to obtain good results.

Consequently, these methods have become a key component of many state-of-the-art large-scale machine learning systems [Bottou, 2010].

It is therefore natural to consider analogous data *subsampling* procedures for MCMC. This is considerably more difficult than in the variational case. Perhaps most analogous to SGD in the MCMC context are *pseudo-marginal* methods [Andrieu et al., 2009], which replace the posterior ratios computed by MH algorithm by any ratio of unbiased (and almost-surely nonnegative) estimators of an unnormalised version of the posterior. However, to obtain such unbiased estimators is far less straightforward than for the variational setting. In general we cannot make use of unbiased estimates of the *log*-target as we did for SGD [Jacob et al., 2015], and when we can it is often difficult to ensure that the resulting Markov chain is not so statistically inefficient as to outweight the improved runtime per iteration that we gained by doing so [Bardenet et al., 2017].

As our first contribution of this thesis, we present an alternative approach to doing subsampling for MCMC. Our method is an MH-style algorithm that combines a factorised acceptance probability [Christen and Fox, 2005, Banterle et al., 2015] with fast methods for sampling non-homogeneous Bernoulli processes [Devroye, 1986] to allow iterating without computing every likelihood factor at each iteration. Unlike SGD, the cost per iteration is not directly specified by the user but emerges as a function of the factorisation of the target distribution that is chosen. However, we provide a generic factorisation based on control variate ideas [Bardenet et al., 2017, Bierkens et al., 2019] that we show yields improved performance over MH under standard posterior concentration assumptions. Moreover, our method is *exact* in the sense that it exactly preserves the posterior distribution as the invariant distribution of the Markov chain produced. This is highly pertinent since it means the method preserves the direct correspondence between computational expenditure and accuracy that as described above constitutes a major reason for the appeal of Monte Carlo methods.

This work is presented in Chapter 2 of this thesis.

1.3 Nested Monte Carlo

Recall the three steps of Bayesian workflow described above: modelling, inference, and evaluation. While our focus in the previous section was primarily on inference, we now turn to consider the intersection between inference and modelling. Here, issues of scale arise that are not simply consequences of the size of datasets used, but are also due to model complexity. In particular, as practitioners seek to address larger and more complicated problems, they inevitably require richer families of models than are typically considered in the classical Bayesian setting. In some cases, these models break certain assumptions that are required by standard inference methods such as Monte Carlo, and new theory and methodology is required to handle them.

An important example of such complex models are *probabilistic programs* [Goodman et al., 2012, Mansinghka et al., 2014, Wood et al., 2014]. These are Bayesian in the sense that they define a generative process for parameters and data jointly, with the posterior distribution still the object of interest. However, unlike classical Bayesian models, probabilistic programs are specified using arbitrary (in many cases Turing-complete) computer code, which may include such highly expressive constructs as branching, recursion, and higher-order functions. Probabilistic programs may therefore represent very rich families of generative processes and so hold the possibility to apply Bayesian methods to a range of new applications [Lake et al., 2015, 2017].

However, the expressiveness of probabilistic programs inevitably comes at the expense of harder inference. This is particularly true for systems that allow models to be *nested*, so that the result of one inference query depends on the result of additional sub-queries. Nesting improves the expressiveness of probabilistic programs by facilitating compositionality, allowing larger models to be built up out of smaller ones in a natural way. It has also been argued as an essential modelling feature for certain complex phenomena such as theory-of-mind problems that involve recursive reasoning about the intentions of other agents [Stuhlmüller and Goodman, 2014]. However, nesting introduces difficulties: in particular, standard Monte Carlo

methods for performing inference on probabilistic programs without nesting can fail to be consistent when applied naïvely to the nested case [Rainforth, 2018].

As our second contribution of this thesis, we therefore study the asymptotic behaviour of nested Monte Carlo schemes. In particular, we consider generically the problem of estimating an expection that depends on the result of a second, nested expectation. We catalogue several instances in which such a problem can be reformulated without nesting. For cases in which this is not possible, we show that nesting Monte Carlo estimators can achieve consistency under fairly weak assumptions. However, we observe that in general this approach requires a number of Monte Carlo samples that increases exponentially with the depth of nesting required in order to achieve a fixed accuracy. Our results therefore highlight certain difficulties that must be confronted if the modelling benefits of nested probabilistic programs are to be fully realised [Rainforth, 2018].

Although motivated by our desire to facilitate inference on larger and more complicated models specified as probabilistic programming, our results are formulated generically and therefore have applications beyond this. As such, we consider examples in Bayesian experimental design [Chaloner and Verdinelli, 1995], as well as deep generative modelling, which is the focus of the next and final section.

This work is presented in Chapter 3 of this thesis.

1.4 Learning Deep Generative Models

At scale, and before considering the tractability of inference, even the modelling stage of the Bayesian workflow can become extremely difficult. While frameworks such as probabilistic programming can at least *in principle* express models for highly complex phenomena, actually writing down a sufficiently useful model can in many cases seem impossible. Consider for example the problem of identifying the objects that make up a 3D scene, which is hugely important for robotics and artificial intelligence. It seems natural to approach this within the Bayesian framework, whereby we postulate some set of latent factors (that may encode, for example, the identities and positions of objects present) that gave rise to our observations

and then perform inference to determine the nature of these factors [Eslami et al., 2016]. However, to specify this model by hand seems impossible to do with enough realism to scale to complicated real-world scenarios.

The abundance of huge datasets suggests an alternative approach to this problem: rather than explicitly hardcoding a model, we may try instead to learn one from data. *Deep generative models* seek to leverage the success of deep learning precisely to this end. Many examples of deep generative models exist, with *variational autoencoders* (VAEs) [Kingma and Welling, 2013, Rezende et al., 2014], *normalising flows* [Rezende and Mohamed, 2015, Papamakarios et al., 2017], *generative adversarial networks* (GANs) [Goodfellow et al., 2014], and *autoregressive models* [Hochreiter and Schmidhuber, 1997, Graves, 2013, Oord et al., 2016a,b, Van den Oord et al., 2016] providing key examples. Broadly speaking, these express the observed data as a (potentially noisy) pushforward of some latent parameter distribution by a neural network, which is trained to minimise some statistical divergence to the data-generating distribution.[2]

Deep generative modelling can be combined naturally with probabilistic programming to produce more flexible and interpretable models. For instance, a practitioner may wish to use a deep generative model to specify an unknown *part* of their model, and hardcode the rest as a probabilistic program [Narayanaswamy et al., 2017] Frameworks such as Edward [Tran et al., 2016], Probabilistic Torch [Narayanaswamy et al., 2017], and Pyro [Bingham et al., 2019] have been explicitly designed to facilitate this approach. Most widely used for this purpose at present are VAEs and normalising flows, since both provide a straightforward mechanism for performing inference on their latent variables given observations: normalising flows by directly inverting the pushforward map, and VAEs via a variational approximation to the posterior that is learnt as part of training. Normalising flows in particular have generated significant interest in part due to their facilitating exact computation of likelihoods, which is possible because their pushforward map is chosen to be bijective.

[2]While autoregressive models do not explicitly model the data as a pushforward measure, they do so implicitly in many cases [Kingma et al., 2016, Papamakarios et al., 2017, Jaini et al., 2019].

However, as our final contribution in this thesis, we show that this exactness comes at a cost of *expressiveness*. In particular, bijectivity entails that normalising flows are homeomorphisms, and hence preserve the topology of the latent space from which they map. Intuitively, this seems problematic if the data is supported on a domain with a different topology to the latents. We show that this is indeed the case. In particular, we show that for a normalising flow to converge to such a target in distribution sense requires the bijection become arbitrarily close to noninvertible somewhere in its domain. We analyse several existing flow-based models to highlight the implications of this, and show that this result limits the expressiveness of flow-based models in practice.

As a solution, we propose *continuously indexed flow* (CIFs). CIFs consist of stacked layers of continuous mixtures of normalising flow, and constitute a hybrid between classical VAE and normalising flow models. We show that CIFs relax the bijectivity constraints on standard normalising flows and thereby improve their expressiveness. This comes at the cost that we are no longer able to perform inference exactly, and must resort to a variational approximation as with VAEs. However, we provide a natural scheme for doing so that exploits the bijectivity structure of the model, and overall demonstrate empirically that CIFs yield improved performance despite sacrificing exactness.

This work is presented in Chapter 4 of this thesis.

1.5 Summary

This thesis addresses problems of scale that occur within the inference and modelling steps of the Bayesian workflow. Our key inference contributions are: a subsampling MCMC algorithm with guarantees of performance benefits for Bayesian posterior distributions; and an asymptotic analysis of nested Monte Carlo schemes, along with prescriptions to ensure consistency under well-specified conditions. For modelling, we identify and analyse the topological limitations of normalising flow models, and propose a new family of models to address these. These contributions are made through the following three papers:

- "Scalable Metropolis-Hastings for Exact Bayesian Inference with Large Datasets", **Rob Cornish**, Paul Vanetti, Alexandre Bouchard-Côté, George Deligiannidis, Arnaud Doucet. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*, 2019

- "On Nesting Monte Carlo Estimators", Tom Rainforth, **Rob Cornish**, Hongseok Yang, Andrew Warrington, Frank Wood. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden*, 2018

- "Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows", **Rob Cornish**, Anthony Caterini, George Deligiannidis, Arnaud Doucet. In *Preprint*, 2019

This thesis is presented in integrated format with these papers included directly in conference form as the three successive chapters below.

2

Scalable Metropolis–Hastings for Exact Bayesian Inference with Large Datasets

Scalable Metropolis–Hastings for Exact Bayesian Inference with Large Datasets

Rob Cornish [1] **Paul Vanetti** [1] **Alexandre Bouchard-Côté** [2] **George Deligiannidis** [1 3] **Arnaud Doucet** [1 3]

Abstract

Bayesian inference via standard Markov Chain Monte Carlo (MCMC) methods is too computationally intensive to handle large datasets, since the cost per step usually scales like $\Theta(n)$ in the number of data points n. We propose the *Scalable Metropolis–Hastings* (SMH) kernel that exploits Gaussian concentration of the posterior to require processing on average only $O(1)$ or even $O(1/\sqrt{n})$ data points per step. This scheme is based on a combination of factorized acceptance probabilities, procedures for fast simulation of Bernoulli processes, and control variate ideas. Contrary to many MCMC subsampling schemes such as fixed step-size Stochastic Gradient Langevin Dynamics, our approach is exact insofar as the invariant distribution is the true posterior and not an approximation to it. We characterise the performance of our algorithm theoretically, and give realistic and verifiable conditions under which it is geometrically ergodic. This theory is borne out by empirical results that demonstrate overall performance benefits over standard Metropolis–Hastings and various subsampling algorithms.

1. Introduction

Bayesian inference is concerned with the posterior distribution $p(\theta|y_{1:n})$, where $\theta \in \Theta = \mathbb{R}^d$ denotes parameters of interest and $y_{1:n} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ are observed data. We assume the prior admits a Lebesgue density $p(\theta)$ and that the data are conditionally independent given θ with likelihoods $p(y_i|\theta)$, which means

$$p(\theta|y_{1:n}) \propto p(\theta) \prod_{i=1}^{n} p(y_i|\theta).$$

[1]University of Oxford, Oxford, United Kingdom [2]University of British Columbia, Vancouver, Canada [3]The Alan Turing Institute, London, United Kingdom. Correspondence to: Rob Cornish <rcornish@robots.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

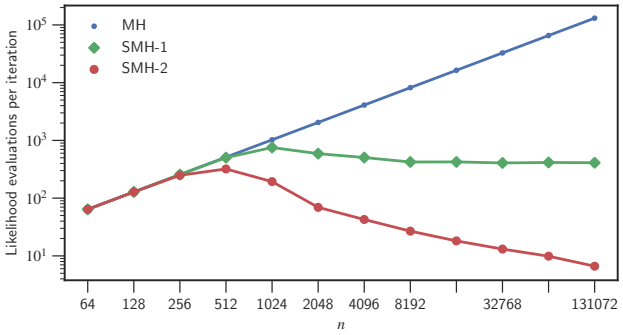


Figure 1. Average number of likelihood evaluations per iteration required by SMH for a 10-dimensional logistic regression posterior as the number of data points n increases. SMH-1 uses a first-order approximation to the target and SMH-2 a second-order one.

In most cases of interest, $p(\theta|y_{1:n})$ does not admit a closed-form expression and so we must resort to a Markov Chain Monte Carlo (MCMC) approach. However, standard MCMC schemes can become very computationally expensive for large datasets. For example, the Metropolis–Hastings (MH) algorithm requires computing a likelihood ratio $p(y_{1:n}|\theta')/p(y_{1:n}|\theta)$ at each iteration. A direct implementation of this algorithm thus requires computational cost $\Theta(n)$ per step, which is prohibitive for large n.

Many ideas for mitigating this cost have been suggested; see Bardenet et al. (2017) for a recent review. Broadly speaking these approaches are distinguished by whether they exactly preserve the true posterior as the invariant distribution of the Markov chain produced. Approximate methods that have been proposed include *divide-and-conquer* schemes, which run parallel MCMC chains on a partition of the data (Neiswanger et al., 2013; Scott et al., 2016). Other approaches replace the likelihood ratio in MH with an approximation computed from a subsample of observations. The error introduced can be controlled heuristically using central limit theorem approximations (Korattikara et al., 2014) or rigorously via concentration inequalities (Bardenet et al., 2014; 2017; Quiroz et al., 2018a). Another popular class of schemes is based on Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011; Dubey et al., 2016; Baker et al., 2018; Brosse et al., 2018; Chatterji et al., 2018),

which is a time-discretized Langevin dynamics where the gradient of the log-likelihood is approximated by subsampling. SGLD is usually implemented using a fixed step-size discretization, which does not exactly preserve the posterior distribution. Finally, Quiroz et al. (2016) and Dang et al. (2017) propose schemes that do not preserve the posterior exactly, but yield consistent estimates of posterior expectations after an importance sampling correction.

In addition to these approximate methods, several MCMC methods exist that do preserve the target as invariant distribution while only requiring access to a subset of the data at each iteration. However, various restrictions of these approaches have so far limited their widespread use. Firefly Monte Carlo (Maclaurin & Adams, 2014) considers an extended target that can be evaluated using a subset of the data at each iteration, but requires the user specify global lower bounds to the likelihood factors that can be difficult to derive. It is as yet also unclear what the convergence properties of this scheme are. Delayed acceptance schemes have been proposed based on a factorized version of the MH acceptance probability (Banterle et al., 2019) and on a random subsample of the data (Payne & Mallick, 2018). These methods allow rejecting a proposal without computing every likelihood term, but still require evaluating each term in order to accept. Quiroz et al. (2018b) combine the latter with the approximate subsampling approach of Quiroz et al. (2018a) to mitigate this problem. Finally, various non-reversible continuous-time MCMC schemes based on Piecewise Deterministic Markov Processes have been proposed which, when applied to large-scale datasets (Bouchard-Côté et al., 2018; Bierkens et al., 2019), only require evaluating the gradient of the log-likelihood for a subset of the data. However, these schemes can be difficult to understand theoretically, falling outside the scope of existing geometric ergodicity results, and can be challenging to implement.

In this paper we present a novel MH-type subsampling scheme that exactly preserves the posterior as the invariant distribution while still enjoying attractive theoretical properties and being straightforward to implement and tune. We make use of a combination of a factorized MH acceptance probability (Ceperley, 1995; Christen & Fox, 2005; Banterle et al., 2019; Michel et al., 2019; Vanetti et al., 2018) and fast methods for sampling non-homogeneous Bernoulli processes (Shanthikumar, 1985; Devroye, 1986; Fukui & Todo, 2009; Michel et al., 2019; Vanetti et al., 2018) to allow iterating without computing every likelihood factor. The combination of these ideas has proven useful for some physics models (Michel et al., 2019), but a naïve application is not efficient for large-scale Bayesian inference. Our contribution here is an MH-style MCMC kernel that realises the potential computational benefits of this method in the Bayesian setting. We refer to this kernel as *Scalable Metropolis-Hastings* (SMH) and, in addition to empirical results, provide a rigorous theoretical analysis of its behaviour under realistic and verifiable assumptions. In particular, we show SMH requires on average only $O(1)$ or even $O(1/\sqrt{n})$ cost per step as illustrated in Figure 1, has a non-vanishing average acceptance probability in the stationary regime, and is geometrically ergodic under mild conditions.

Key to our approach is the use of *control variate* ideas, which allow us to exploit the concentration around the mode frequently observed for posterior distributions with large datasets. Control variate ideas based on posterior concentration have been used successfully for large-scale Bayesian analysis in numerous recent contributions (Dubey et al., 2016; Bardenet et al., 2017; Baker et al., 2018; Brosse et al., 2018; Bierkens et al., 2019; Chatterji et al., 2018; Quiroz et al., 2018a). In our setting, this may be understood as making use of a computationally cheap approximation of the posterior.

The Supplement contains all our proofs as well as a guide to our notation in Section A.

2. Factorised Metropolis-Hastings

We first review the use of a factorised acceptance probability inside an MH-style algorithm. For now we assume a generic target $\pi(\theta)$ before specialising to the Bayesian setting below.

2.1. Transition Kernel

Assume our target $\pi(\theta)$ and proposal $q(\theta, \theta')$ factorise like

$$\pi(\theta) \propto \prod_{i=1}^{m} \pi_i(\theta) \qquad q(\theta, \theta') \propto \prod_{i=1}^{m} q_i(\theta, \theta')$$

for some $m \geq 1$ and some choice of non-negative functions π_i and q_i. These factors are not themselves required to be integrable; for instance, we may take any $\pi_i, q_i \equiv 1$. Define the *Factorised Metropolis-Hastings* (FMH) kernel

$$P_{\mathrm{FMH}}(\theta, A) := \left(1 - \int q(\theta, \theta')\alpha_{\mathrm{FMH}}(\theta, \theta')d\theta'\right)\mathbb{I}_A(\theta) + \int_A q(\theta, \theta')\alpha_{\mathrm{FMH}}(\theta, \theta')d\theta', \quad (1)$$

where $\theta \in \Theta$, $A \subseteq \Theta$ is measurable, and the FMH acceptance probability is defined

$$\alpha_{\mathrm{FMH}}(\theta, \theta') := \prod_{i=1}^{m} \underbrace{1 \wedge \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')}}_{=:\alpha_{\mathrm{FMH}\,i}(\theta, \theta')}. \quad (2)$$

It is straightforward and well-known that P_{FMH} is π-reversible; see Section B.1 in the Supplement for a proof. Factorised acceptance probabilities have appeared numerous times in the literature and date back at least to (Ceperley, 1995). The MH acceptance probability α_{MH} and kernel P_{MH} correspond to α_{FMH} and P_{FMH} when $m = 1$.

2.2. Poisson Subsampling Implementation

The acceptance step of P_{FMH} can be implemented by sampling directly m independent Bernoulli trials with success probability $1 - \alpha_{\mathrm{FMH}i}$, and returning θ' if every trial is a failure. Since we can reject θ' as soon as a single success occurs, this allows us potentially to reject θ' without computing each factor at each iteration (Christen & Fox, 2005; Banterle et al., 2019).

However, although this can lead to efficiency gains in some contexts, it remains of limited applicability for Bayesian inference with large datasets since we are still forced to compute every factor whenever we accept a proposal. It was realized independently by Michel et al. (2019) and Vanetti et al. (2018) that if one has access to lower bounds on $\alpha_{\mathrm{FMH}i}(\theta, \theta')$, hence to an upper bound on $1-\alpha_{\mathrm{FMH}i}(\theta, \theta')$, then techniques for fast simulation of Bernoulli random variables can be used that potentially avoid this problem. One such technique is given by the discrete-time thinning algorithms introduced in (Shanthikumar, 1985); see also (Devroye, 1986, Chapter VI Sections 3.3-3.4). This is used in (Michel et al., 2019).

We use here an original variation of a scheme developed in (Fukui & Todo, 2009). Denote

$$\lambda_i(\theta, \theta') := -\log \alpha_{\mathrm{FMH}i}(\theta, \theta'),$$

and assume we have the bounds

$$\lambda_i(\theta, \theta') \leq \varphi(\theta, \theta')\psi_i := \overline{\lambda}_i(\theta, \theta') \tag{3}$$

for nonnegative φ, ψ_i. This condition holds for a variety of statistical models: for instance, if π_i is log-Lipschitz and q is symmetric with (say) $q_i = q^{1/m}$, then

$$\lambda_i(\theta, \theta') \leq K_i\|\theta - \theta'\|. \tag{4}$$

This case illustrates that (3) is usually a *local* constraint on the target and therefore not as strenuous as the global lower-bounds required by Firefly (Maclaurin & Adams, 2014). We exploit this to provide a methodology for producing φ and ψ mechanically when we consider Bayesian targets in Section 3. Letting $\overline{\lambda}(\theta, \theta') := \sum_{i=1}^m \overline{\lambda}_i(\theta, \theta')$, it follows that if

$$N \sim \mathrm{Poisson}\left(\overline{\lambda}(\theta, \theta')\right)$$

$$X_1, \cdots, X_N \overset{\mathrm{iid}}{\sim} \mathrm{Categorical}((\overline{\lambda}_i(\theta, \theta')/\overline{\lambda}(\theta, \theta'))_{1\leq i\leq m})$$

$$B_j \sim \mathrm{Bernoulli}(\lambda_{X_j}(\theta, \theta')/\overline{\lambda}_{X_j}(\theta, \theta')) \text{ independently for } 1 \leq j \leq N$$

then $\mathbb{P}(B = 0) = \alpha_{\mathrm{FMH}}(\theta, \theta')$ where $B = \sum_{j=1}^N B_j$ (and $B = 0$ if $N = 0$). See Proposition C.1 in the Supplement for a proof. These steps may be interpreted as sampling a discrete Poisson point process with intensity $\lambda_i(\theta, \theta')$ on $i \in \{1, \cdots, m\}$ via thinning (Devroye, 1986). Thus, to perform the FMH acceptance step, we can simulate these B_j and check whether each is 0.

We may exploit (3) to sample each X_j and B_j in $O(1)$ time per MCMC step as $m \to \infty$ after paying some once-off setup costs. Note that

$$\overline{\lambda}(\theta, \theta') = \varphi(\theta, \theta') \sum_{i=1}^m \psi_i, \tag{5}$$

so that we may compute $\overline{\lambda}(\theta, \theta')$ in $O(1)$ time per iteration by simply evaluating $\varphi(\theta, \theta')$ if we pre-compute $\sum_{i=1}^m \psi_i$ ahead of our run. This incurs a one-time cost of $\Theta(m)$, but assuming our run is long enough this will be negligible overall. Similarly, note that

$$\frac{\overline{\lambda}_i(\theta, \theta')}{\overline{\lambda}(\theta, \theta')} = \frac{\psi_i}{\sum_{j=1}^m \psi_j},$$

so that $\mathrm{Categorical}((\overline{\lambda}_i(\theta, \theta')/\overline{\lambda}(\theta, \theta'))_{1\leq i\leq m})$ does not depend on θ, θ'. Thus, we can sample each X_i in $O(1)$ time using Walker's alias method (Walker, 1977; Kronmal & Peterson, 1979) having paid another once-off $\Theta(m)$ cost.

Algorithm 1 shows how to implement P_{FMH} using this approach. Observe that if $N < m$ we are guaranteed not to evaluate every target factor even if we accept the proposal θ'. Of course, since N is random, in general it is not obvious that $N \ll m$ will necessarily hold on average, and indeed this will not be so for a naïve factorisation. We show in Section 3 how to use Algorithm 1 as the basis of an efficient subsampling method for Bayesian inference.

In many cases we will not have bounds of the form (3) for every factor. However, Algorithm 1 can still be useful provided the computational cost for computing these extra factors is $O(1)$. In this case we can directly simulate a Bernoulli trial for each additional factor, which by assumption does not change the asymptotic complexity of this method.

2.3. Geometric Ergodicity

We consider now the theoretical implications of using P_{FMH} rather than P_{MH}. We refer the reader to Section B.2 in the Supplement for a review of the relevant definitions and theory of Markov chains. It is straightforward to show and well-known that the following holds.

Proposition 2.1. *For all* $\theta, \theta' \in \Theta$, $\alpha_{\mathrm{FMH}}(\theta, \theta') \leq \alpha_{\mathrm{MH}}(\theta, \theta')$.

See Section B in the Supplement for a proof. As such, we do not expect FMH to enjoy better convergence properties than MH. Indeed, Proposition 2.1 immediately entails that FMH produces ergodic averages of higher asymptotic variance

Algorithm 1 Efficient implementation of the FMH kernel. Setup() is called once prior to starting the MCMC run.

function Setup()
  $\Psi \leftarrow \sum_{i=1}^{m} \psi_i$
  $\tau \leftarrow \text{AliasTable}((\psi_i/\Psi)_{1 \le i \le m})$
end function

function FmhKernel(θ)
  $\theta' \sim q(\theta, \cdot)$
  $N \sim \text{Poisson}(\varphi(\theta, \theta')\Psi)$
  for $j \in 1, ..., N$ **do**
    $X_j \sim \tau$
    $B_j \sim \text{Bernoulli}(\lambda_{X_j}(\theta, \theta')/\overline{\lambda}_{X_j}(\theta, \theta'))$
    if $B_j = 1$ **then**
      return θ
    end if
  end for
  return θ'
end function

than standard MH (Peskun, 1973; Tierney, 1998). Moreover P_{FMH} can fail to be geometrically ergodic even when P_{MH} is, as noticed by Banterle et al. (2019). Geometric ergodicity is a desirable property of MCMC algorithms because it ensures the central limit theorem holds for some ergodic averages (Roberts & Rosenthal, 1997, Corollary 2.1). The central limit theorem in turn is the foundation of principled stopping criteria based on Monte Carlo standard errors (Jones & Hobert, 2001).

To address the fact that P_{FMH} might not be geometrically ergodic, we introduce the *Truncated FMH* (TFMH) kernel P_{TFMH} which is obtained by simply replacing in (1) the term $\alpha_{\text{FMH}}(\theta, \theta')$ with the acceptance probability

$$\alpha_{\text{TFMH}}(\theta, \theta') := \begin{cases} \alpha_{\text{FMH}}(\theta, \theta'), & \overline{\lambda}(\theta, \theta') < R \\ \alpha_{\text{MH}}(\theta, \theta'), & \text{otherwise,} \end{cases} \tag{6}$$

for some choice of $R \in [0, \infty]$. Observe that FMH is a special case of TFMH with $R = \infty$. When $\overline{\lambda}(\theta, \theta')$ is symmetric in θ and θ', Proposition B.3 in the Supplement shows that P_{TFMH} is still π-reversible. The following theorem shows that under mild conditions TFMH inherits the desirable convergence properties of MH.

Theorem 2.1. *If P_{MH} is φ-irreducible, aperiodic, and geometrically ergodic, then P_{TFMH} is too if*

$$\delta := \inf_{\overline{\lambda}(\theta,\theta') < R} \alpha_{\text{FMH}}(\theta, \theta') \vee \alpha_{\text{FMH}}(\theta', \theta) > 0. \tag{7}$$

In this case, $\text{Gap}(P_{\text{FMH}}) \ge \delta \text{Gap}(P_{\text{MH}})$*, and for* $f \in L^2(\pi)$

$$\text{var}(f, P_{\text{TFMH}}) \le (\delta^{-1} - 1)\text{var}(f, \pi) + \delta^{-1}\text{var}(f, P_{\text{MH}}).$$

Here $\text{Gap}(P)$ denotes the spectral gap and $\text{var}(f, P)$ the asymptotic variance of the ergodic averages of f. See Section B.2 in the Supplement for full definitions and a proof. Proposition B.1 in the Supplement shows that $\alpha_{\text{FMH}}(\theta, \theta') \vee \alpha_{\text{FMH}}(\theta', \theta) = \alpha_{\text{FMH}}(\theta, \theta')/\alpha_{\text{MH}}(\theta, \theta')$, and hence (7) quantifies the worst-case cost we pay for using the FMH acceptance probability rather than the MH one. The condition (7) is easily seen to hold in the common case that each π_i is bounded away from 0 and ∞ on $\{\theta, \theta' \in \Theta \mid \overline{\lambda}(\theta, \theta') < R\}$, which is a fairly weak requirement when $R < \infty$.

Recall from the previous section that P_{FMH} requires computing $N \sim \text{Poisson}(\overline{\lambda}(\theta, \theta'))$ factors for a given θ, θ'. In this way, TFMH yields the additional benefit of controlling the maximum expected number of factors we will need to compute via the choice of R. An obvious choice is to take $R = m$, which ensures we will not compute more factors for FMH than for MH on average. Thus, overall, TFMH yields the computational benefits of α_{FMH} when our bounds (3) are tight (usually near the mode), and otherwise falls back to MH as a default (usually in the tails).

3. FMH for Bayesian Big Data

We now consider the specific application of FMH to the problem of Bayesian inference for large datasets, where $\pi(\theta) \propto p(\theta) \prod_{i=1}^{n} p(y_i|\theta)$. It is frequently observed that such targets concentrate at a rate $1/\sqrt{n}$ around the mode as $n \to \infty$, in what is sometimes referred to as the Bernstein-von Mises phenomenon. We describe here how to leverage this phenomenon to devise an effective subsampling algorithm based on Algorithm 1. Our approach is based on control variate ideas similar to Dubey et al. (2016); Bardenet et al. (2017); Bierkens et al. (2019); Baker et al. (2018); Chatterji et al. (2018); Quiroz et al. (2018a). We emphasise that all these techniques also rely on a posterior concentration assumption but none of them only requires processing $O(1/\sqrt{n})$ data points per iteration as we do.

To see why this approach is needed, observe that the most natural factorisations of the posterior have $m \asymp n$. This introduces a major pitfall: each new factor introduced can only lower the value of $\alpha_{\text{FMH}}(\theta, \theta')$, which in the aggregate can therefore mean $\alpha_{\text{FMH}}(\theta, \theta') \to 0$ as $n \to \infty$.

Consider heuristically a naïve application of Algorithm 1 to π. Assuming a flat prior for simplicity, the obvious factorisation takes $m = n$ and each $\pi_i(\theta) = p(y_i|\theta)$. Suppose the likelihoods are log-Lipschitz and that we use the bounds (4) derived above. For smooth likelihoods, if the Lipschitz constants K_i are chosen minimally, these bounds will be tight in the limit as $\|\theta - \theta'\| \to 0$. Consequently, if we scale $\|\theta - \theta'\|$ as $1/\sqrt{n}$ to match the concentration of the target,

then $\alpha_{\mathrm{FMH}}(\theta,\theta') \asymp \exp(-\overline{\lambda}(\theta,\theta')) \to 0$ since

$$\overline{\lambda}(\theta,\theta') = \underbrace{\|\theta-\theta'\|}_{=\Theta(1/\sqrt{n})} \underbrace{\sum_{i=1}^{n} K_i}_{=\Theta(n)} = \Theta(\sqrt{n}).$$

Recall that Algorithm 1 requires the computation of at most $N \sim \mathrm{Poisson}(\overline{\lambda}(\theta,\theta'))$ factors, and hence in this case we do obtain a reduced expected cost per iteration of $\Theta(\sqrt{n})$ as opposed to $\Theta(n)$. Nevertheless, we found empirically that the increased asymptotic variance produced by the decaying acceptance probability entails an overall loss of performance compared with standard MH. We could consider using a smaller stepsize such as $\|\theta-\theta'\| = O(1/n)$ which would give a stable acceptance probability, but then our proposal would not match the $1/\sqrt{n}$ concentration of the posterior. We again found this increases the asymptotic variance to the extent that it negates the benefits of subsampling overall.

3.1. Scalable Metropolis–Hastings

Our approach is based on controlling $\overline{\lambda}(\theta,\theta')$, which ensures both a low computational cost and a large acceptance probability. We assume an initial factorisation

$$\pi(\theta) \propto p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \propto \prod_{i=1}^{m} \widetilde{\pi}_i(\theta) \tag{8}$$

for some m (not necessarily equal to n) and $\widetilde{\pi}_i$ (e.g. using directly the factorisation of prior and likelihoods). Let

$$U_i(\theta) := -\log \widetilde{\pi}_i(\theta) \qquad U(\theta) := \sum_{i=1}^{m} U_i(\theta).$$

We choose some fixed $\widehat{\theta} \in \Theta$ not depending on i that is near the mode of π like Dubey et al. (2016); Bardenet et al. (2017); Bierkens et al. (2019); Baker et al. (2018); Chatterji et al. (2018); Quiroz et al. (2018a). Assuming sufficient differentiability, we then approximate U_i with a k-th order Taylor expansion around $\widehat{\theta}$, which we denote by

$$\widehat{U}_{k,i}(\theta) \approx U_i(\theta).$$

We also define

$$\widehat{\pi}_{k,i}(\theta) := \exp(-\widehat{U}_{k,i}(\theta)) \approx \widetilde{\pi}_i(\theta).$$

In practice we are exclusively interested in the cases $k=1$ and $k=2$, which correspond to first and second-order approximations respectively. Explicitly, in these cases

$$\begin{aligned}\widehat{U}_{1,i}(\theta) &= U_i(\widehat{\theta}) + \nabla U_i(\widehat{\theta})^\top(\theta-\widehat{\theta}),\\ \widehat{U}_{2,i}(\theta) &= \widehat{U}_{1,i}(\theta) + \frac{1}{2}(\theta-\widehat{\theta})^\top \nabla^2 U_i(\widehat{\theta})(\theta-\widehat{\theta}),\end{aligned}$$

where ∇ denotes the gradient and ∇^2 the Hessian. Letting

$$\widehat{U}_k(\theta) := \sum_{i=1}^{m} \widehat{U}_{k,i}(\theta) \qquad \widehat{\pi}_k(\theta) := \exp(-\widehat{U}_k(\theta)),$$

additivity of the Taylor expansion further yields

$$\begin{aligned}\widehat{U}_1(\theta) &= U(\widehat{\theta}) + \nabla U(\widehat{\theta})^\top(\theta-\widehat{\theta})\\ \widehat{U}_2(\theta) &= \widehat{U}_1(\theta) + \frac{1}{2}(\theta-\widehat{\theta})^\top \nabla^2 U(\widehat{\theta})(\theta-\widehat{\theta}).\end{aligned} \tag{9}$$

Thus when $\nabla^2 U(\widehat{\theta}) \succ 0$ (i.e. symmetric positive-definite), $\widehat{\pi}_2(\theta)$ is seen to be a Gaussian approximation to π around the (approximate) mode $\widehat{\theta}$.

We use the $\widehat{\pi}_{k,i}$ to define the *Scalable Metropolis-Hastings* (SMH or SMH-k) acceptance probability

$$\alpha_{\text{SMH-}k}(\theta,\theta') := \left(1 \wedge \frac{\widehat{\pi}_k(\theta')q(\theta',\theta)}{\widehat{\pi}_k(\theta)q(\theta,\theta')}\right)\prod_{i=1}^{m} 1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widehat{\pi}_{k,i}(\theta')\widetilde{\pi}_i(\theta)}. \tag{10}$$

Note that SMH-k is a special case of FMH with $m+1$ factors given by

$$\pi = \underbrace{\widehat{\pi}_k}_{=\pi_{m+1}} \prod_{i=1}^{m} \underbrace{\frac{\widetilde{\pi}_i}{\widehat{\pi}_{k,i}}}_{=\pi_i} \qquad q = \underbrace{q}_{=q_{m+1}} \prod_{i=1}^{m} \underbrace{1}_{=q_i} \tag{11}$$

and hence defines a valid acceptance probability. (Note that $\widehat{\pi}_1$ is not integrable, but recall this is not required of FMH factors.) We could consider any factorisation of q, but we will not make use of this generality.

$\widehat{\pi}_k(\theta)$ can be computed in constant time after precomputing the relevant partial derivatives at $\widehat{\theta}$ before our MCMC run. This allows us to deal with $1 \wedge \widehat{\pi}_k(\theta')q(\theta',\theta)/\widehat{\pi}_k(\theta)q(\theta,\theta')$ by directly simulating a Bernoulli trial with this value as its success probability. For the remaining factors we have

$$\lambda_i(\theta,\theta') = -\log\left(1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widetilde{\pi}_i(\theta)\widehat{\pi}_{k,i}(\theta')}\right).$$

We can obtain a bound of the form (3) provided U_i is $(k+1)$-times continuously differentiable. In this case, if we can find constants

$$\overline{U}_{k+1,i} \geq \sup_{\substack{\theta\in\Theta\\ |\beta|=k+1}} |\partial^\beta U_i(\theta)|, \tag{12}$$

(here β is multi-index notation; see Section A of the Supplement) it follows that

$$\overline{\lambda}(\theta,\theta') := \underbrace{(\|\theta-\widehat{\theta}\|_1^{k+1} + \|\theta'-\widehat{\theta}\|_1^{k+1})}_{=\varphi(\theta,\theta')} \sum_{i=1}^{m} \underbrace{\frac{\overline{U}_{k+1,i}}{(k+1)!}}_{=\psi_i} \tag{13}$$

defines an upper bound of the required form (5). See Proposition D.1 in the Supplement for a derivation. Observe this is symmetric in θ and θ' and therefore can be used to define a truncated version of SMH as described in Section 2.3.

Although we concentrate on Taylor expansions here, other choices of $\widehat{\pi}_i$ may be useful. For instance, it may be possible to make $\widetilde{\pi}_i/\widehat{\pi}_i$ log-Lipschitz or log-concave and obtain better bounds. However, Taylor expansions have the advantage of generality and (13) is sufficiently tight for us.

Heuristically, if the posterior concentrates like $1/\sqrt{n}$, if we scale our proposal like $1/\sqrt{n}$, and if $\widehat{\theta}$ is not too far (specifically $O(1/\sqrt{n})$) from the mode, then both $\|\theta - \widehat{\theta}\|$ and $\|\theta' - \widehat{\theta}\|$ will be $O(1/\sqrt{n})$, and $\varphi(\theta, \theta')$ will be $O(n^{-(k+1)/2})$. If moreover $m \asymp n$, then the summation will be $O(n)$ and hence overall $\overline{\lambda}(\theta, \theta') = O(n^{(1-k)/2})$. When $k = 1$ this is $O(1)$ and when $k = 2$ this is $O(1/\sqrt{n})$, which entails a substantial improvement over the naïve approach. In particular, we expect stable acceptance probabilities in both cases, constant expected cost in n for $k = 1$, and indeed $O(1/\sqrt{n})$ *decreasing* cost for $k = 2$. We make this argument rigorous in Theorem 3.1 below.

Beyond what is already needed for MH, $\overline{U}_{k+1,i}$ and $\widehat{\theta}$ are all the user must provide for our method. In practice neither of these seems problematic in typical settings. We have found deriving $\overline{U}_{k+1,i}$ to be a fairly mechanical procedure, and give examples for two models in Section 4. Likewise, while computing $\widehat{\theta}$ does entail some cost, we have found that standard gradient descent finds an adequate result in time negligible compared with the full MCMC run.

3.2. Choice of Proposal

We now consider the choice of proposal q and its implications for the acceptance probability. As mentioned, it is necessary to ensure that, roughly speaking, $\|\theta - \theta'\| = O(n^{-1/2})$ to match the concentration of the target. In this section we describe heuristically how to ensure this. Theorem 3.1 below and Section F.1.2 in the Supplement give precise statements of what is required.

Two main classes of q are of interest to us. When q is symmetric, (10) simplifies to

$$\alpha_{\text{SMH-}k}(\theta, \theta') = \left(1 \wedge \frac{\widehat{\pi}_k(\theta')}{\widehat{\pi}_k(\theta)}\right) \prod_{i=1}^{m} 1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widetilde{\pi}_i(\theta)\widehat{\pi}_{k,i}(\theta')}. \tag{14}$$

We can realise this with the correct scaling with for example

$$q(\theta, \theta') = \text{Normal}(\theta' \mid \theta, \frac{\sigma^2}{n} I_d), \tag{15}$$

where $\sigma > 0$ is fixed in n. Alternatively, we can more closely match the covariance of our proposal to the covariance of our target with

$$q(\theta, \theta') = \text{Normal}(\theta' \mid \theta, \sigma^2[\nabla^2 U(\widehat{\theta})]^{-1}). \tag{16}$$

Under usual circumstances $[\nabla^2 U(\widehat{\theta})]^{-1}$ is approximately (since in general this will include a non-flat prior term) proportional to the inverse observed information matrix, and hence the correct $O(n^{-1/2})$ scaling is achieved automatically. See Section F.1.2 in the Supplement for more details.

We can improve somewhat on a symmetric proposal if we choose q to be $\widehat{\pi}_k$-*reversible* in the sense that

$$\widehat{\pi}_k(\theta) q(\theta, \theta') = \widehat{\pi}_k(\theta') q(\theta', \theta)$$

for all θ, θ'; see, e.g., (Tierney, 1994; Neal, 1999; Kamatani, 2018). In this case we obtain

$$\alpha_{\text{SMH-}k}(\theta, \theta') = \prod_{i=1}^{m} 1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widetilde{\pi}_i(\theta)\widehat{\pi}_{k,i}(\theta')}.$$

Note that using a $\widehat{\pi}_k$-reversible proposal allows us to drop the first term in (14), and hence obtain a higher acceptance probability for the same θ, θ'. Moreover, when $k = 2$, we see from (9) that a $\widehat{\pi}_k$-reversible proposal corresponds to an MCMC kernel that targets a Gaussian approximation to π, and may therefore be more suited to the geometry of π than a symmetric one.

We now consider how to produce $\widehat{\pi}_k$-reversible proposals. For q of the form

$$q(\theta, \theta') = \text{Normal}(\theta' \mid A\theta + b, C)$$

where $A, C \in \mathbb{R}^{d \times d}$ with $C \succ 0$ and $b \in \mathbb{R}^d$, Theorem E.1 in the Supplement gives necessary and sufficient conditions for $\widehat{\pi}_1$ and $\widehat{\pi}_2$-reversibility. Specific useful choices that satisfy these conditions and ensure the correct scaling are then as follows. For $\widehat{\pi}_1$ we can use for example

$$A = I_d \qquad b = -\frac{\sigma}{2n} \nabla U(\widehat{\theta}) \qquad C = \frac{\sigma}{n} I_d \tag{17}$$

for some $\sigma > 0$, where $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix. For $\widehat{\pi}_2$, assuming $\nabla^2 U(\widehat{\theta}) \succ 0$ (which will hold if $\widehat{\theta}$ is sufficiently close to the mode), we can use a variation of the *preconditioned-Crank Nicholson* proposal (pCN) (Neal, 1999) defined by taking

$$A = \sqrt{\rho} I_d \qquad C = (1 - \rho)[\nabla^2 U(\widehat{\theta})]^{-1}$$
$$b = (1 - \sqrt{\rho})(\widehat{\theta} - [\nabla^2 U(\widehat{\theta})]^{-1} \nabla U(\widehat{\theta}))$$

where $\rho \in [0, 1)$. When $\rho = 0$ this corresponds to an independent Gaussian proposal: $\theta' \sim \widehat{\pi}_2$. Note that this can be re-interpreted as the exact discretization of an Hamiltonian dynamics for the Gaussian $\widehat{\pi}_2$.

3.3. Performance

We now show rigorously that SMH addresses the issues of a naive approach and entails an overall performance benefit. In our setup we assume some unknown data-generating distribution P_0, with data $Y_i \overset{\text{iid}}{\sim} P_0$. We denote the (random) targets by $\pi^{(n)}(\theta) := p(\theta|Y_{1:n})$, for which we assume a factorisation (8) involving $m^{(n)}$ terms. We denote the mode of $\pi^{(n)}$ by $\theta^{(n)}_{\text{MAP}}$, and our estimate of the mode by $\widehat{\theta}^{(n)}$. Observe that $\theta^{(n)}_{\text{MAP}} \equiv \theta^{(n)}_{\text{MAP}}(Y_{1:n})$ is a deterministic function of the data, and we assume this holds for $\widehat{\theta}^{(n)} \equiv \widehat{\theta}^{(n)}(Y_{1:n})$ also. In general $\widehat{\theta}^{(n)}$ may depend on additional randomness, say $W_{1:n}$, if for instance it is the output of a stochastic gradient descent algorithm. In that case, our statements should involve conditioning on $W_{1:n}$ but are otherwise unchanged.

Given n data points we denote the proposal by $q^{(n)}$, and model the behaviour of our chain at stationarity by considering $\theta^{(n)} \sim \pi^{(n)}$ and $\theta'^{(n)} \sim q^{(n)}(\theta^{(n)}, \cdot)$ sampled independently of all other randomness given $Y_{1:n}$. The following theorem allows us to show that both the computational cost and the acceptance probability of SMH remain stable as $n \to \infty$. See Section F in the Supplement for a proof.

Theorem 3.1. *Suppose each U_i is $(k+1)$-times continuously differentiable, each $\overline{U}_{k+1,i} \in L^{k+2}$, and $\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{k+1,i}|Y_{1:n}] = O_{P_0}(n)$. Likewise, assume each of $\|\theta^{(n)} - \theta^{(n)}_{\text{MAP}}\|$, $\|\theta^{(n)} - \theta'^{(n)}\|$, and $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\text{MAP}}\|$ is in L^{k+2}, and each of $\mathbb{E}[\|\theta^{(n)} - \theta^{(n)}_{\text{MAP}}\|^{k+1}|Y_{1:n}]$, $\mathbb{E}[\|\theta^{(n)} - \theta'^{(n)}\|^{k+1}|Y_{1:n}]$, and $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\text{MAP}}\|^{k+1}$ is $O_{P_0}(n^{-(k+1)/2})$ as $n \to \infty$. Then $\overline{\lambda}$ defined by (13) satisfies*

$$\mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = O_{P_0}(n^{(1-k)/2}).$$

For given $\theta^{(n)}$ and $\theta'^{(n)}$, recall that the method described in Section 2.2 requires the computation of at most $N^{(n)} \sim \text{Poisson}(\overline{\lambda}(\theta^{(n)}, \theta'^{(n)}))$ factors. Under the conditions of Theorem 3.1, we therefore have

$$\begin{aligned}\mathbb{E}[N^{(n)}|Y_{1:n}] &= \mathbb{E}[\mathbb{E}[N^{(n)}|\theta^{(n)}, \theta'^{(n)}, Y_{1:n}]|Y_{1:n}] \\ &= \mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] \\ &= O_{P_0}(n^{(1-k)/2}).\end{aligned}$$

In other words, with arbitrarily high probability with respect to the data-generating distribution, SMH requires processing on average only $O(1)$ data points per step for a first-order approximation, and $O(1/\sqrt{n})$ for a second-order one.

This result also ensures that the acceptance probability for SMH does not vanish as $n \to \infty$. Denoting by $\widehat{\pi}^{(n)}_k$ our approximation in the case of n data points, observe that

$$\begin{aligned}0 &\leq \mathbb{E}[-\log \alpha_{\text{FMH}}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] \\ &\leq \mathbb{E}[-\log(1 \wedge \frac{\widehat{\pi}^{(n)}_k(\theta'^{(n)})q^{(n)}(\theta'^{(n)}, \theta^{(n)})}{\widehat{\pi}^{(n)}_k(\theta^{(n)})q^{(n)}(\theta^{(n)}, \theta'^{(n)})})|Y_{1:n}] \\ &\qquad + \mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}]\end{aligned}$$

Here the second right-hand side term is $O_{P_0}(n^{(1-k)/2})$ by Theorem 3.1. For a $\widehat{\pi}_k$-reversible proposal the first term is simply 0, while for a symmetric proposal Theorem F.2 in the Supplement shows it is $O_{P_0}(1)$. In either case, we see that the acceptance probability is stable in the limit of large n. In the case of a $\widehat{\pi}_2$-reversible proposal, we in fact have $\mathbb{E}[\alpha_{\text{FMH}}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] \xrightarrow{P_0} 1$.

Note that both these implications also apply if we use a truncated version of SMH as per Section 2.3. This holds since in general TFMH ensures both that the expected number of factor evaluations is not greater than for FMH, and that the acceptance probability is not less than for FMH.

The conditions of Theorem 3.1 hold in realistic scenarios. The integrability assumptions are mild and mainly technical. We will see in Section 4 that in practice $\overline{U}_{k+1,i} \equiv \overline{U}_{k+1}(Y_i)$ is usually a function of Y_i, in which case

$$\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{k,i}|Y_{1:n}] = \sum_{i=1}^{n} \overline{U}_{k+1}(Y_i) = O_{P_0}(n)$$

by the law of large numbers. In general, we might also have one $\overline{U}_{k,i}$ for the prior also, but the addition of this term still gives the same asymptotic behaviour.

The condition $\mathbb{E}[\|\theta^{(n)} - \theta^{(n)}_{\text{MAP}}\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2})$ essentially states that the posterior must concentrate at rate $O(1/\sqrt{n})$ around the mode. This is a consequence of standard, widely-applicable assumptions that are used to prove Bernstein-von Mises. See Section F.1.1 of the Supplement for more details. Note in particular that we do not require our model to be well-specified (i.e. we do not need $P_0 = p(y|\theta_0)$ for some $\theta_0 \in \Theta$). The remaining two O_{P_0} conditions correspond to the heuristic conditions given in Section 3.1. In particular, the proposal should scale like $1/\sqrt{n}$. We show in Section F.1.2 of the Supplement that this condition holds for the proposals described in Section 3.2. Likewise, $\widehat{\theta}$ should be distance $O(1/\sqrt{n})$ from the mode. When the posterior is log-concave it can be shown this holds for instance for stochastic gradient descent after performing a single pass through the data (Baker et al., 2018, Section 3.4). In practice, we interpret this condition to mean that $\widehat{\theta}$ should be as close as possible to θ_{MAP}, but that some small margin for error is acceptable.

4. Experimental Results

In this section we apply SMH to Bayesian logistic regression. A full description of the model and upper bounds (12) we used is given in Section G.2 of the Supplement. We also provide there an additional application our method to robust linear regression. We chose these models due to the availability of lower bounds on the likelihoods required by Firefly.

In our experiments we took $d = 10$. For both SMH-1 and SMH-2 we used truncation as described in Section 2.3, with $R = n$. Our estimate of the mode $\widehat{\theta}$ was computed using stochastic gradient descent. We compare our algorithms to standard MH, Firefly, and Zig-Zag (Bierkens et al., 2019), which all have the exact posterior as the invariant distribution. We used the MAP-tuned variant of Firefly (which also makes use of $\widehat{\theta}$) with implicit sampling (this uses an algorithmic parameter $q_{d\to b} = 10^{-3}$; the optimal choice of $q_{d\to b}$ is an open question) and the lower bounds specified in Section 3.1 of Maclaurin & Adams (2014).

Figure 1 (in Section 1) shows the average number of likelihood evaluations per step and confirms the predictions of Theorem 3.1. Figure 2 displays the effective sample sizes (ESS) for the posterior mean estimate of one regression coefficient, rescaled by execution time. For large n, SMH-2 significantly outperforms competing techniques. For all methods except Zig-Zag we used the proposal (16) with $\sigma = 1$, which automatically scales according to the concentration of the target.

We also separately considered the performance of the pCN proposal. Figure 3 shows the effect of varying ρ. As the target concentrates, the Gaussian approximation of the target improves and an independent proposal ($\rho = 0$) becomes optimal. Finally, we also illustrate the average acceptance rate when varying ρ in Figure 4.

Since SMH-2 makes use of a Gaussian approximation to the posterior $\widehat{\pi}_2$, we finally consider the benefit that our method yields over simply using $\widehat{\pi}_2$ directly. Observe in Figure 4 that the acceptance probability of MH with the independent proposal differs non-negligibly from 1 for reasonably large values of n, which indicates that our method yields a non-trivial increase in accuracy. For very large n, the discrepancy vanishes as expected and SMH and other subsampling methods based on control variates become less useful in practice. See Section G of the Supplement for further results along these lines. We believe however that our approach could form the basis of subsampling methods in more general and interesting settings such as random effect models and leave this as an important piece of future work.

Code to reproduce our experiments is available at `github.com/pjcv/smh`.

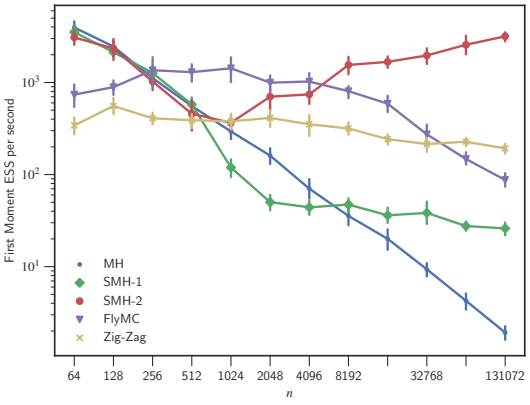


Figure 2. ESS for first regression coefficient, scaled by execution time (higher is better).

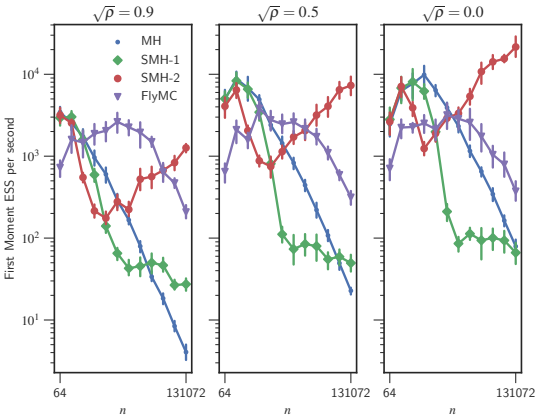


Figure 3. Effect of ρ on ESS. ESS for first regression coefficient, scaled by execution time (higher is better).

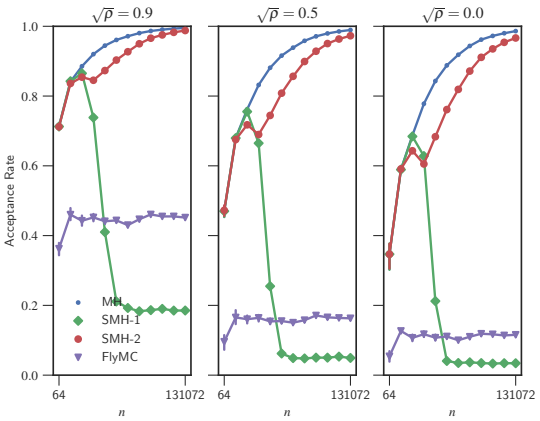


Figure 4. Acceptance rates for pCN proposals.

Acknowledgements

Rob Cornish is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems (EP/L015897/1) and NVIDIA. Arnaud Doucet is partially supported by the U.S. Army Research Laboratory, the U. S. Army Research Office, and by the U.K. Ministry of Defence (MoD) and the U.K. EPSRC under grant numbers EP/R013616/1 and EP/R034710/1.

References

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 2018. URL https://doi.org/10.1007/s11222-018-9826-2. to appear.

Banterle, M., Grazian, C., Lee, A., and Robert, C. P. Accelerating Metropolis-Hastings algorithms by Delayed Acceptance. *Foundations of Data Science*, 1(2):103–128, 2019.

Bardenet, R., Doucet, A., and Holmes, C. Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–405–I–413. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3044852.

Bardenet, R., Doucet, A., and Holmes, C. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017. URL http://jmlr.org/papers/v18/15-205.html.

Bierkens, J., Fearnhead, P., and Roberts, G. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics*, 47:1288–1320, 2019.

Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. The Bouncy Particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113:855–867, 2018.

Brosse, N., Durmus, A., and Moulines, E. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 8268–8278, 2018.

Ceperley, D. M. Path integrals in the theory of condensed helium. *Reviews of Modern Physics*, 67(2):279, 1995.

Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L., and Jordan, M. I. On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*, 2018.

Christen, J. A. and Fox, C. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005. ISSN 10618600. URL http://www.jstor.org/stable/27594150.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. Hamiltonian Monte Carlo with energy conserving subsampling. *arXiv preprint arXiv:1708.00955*, 2017.

Devroye, L. *Non-Uniform Random Variate Generation*. Springer, 1986.

Dubey, K. A., Reddi, S. J., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 1154–1162, 2016.

Fukui, K. and Todo, S. Order-n cluster Monte Carlo method for spin systems with long-range interactions. *Journal of Computational Physics*, 228(7):2629–2642, 2009.

Jones, G. L. and Hobert, J. P. Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo. *Statistical Science*, 16(4):312–334, 2001. ISSN 0883-4237.

Kamatani, K. Efficient strategy for the Markov chain Monte Carlo in high-dimension with heavy-tailed target probability distribution. *Bernoulli*, 24(4B):3711–3750, 2018.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *International Conference on Machine Learning*, pp. 181–189, 2014.

Kronmal, R. A. and Peterson, A. V. J. On the alias method for generating random variables from a discrete distribution. *The American Statistician*, 33(4):214–218, 1979.

Maclaurin, D. and Adams, R. P. Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pp. 543–552, Arlington, Virginia, United States, 2014. AUAI Press. ISBN 978-0-9749039-1-0. URL http://dl.acm.org/citation.cfm?id=3020751.3020808.

Michel, M., Tan, X., and Deng, Y. Clock Monte Carlo methods. *Physical Review E*, 99(1):010105, 2019.

Neal, R. M. Regression and classification using Gaussian process priors. In *Bayesian Statistics 6*, pp. 475–501. Oxford Univ. Press, 1999.

Neiswanger, W., Wang, C., and Xing, E. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.

Payne, R. D. and Mallick, B. K. Two-stage Metropolis–Hastings for tall data. *Journal of Classification*, 35 (1):29–51, 2018. ISSN 1432-1343. doi: 10.1007/s00357-018-9248-z. URL https://doi.org/10.1007/s00357-018-9248-z.

Peskun, P. H. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973. ISSN 00063444. URL http://www.jstor.org/stable/2335011.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. The block-Poisson estimator for optimally tuned exact subsampling MCMC. *arXiv e-prints*, art. arXiv:1603.08232, 2016.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 2018a. to appear.

Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. Speeding up MCMC by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics*, 27(1):12–22, 2018b. doi: 10.1080/10618600.2017.1307117. URL https://doi.org/10.1080/10618600.2017.1307117.

Roberts, G. and Rosenthal, J. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997. doi: 10.1214/ECP.v2-981. URL https://doi.org/10.1214/ECP.v2-981.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016. URL http://www.tandfonline.com/doi/full/10.1080/17509653.2016.1142191.

Shanthikumar, J. Discrete random variate generation using uniformization. *European Journal of Operational Research*, 21(3):387–398, 1985.

Tierney, L. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

Tierney, L. A note on Metropolis–Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998.

Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Piecewise deterministic Markov chain Monte Carlo. *arXiv preprint arXiv:1707.05296v2*, 2018.

Walker, A. J. An efficient method for generating discrete random variables with general distributions. *ACM Trans. Math. Softw.*, 3(3):253–256, 1977. ISSN 0098-3500. doi: 10.1145/355744.355749. URL http://doi.acm.org/10.1145/355744.355749.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104568.

Scalable Metropolis–Hastings for Exact Bayesian Inference with Large Datasets: Supplementary Material

Rob Cornish [1] **Paul Vanetti** [1] **Alexandre Bouchard-Côté** [2] **George Deligiannidis** [1 3] **Arnaud Doucet** [1 3]

A. Guide to Notation

$a \wedge b$	$\min\{a, b\}$
$a \vee b$	$\max\{a, b\}$
$B(x, K)$	Euclidean ball centered at x of radius K
$\mathbb{I}_A$	Indicator function of the set A
$\partial_j F(x)$	j-th partial derivative of F at x, i.e. $\partial F(x)/\partial x_j$
$\nabla F(x)$	Gradient of F at x
$\nabla^2 F(x)$	Hessian of F at x
$\|\cdot\|$	The ℓ^2 norm
$\|\cdot\|_1$	The ℓ^1 norm
$\|\cdot\|_\infty$	The supremum norm
$\|\cdot\|_{\text{op}}$	The operator norm with respect to $\|\cdot\|$ on the domain and range
I_d	Identity matrix
$A \prec B$	$B - A$ is symmetric positive-definite
$A \preceq B$	$B - A$ is symmetric nonnegative-definite
$a(x) \asymp b(x)$ as $x \to x_0$	$\lim_{x \to x_0} a(x)/b(x) = 1$
$a(x) = O(b(x))$ as $x \to x_0$	$\limsup_{x \to x_0} \lvert a(x)/b(x)\rvert < \infty$
$a(x) = \Theta(b(x))$ as $x \to x_0$	$a(x) = O(b(x))$ as $x \to x_0$ and $\liminf_{x \to x_0} \lvert a(x)/b(x)\rvert > 0$. (Note that similar notation is used for our state space Θ, but the meaning will always be clear from context.)
$x \ll y$	(Informal) x is much smaller than y
$x \approx y$	(Informal) x is approximately equal to y
Leb	The Lebesgue measure
a.s.	Almost surely
i.i.d.	Independent and identically distributed
$X_n \xrightarrow{\mathbb{P}} X$	X_n converges to X in $\mathbb{P}$-probability
$X_n = O_{\mathbb{P}}(a_n)$	X_n/a_n is $\mathbb{P}$-tight, i.e. for all $\epsilon > 0$ there exists $c > 0$ such that $\mathbb{P}(\lvert X_n/a_n\rvert < c) > 1 - \epsilon$ for all n
$X_n = o_{\mathbb{P}}(a_n)$	$X_n/a_n \xrightarrow{\mathbb{P}} 0$
$\mathbb{E}[X]$	Expectation of a random variable X
$\mathbb{E}[X; A]$	$\mathbb{E}[X\mathbb{I}_A]$
L^p	The space of random variables X such that $\mathbb{E}[\lvert X\rvert^p] < \infty$
$L^p(\mu)$	The space of real-valued test functions f such that $f(X) \in L^p$ where $X \sim \mu$

We also use multi-index notation to express higher-order derivatives succinctly. Specifically, for $\beta = (\beta_1, \cdots, \beta_d) \in \mathbb{Z}^d_{\geq 0}$

[1]University of Oxford, Oxford, United Kingdom [2]University of British Columbia, Vancouver, Canada [3]The Alan Turing Institute, London, United Kingdom. Correspondence to: Rob Cornish <rcornish@robots.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

and $\theta = (\theta_1, \cdots, \theta_d) \in \Theta$, we define

$$|\beta| := \sum_{i=1}^{d} \beta_i \qquad \beta! := \prod_{i=1}^{d} \beta_i! \qquad \theta^\beta := \prod_{i=1}^{d} \theta_i^{\beta_i} \qquad \partial^\beta := \frac{\partial^{|\beta|}}{\partial^{\beta_1} \dots \partial^{\beta_d}}.$$

B. Factorised Metropolis–Hastings

Note that the definition (2) of $\alpha_{\mathrm{FMH}}(\theta, \theta')$ technically does not apply when $\pi(\theta)q(\theta, \theta') = 0$. For concreteness, like Hastings (1970), we therefore define explicitly

$$\alpha_{\mathrm{FMH}}(\theta, \theta') := \begin{cases} \prod_{i=1}^{m} 1 \wedge \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')} & \text{if each } \pi_i(\theta)q_i(\theta, \theta') \neq 0 \\ 1 & \text{otherwise,} \end{cases}$$

and take $\alpha_{\mathrm{MH}}(\theta, \theta')$ to be the case when $m = 1$. We still take $\alpha_{\mathrm{TFMH}}(\theta, \theta')$ to be defined by (6). We first establish a useful preliminary Proposition.

Proposition B.1. *For all $\theta, \theta' \in \Theta$, $\alpha_{\mathrm{FMH}}(\theta, \theta') = \alpha_{\mathrm{MH}}(\theta, \theta')(\alpha_{\mathrm{FMH}}(\theta, \theta') \vee \alpha_{\mathrm{FMH}}(\theta', \theta))$.*

Proof. The cases where $\pi_i(\theta)q_i(\theta, \theta') = 0$ or $\pi_i(\theta')q_i(\theta', \theta) = 0$ for some i are immediate from the definition above. Otherwise, since $(1 \wedge c)^{-1} = 1 \vee c^{-1}$ for all $c > 0$,

$$\begin{aligned} \alpha_{\mathrm{MH}}(\theta, \theta')^{-1} &= \left(1 \wedge \prod_{i=1}^{m} \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')}\right)^{-1} \\ &= 1 \vee \prod_{i=1}^{m} \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}, \end{aligned}$$

and hence

$$\begin{aligned} \frac{\alpha_{\mathrm{FMH}}(\theta, \theta')}{\alpha_{\mathrm{MH}}(\theta, \theta')} &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \left(\alpha_{\mathrm{FMH}}(\theta, \theta') \prod_{i=1}^{m} \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}\right) \\ &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \left(\prod_{i=1}^{m} (1 \wedge \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')}) \prod_{i=1}^{m} \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}\right) \\ &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \left(\prod_{i=1}^{m} 1 \wedge \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}\right) \\ &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \alpha_{\mathrm{FMH}}(\theta', \theta) \end{aligned}$$

which gives the result. □

Corollary B.1. *For all $\theta, \theta' \in \Theta$, $\alpha_{\mathrm{FMH}}(\theta, \theta') \leq \alpha_{\mathrm{MH}}(\theta, \theta')$.*

B.1. Reversibility

To show reversibility for P_{FMH} and P_{TFMH}, we will use the standard result (see e.g. (Geyer, 1998, Lemma 3.4)) that a kernel of the form

$$P(\theta, A) = \left(1 - \int q(\theta, \theta')\alpha(\theta, \theta')d\theta'\right) \mathbb{I}_A(\theta) + \int_A q(\theta, \theta')\alpha(\theta, \theta')d\theta'$$

is reversible if $\pi(\theta)q(\theta, \theta')\alpha(\theta, \theta')$ is symmetric in θ and θ'. It is straightforward to show for instance that

$$\pi(\theta)q(\theta, \theta')\alpha_{\mathrm{MH}}(\theta, \theta') = \pi(\theta')q(\theta', \theta)\alpha_{\mathrm{MH}}(\theta', \theta), \tag{B.1}$$

which is immediate if either $\pi(\theta) = 0$ or $\pi(\theta') = 0$, and otherwise

$$\begin{aligned} \pi(\theta)q(\theta, \theta')\alpha_{\mathrm{MH}}(\theta, \theta') &= \pi(\theta)q(\theta, \theta') \left(1 \wedge \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right) \\ &= \pi(\theta)q(\theta, \theta') \wedge \pi(\theta')q(\theta', \theta). \end{aligned}$$

We use this result to establish reversiblity of P_{FMH}. This result is standard but we include it here for completeness.

Proposition B.2. *P_{FMH} is π-reversible.*

Proof. By Proposition B.1

$$\pi(\theta)q(\theta,\theta')\alpha_{\mathrm{FMH}}(\theta,\theta') = \pi(\theta)q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')(\alpha_{\mathrm{FMH}}(\theta,\theta') \vee \alpha_{\mathrm{FMH}}(\theta',\theta)),$$

which is symmetric in θ and θ' by (B.1). □

Proposition B.3. *If $\overline{\lambda}(\theta,\theta')$ is symmetric in θ and θ', then P_{TFMH} is π-reversible.*

Proof. Simply write

$$\alpha_{\mathrm{TFMH}}(\theta,\theta') = \mathbb{I}(\overline{\lambda}(\theta,\theta') < R)\alpha_{\mathrm{FMH}}(\theta,\theta') + \mathbb{I}(\overline{\lambda}(\theta,\theta') \geq R)\alpha_{\mathrm{MH}}(\theta,\theta').$$

The result then follows from the symmetry of the indicator functions, (B.1), and the proof of Proposition B.2. □

B.2. Ergodic Properties

We provide a brief background to the theory of φ-irreducible Markov Chains. See (Meyn & Tweedie, 2009) for a comprehensive treatment.

For a transition kernel P, we inductively define the transition kernel P^k for $k \geq 1$ by setting $P^1 := P$ and

$$P^k(\theta, A) := \int P(\theta, d\theta')P^{k-1}(\theta', A)d\theta'$$

for $k > 1$, where $\theta \in \Theta$ and $A \subseteq \Theta$ is measurable. Given a nontrivial measure φ on Θ, we say P is *φ-irreducible* if $\varphi(A) > 0$ implies $P^k(\theta, A) > 0$ for some $k \geq 1$. For φ-irreducible P, we define a *k-cycle* of P to be a partition $D_1, \cdots, D_k, N$ of Θ such that $\varphi(N) = 0$, and for all $1 \leq i \leq k$, if $\theta \in D_i$ then $P(\theta, D_{i+1}) = 1$. (Here $i+1$ is meant modulo k.) If there exists a k-cycle with $k > 1$, we say that P is *periodic*; otherwise it is *aperiodic*.

If P is φ-irreducible and aperiodic and has invariant distribution π, we say P is *geometrically ergodic* if there exists constants $\rho < 1$, $C < \infty$, and a π-a.s. finite function $V \geq 1$ such that

$$\|P^k(\theta, \cdot) - \pi\|_V \leq C\, V(\theta)\rho^k$$

for all $\theta \in \Theta$ and $k \geq 1$. Here $\|\cdot\|_V$ denotes the V-norm on signed measures defined by

$$\|\mu\|_V = \sup_{|f| \leq V} |\pi(f)|,$$

where $\pi(f) := \int f(\theta)\pi(d\theta)$. By (Roberts & Rosenthal, 1997, Proposition 2.1), this is equivalent to the apparently weaker condition that there exist some constant $\rho > 0$ and π-a.s. finite function M such that

$$\|P^k(\theta, \cdot) - \pi\|_{\mathrm{TV}} \leq M(\theta)\rho^k$$

for all $\theta \in \Theta$ and $k \geq 1$, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance on signed measures.

Our interest in geometric ergodicity is largely due to the implications it has for the *asymptotic variance* of the ergodic averages produced by a transition kernel. Suppose $(\theta_k)_{k\geq 1}$ is a stationary Markov chain with transition kernel P having invariant distribution π. For $f \in L^2(\pi)$, the asymptotic variance for the ergodic averages of f is defined

$$\mathrm{var}(f, P) := \lim_{k\to\infty} \mathrm{Var}\left(\sqrt{k}(\frac{1}{k}\sum_{i=1}^{k} f(\theta_k) - \pi(f))\right) = \lim_{k\to\infty} \frac{1}{k}\mathrm{Var}(\sum_{i=1}^{k} f(\theta_k)).$$

We abuse notation a little and denote the variance of $f(\theta)$ where $\theta \sim \pi$ by $\mathrm{var}(f, \pi)$.

Of interest is also the (right) *spectral gap*, which for a π-reversible transition kernel P is defined

$$\mathrm{Gap}(P) := \inf_{f \in L^2(\pi):\pi(f)=0} \frac{\int\int \frac{1}{2}(f(\theta)-f(\theta'))^2 \pi(d\theta)P(\theta,d\theta')}{\int f(\theta)^2\pi(d\theta)}.$$

Finally, it is convenient to define the MH rejection probability

$$r_{\mathrm{MH}}(\theta) := 1 - \int q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta',$$

and similarly r_{FMH} and r_{TFMH} for FMH and TFMH.

Proposition B.4. *P_{TFMH} is φ-irreducible and aperiodic whenever P_{MH} is.*

Proof. We use throughout the easily verified facts $\alpha_{\mathrm{FMH}}(\theta,\theta') \leq \alpha_{\mathrm{TFMH}}(\theta,\theta') \leq \alpha_{\mathrm{MH}}(\theta,\theta')$ and $r_{\mathrm{FMH}}(\theta) \geq r_{\mathrm{TFMH}}(\theta) \geq r_{\mathrm{MH}}(\theta)$ for all $\theta,\theta' \in \Theta$. See Proposition B.1.

For φ-irreducibility, first note that if $\alpha_{\mathrm{MH}}(\theta,\theta') > 0$ then $\alpha_{\mathrm{TFMH}}(\theta,\theta') > 0$. This holds since if $\alpha_{\mathrm{TFMH}}(\theta,\theta') = 0$, then either $\alpha_{\mathrm{MH}}(\theta,\theta') = 0$ or $\alpha_{\mathrm{FMH}}(\theta,\theta') = 0$. In the latter case we must have some $\pi_i(\theta')q_i(\theta',\theta) = 0$, so that $\pi(\theta')q(\theta,\theta') = 0$, and hence again $\alpha_{\mathrm{MH}}(\theta,\theta') = 0$.

We now show by induction on $k \in \mathbb{Z}_{\geq 1}$ that for all $\theta \in \Theta$, $P^k_{\mathrm{MH}}(\theta,A) > 0$ implies $P^k_{\mathrm{TFMH}}(\theta,A) > 0$. For $k = 1$, suppose $P_{\mathrm{MH}}(\theta,A) > 0$. Then either $r_{\mathrm{MH}}(\theta)\mathbb{I}_A(\theta) > 0$ or $\int_A q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' > 0$. In the former case we we have

$$r_{\mathrm{TFMH}}(\theta)\mathbb{I}_A(\theta) \geq r_{\mathrm{MH}}(\theta)\mathbb{I}_A(\theta) > 0.$$

In the latter case the above considerations give

$$\begin{aligned}\mathrm{Leb}(\{\theta' \in A \mid q(\theta,\theta')\alpha_{\mathrm{TFMH}}(\theta,\theta') > 0\}) &= \mathrm{Leb}(\{\theta' \in A \mid q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta') > 0\}) \\ &> 0.\end{aligned}$$

Either way we have $P_{\mathrm{TFMH}}(\theta,A) > 0$.

Suppose now $P^{k-1}_{\mathrm{MH}}(\theta,A) > 0$ implies $P^{k-1}_{\mathrm{TFMH}}(\theta,A) > 0$. Then observe

$$P^k_{\mathrm{MH}}(\theta,A) = r_{\mathrm{MH}}(\theta)P^{k-1}_{\mathrm{MH}}(\theta,A) + \int q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')P^{k-1}_{\mathrm{MH}}(\theta',A)d\theta'$$

and likewise *mutatis mutandis* for $P^k_{\mathrm{TFMH}}(\theta,A)$. Thus if $P^k_{\mathrm{MH}}(\theta,A) > 0$, one possibility is $r_{\mathrm{MH}}(\theta)P^{k-1}_{\mathrm{MH}}(\theta,A) > 0$, which implies $r_{\mathrm{TFMH}}(\theta) > 0$ and, by the induction hypothesis, $P^{k-1}_{\mathrm{TFMH}}(\theta,A) > 0$. The only other possibility is

$$\begin{aligned}\mathrm{Leb}(\{\theta' \in \Theta \mid q(\theta,\theta')\alpha_{\mathrm{TFMH}}(\theta,\theta')P^{k-1}_{\mathrm{TFMH}}(\theta',A) > 0\}) &= \mathrm{Leb}(\{\theta' \in \Theta \mid q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')P^{k-1}_{\mathrm{MH}}(\theta',A) > 0\}) \\ &> 0,\end{aligned}$$

again by the induction hypothesis. Either way, as desired $P^k_{\mathrm{TFMH}}(\theta,A) > 0$. It now follows that P_{TFMH} is φ-irreducible when P_{MH} is.

Now suppose P_{MH} and hence P_{TFMH} is φ-irreducible. If P_{TFMH} is periodic, then there exists a k-cycle $D_1,\cdots,D_k,N$ for P_{TFMH} with $k > 1$. But now if $\theta \in D_i$, then $\mathbb{I}_{D_{i+1}}(\theta) = 0$ and so

$$\begin{aligned}P_{\mathrm{MH}}(\theta,D_{i+1}) &= \int_{D_{i+1}} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' \\ &\geq \int_{D_{i+1}} q(\theta,\theta')\alpha_{\mathrm{TFMH}}(\theta,\theta')d\theta' \\ &= P_{\mathrm{TFMH}}(\theta,D_{i+1}) \\ &= 1.\end{aligned}$$

Thus the same partition is a k-cycle for P_{MH} which is therefore periodic. □

Theorem B.1. *If P_{MH} is φ-irreducible, aperiodic, and geometrically ergodic, then P_{TFMH} is too if*

$$\delta := \inf_{\overline{\lambda}(\theta,\theta')<R} \alpha_{\mathrm{FMH}}(\theta,\theta') \vee \alpha_{\mathrm{FMH}}(\theta',\theta) > 0.$$

In this case, $\mathrm{Gap}(P_{\mathrm{FMH}}) \geq \delta \mathrm{Gap}(P_{\mathrm{MH}})$*, and for* $f \in L^2(\pi)$

$$\mathrm{var}(f, P_{\mathrm{TFMH}}) \leq (\delta^{-1} - 1)\mathrm{var}(f, \pi) + \delta^{-1}\mathrm{var}(f, P_{\mathrm{MH}}).$$

Proof. Our proof of this result is similar to (Banterle et al., 2015, Proposition 1), but differs in its use of Proposition B.1 to express the relationship between MH and FMH exactly.

For $\theta \in \Theta$, let

$$\mathcal{R}(\theta) := \{\theta' \in \Theta \mid \overline{\lambda}(\theta,\theta') < R\}.$$

Whenever $\theta \in \Theta$ and $A \subseteq \Theta$ is measurable,

$$\begin{aligned}
P_{\mathrm{TFMH}}(\theta, A) &= r_{\mathrm{TFMH}}(\theta)\mathbb{I}_A(\theta) + \int_{\mathcal{R}(\theta)\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')(\alpha_{\mathrm{FMH}}(\theta,\theta') \vee \alpha_{\mathrm{FMH}}(\theta',\theta))d\theta' \\
&\quad + \int_{\mathcal{R}(\theta)^c\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' \\
&\geq r_{\mathrm{MH}}(\theta)\mathbb{I}_A(\theta) + \delta\int_{\mathcal{R}(\theta)\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' + \int_{\mathcal{R}(\theta)^c\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' \\
&\geq \delta P_{\mathrm{MH}}(\theta, A).
\end{aligned}$$

The last line follows since certainly $\delta \leq 1$.

Suppose $\delta > 0$. If P_{MH} is geometrically ergodic, then (Jones et al., 2014, Theorem 1) entails that P_{TFMH} is geometrically ergodic also. The remaining claims follow directly from (Andrieu et al., 2018, Lemma 32). □

C. Fast Simulation of Bernoulli Random Variables

For sake of completeness, we provide here the proof of validity of Algorithm 1. It combines the Fukui-Todo procedure (Fukui & Todo, 2009) with a thinning argument.

Proposition C.1. *If*

$$N \sim \mathrm{Poisson}\left(\overline{\lambda}(\theta,\theta')\right)$$

$$X_1, \cdots, X_N \overset{\mathrm{iid}}{\sim} \mathrm{Categorical}((\overline{\lambda}_i(\theta,\theta')/\overline{\lambda}(\theta,\theta'))_{1\leq i\leq m})$$

$$B_j \sim \mathrm{Bernoulli}(\lambda_{X_j}(\theta,\theta')/\overline{\lambda}_{X_j}(\theta,\theta')) \text{ independently for } 1 \leq j \leq N$$

then $\mathbb{P}(B = 0) = \alpha_{\mathrm{FMH}}(\theta,\theta')$ *where* $B = \sum_{j=1}^N B_j$ *(and* $B = 0$ *if* $N = 0$*).*

Proof. Letting

$$\lambda(\theta,\theta') := \sum_{i=1}^m \lambda_i(\theta,\theta'),$$

our goal is to show that $\mathbb{P}(B = 0) = \exp(-\lambda(\theta,\theta'))$. For brevity we omit all dependences on θ and θ' in the following.

Observe the random variables B_j's are i.i.d. with

$$\mathbb{P}(B_j = 0) = \sum_{i=1}^m \underbrace{\mathbb{P}(X_j = i)}_{=\overline{\lambda}_i/\overline{\lambda}} \underbrace{\mathbb{P}(B_j = 0 | X_j = i)}_{=1-\lambda_i/\overline{\lambda}_i} = \frac{\overline{\lambda} - \lambda}{\overline{\lambda}}.$$

Thus

$$\begin{aligned}\mathbb{P}(B=0) &= \sum_{\ell=0}^{\infty} \underbrace{\mathbb{P}(N=\ell)}_{=\exp(-\overline{\lambda})\overline{\lambda}^{\ell}/\ell!} \mathbb{P}(B_1=0)^{\ell} \\ &= \exp(-\overline{\lambda}) \sum_{\ell=0}^{\infty} \frac{(\overline{\lambda}-\lambda)^{\ell}}{\ell!} \\ &= \exp(-\lambda)\end{aligned}$$

as desired. □

D. Upper Bounds

We refer the reader to Section A for an explanation of multi-index notation β.

Proposition D.1. *If each U_i is $(k+1)$-times continuously differentiable with*

$$\overline{U}_{k+1,i} \geq \sup_{\substack{\theta\in\Theta \\ |\beta|=k+1}} |\partial^{\beta} U_i(\theta)|,$$

then

$$-\log\left(1 \wedge \frac{\pi_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\pi_i(\theta)\widehat{\pi}_{k,i}(\theta')}\right) \leq (\|\theta-\widehat{\theta}\|_1^{k+1} + \|\theta'-\widehat{\theta}\|_1^{k+1}) \frac{\overline{U}_{k+1,i}}{(k+1)!}.$$

Proof. We have

$$\begin{aligned}-\log\left(1 \wedge \frac{\pi_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\pi_i(\theta)\widehat{\pi}_{k,i}(\theta')}\right) &= 0 \vee (U_i(\theta') - \widehat{U}_{k,i}(\theta') - U_i(\theta) + \widehat{U}_{k,i}(\theta)) \\ &\leq |U_i(\theta') - \widehat{U}_{k,i}(\theta')| + |U_i(\theta) - \widehat{U}_{k,i}(\theta)|.\end{aligned}$$

Notice that $U_i(\theta) - \widehat{U}_{k,i}(\theta)$ is just the remainder of a Taylor expansion. As such, for each θ, Taylor's remainder theorem gives for some $\widetilde{\theta} \in \Theta$

$$\begin{aligned}|U_i(\theta) - \widehat{U}_{k,i}(\theta)| &= \left|\frac{1}{(k+1)!} \sum_{|\beta|=k+1} \partial^{\beta} U_i(\widetilde{\theta})(\theta-\widehat{\theta})^{\beta}\right| \\ &\leq \frac{\overline{U}_{k+1,i}}{(k+1)!} \sum_{|\beta|=k+1} \frac{|(\theta-\widehat{\theta})^{\beta}|}{\beta!} \\ &\leq \frac{\overline{U}_{k+1,i}}{(k+1)!} \|\theta-\widehat{\theta}\|_1^{k+1}.\end{aligned}$$

The result now follows. □

E. Reversible Proposals

E.1. General Conditions for Reversibility

We can handle both the first and second-order cases with the following Proposition.

Proposition E.1. *Suppose*

$$q(\theta,\theta') = \mathrm{Normal}(\theta' \mid A\theta + b, C)$$

and

$$-\log\hat{\pi}(\theta) = \frac{1}{2}\theta^{\top} D\theta + e^{\top}\theta + \mathrm{const}$$

where $A, C, D \in \mathbb{R}^{d\times d}$ with $C \succ 0$, and $b, e \in \mathbb{R}^d$. Then q is $\hat{\pi}$-reversible if and only if the following conditions hold:

$$\begin{aligned} A^\top C^{-1} &= C^{-1}A && \text{(E.1)} \\ A^2 &= I_d - CD && \text{(E.2)} \\ (A^\top + I_d)b &= -Ce, && \text{(E.3)} \end{aligned}$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix.

Proof. Let

$$F(\theta, \theta') := -\log \hat{\pi}(\theta) - \log q(\theta, \theta').$$

Note that q is $\hat{\pi}$-reversible precisely when F is symmetric in its arguments. Since F is a polynomial of the form

$$F(\theta, \theta') = \frac{1}{2}\theta^\top J\theta + \frac{1}{2}\theta'^\top K\theta' + \theta^\top L\theta' + m^\top\theta + n^\top\theta' + \text{const}, \tag{E.4}$$

where $J, K, L \in \mathbb{R}^{d\times d}$ and $m, n \in \Theta$, then by equating coefficients it follows that $F(\theta, \theta') = F(\theta', \theta)$ precisely when

$$\begin{aligned} J &= K && \text{(E.5)} \\ L &= L^\top && \text{(E.6)} \\ m &= n. && \text{(E.7)} \end{aligned}$$

Now, we can expand

$$\begin{aligned} -\log q(\theta, \theta') &= \frac{1}{2}(\theta' - A\theta - b)^\top C^{-1}(\theta' - A\theta - b) + \text{const} \\ &= \frac{1}{2}\theta'^\top C^{-1}\theta' - (A\theta + b)^\top C^{-1}\theta' + \frac{1}{2}(A\theta + b)^\top C^{-1}(A\theta + b) + \text{const} \\ &= \frac{1}{2}\theta^\top A^\top C^{-1}A\theta + \frac{1}{2}\theta'^\top C^{-1}\theta' - \theta^\top A^\top C^{-1}\theta' + b^\top C^{-1}A\theta - b^\top C^{-1}\theta' + \frac{1}{2}b^\top C^{-1}b \\ &\quad + \text{const} \end{aligned}$$

Since $-\log q(\theta, \theta')$ must be the only source of terms in (E.4) containing both θ and θ', we see immediately that

$$L = -A^\top C^{-1},$$

and thus from (E.6) we have $-A^\top C^{-1} = -(C^{-1})^\top A$. Since $C \succ 0$, C^{-1} is symmetric and this condition becomes (E.1). Next we see that

$$\begin{aligned} J &= A^\top C^{-1}A + D \\ K &= C^{-1}, \end{aligned}$$

and from (E.5) and (E.1) we require $C^{-1}A^2 + D = C^{-1}$, or equivalently (E.2). Finally, since

$$\begin{aligned} m &= A^\top C^{-1}b + e \\ n &= -C^{-1}b, \end{aligned}$$

we require from (E.7) that $A^\top C^{-1}b + e = -C^{-1}b$, which combined with (E.1) gives (E.3).

Since (E.5), (E.6), and (E.7) are necessary and sufficient for symmetry of F, we see that (E.1), (E.2), and (E.3) are necessary and sufficient for reversibility also. □

We now specialise this to the first and second-order cases.

E.2. First-Order Case

When $k = 1$ we have

$$-\log \widehat{\pi}(\theta) = \widehat{U}_1(\theta) = U(\widehat{\theta}) + \nabla U(\widehat{\theta})^\top(\theta - \widehat{\theta}),$$

so that

$$\begin{aligned} D &= 0 \\ e &= \nabla U(\widehat{\theta}), \end{aligned}$$

and conditions (E.1), (E.2), and (E.3) become

$$\begin{aligned} A^\top C^{-1} &= C^{-1} A \\ A^2 &= I_d \\ (A^\top + I_d)b &= -C\nabla U(\widehat{\theta}). \end{aligned}$$

E.3. Second-Order Case

When $k = 2$,

$$-\log \widehat{\pi}(\theta) = \widehat{U}_2(\theta) = U(\widehat{\theta}) + \nabla U(\widehat{\theta})^\top(\theta - \widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^\top \nabla^2 U(\widehat{\theta})(\theta - \widehat{\theta}).$$

In this case

$$\begin{aligned} D &= \nabla^2 U(\widehat{\theta}) \\ e &= \nabla U(\widehat{\theta}) - \nabla^2 U(\widehat{\theta})^\top \widehat{\theta}, \end{aligned}$$

so conditions (E.1), (E.2), and (E.3) become

$$\begin{aligned} A^\top C^{-1} &= C^{-1} A \\ A^2 &= I_d - C\nabla^2 U(\widehat{\theta}) \\ (A^\top + I_d)b &= C(\nabla^2 U(\widehat{\theta})^\top \widehat{\theta} - \nabla U(\widehat{\theta})). \end{aligned}$$

A common setting has $\nabla^2 U(\widehat{\theta}) \succ 0$, $A = A^\top$, and $A + I_d$ invertible. In this case the latter two conditions become

$$\begin{aligned} C &= (I_d - A^2)[\nabla^2 U(\widehat{\theta})]^{-1} \\ b &= (I_d - A)(\widehat{\theta} - [\nabla^2 U(\widehat{\theta})]^{-1}\nabla U(\widehat{\theta})). \end{aligned}$$

E.4. Decreasing Norm Property

Under usual circumstances for both first and second-order approximations, when $\|\theta\|$ is large, a $\widehat{\pi}$-reversible q will propose $\theta' \sim q(\theta, \cdot)$ with smaller norm than θ. This is made precise in the following Proposition:

Proposition E.2. *Suppose*

$$q(\theta, \theta') = \mathrm{Normal}(\theta' \mid A\theta + b, C)$$

and

$$-\log \hat{\pi}(\theta) = \frac{1}{2}\theta^\top D\theta + e^\top \theta + \mathrm{const},$$

where $A = A^\top$ is symmetric, $C \succ 0$, and $D \succeq 0$. If q is $\hat{\pi}$-reversible, then $\|A\|_{\mathrm{op}} \leq 1$. If $D \succ 0$ is strict, then $\|A\|_{\mathrm{op}} < 1$ is strict too. In this case, if $\theta' \sim q(\theta, \cdot)$, then $\|\theta\| - \|\theta'\| \to \infty$ in probability as $\|\theta\| \to \infty$.

Proof. By (E.2), we must have $CD = I_d - A^2$. Since $A = A^\top$, this entails $CD = (CD)^\top = DC$ and hence $CD \succeq 0$ since $D, C \succeq 0$. Thus $-CD \preceq 0$ and

$$A^2 = I_d - CD \preceq I_d.$$

Therefore each eigenvalue σ of A must have $|\sigma| \leq 1$, since σ^2 is an eigenvalue of A^2. But A is diagonalisable since it is symmetric, and hence $\|A\|_{\mathrm{op}} \leq 1$.

If $D \succ 0$ is strict, then the above matrix inequalities become strict also, and it follows that each $|\sigma| < 1$ and hence $\|A\|_{\mathrm{op}} < 1$. In this case, suppose $\theta' \sim q(\theta, \cdot)$, and fix $K > 0$ arbitrarily. Let $\epsilon > 0$, and choose $L > 0$ large enough that

$$\mathbb{P}(\theta' \in B(A\theta + b, L)) > 1 - \epsilon.$$

As $\|\theta\| \to \infty$,

$$\|\theta\| - \|A\theta + b\| \geq \|\theta\|(1 - \|A\|_{\mathrm{op}}) + \|b\| \to \infty$$

since $1 - \|A\|_{\mathrm{op}} > 0$, so if $\theta' \in B(A\theta + b, L)$, then $\|\theta\| - \|\theta'\| \to \infty$ also. Thus

$$\mathbb{P}(\|\theta\| - \|\theta'\| > K) > 1 - \epsilon$$

for all $\|\theta\|$ large enough. Taking $\epsilon \to 0$ gives the result. □

In practice the assumption $D \succeq 0$ makes sense, since $\widehat{\theta}$ is chosen near a minimum of U and since D is the Hessian of $\widehat{U}_k \approx U$ for $k = 1, 2$. Likewise, all sensible proposals (certainly including pCN) that we have found are such that A is symmetric, though we acknowledge the possibility that it may be desirable to violate this in some cases.

F. Performance Gains

Lemma F.1. *Suppose that $0 \leq X_n \in L^p$ and $\mathcal{F}_n$ is some σ-algebra for every $n \in \mathbb{Z}_{\geq 1}$. If $\mathbb{E}[X_n^p|\mathcal{F}_n] = O_{\mathbb{P}}(a_n)$, then $\mathbb{E}[X_n^\ell|\mathcal{F}_n] = O_{\mathbb{P}}(a_n^{\ell/p})$ for all $1 \leq \ell \leq p$. If moreover $0 \leq Y_n \in L^p$ gives $\mathbb{E}[Y_n^p|\mathcal{F}_n] = O_{\mathbb{P}}(a_n)$, then $\mathbb{E}[(X_n + Y_n)^p|\mathcal{F}_n] = O_{\mathbb{P}}(a_n)$.*

Proof. The first part is just Jensen's inequality:

$$\mathbb{E}[X_n^\ell|\mathcal{F}_n] \leq \mathbb{E}[X_n^p|\mathcal{F}_n]^{\ell/p} = O_{P_0}(a_n)^{\ell/p} = O_{P_0}(a_n^{\ell/p}).$$

The second part follows from the C_p-inequality, which gives

$$\mathbb{E}[(X_n + Y_n)^p|\mathcal{F}_n] \leq 2^{p-1}\left(\mathbb{E}[X_n^p|\mathcal{F}_n] + \mathbb{E}[Y_n^p|\mathcal{F}_n]\right) = 2^{p-1}(O_{\mathbb{P}}(a_n) + O_{\mathbb{P}}(a_n)) = O_{\mathbb{P}}(a_n).$$

□

Theorem F.1. *Suppose each U_i is $(k+1)$-times continuously differentiable, each $\overline{U}_{k+1,i} \in L^{k+2}$, and $\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{k+1,i}|Y_{1:n}] = O_{P_0}(n)$. Likewise, assume each of $\|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|$, $\|\theta^{(n)} - \theta'^{(n)}\|$, and $\|\widehat{\theta}^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|$ is in L^{k+2}, and each of $\mathbb{E}[\|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|^{k+1}|Y_{1:n}]$, $\mathbb{E}[\|\theta^{(n)} - \theta'^{(n)}\|^{k+1}|Y_{1:n}]$, and $\|\widehat{\theta}^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|^{k+1}$ is $O_{P_0}(n^{-(k+1)/2})$ as $n \to \infty$. Then $\overline{\lambda}$ defined by (13) satisfies*

$$\mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = O_{P_0}(n^{(1-k)/2}).$$

Proof. Write

$$\overline{\lambda}(\theta^{(n)}, \theta'^{(n)}) = \varphi(\theta^{(n)}, \theta'^{(n)}) \sum_{i=1}^{m^{(n)}} \psi_i.$$

with φ and ψ defined by (13) also. Observe that

$$\begin{aligned}
\varphi(\theta^{(n)}, \theta'^{(n)}) &= \|\theta^{(n)} - \widehat{\theta}^{(n)}\|_1^{k+1} + \|\theta'^{(n)} - \widehat{\theta}^{(n)}\|_1^{k+1} \\
&\leq (\|\theta^{(n)} - \widehat{\theta}^{(n)}\|_1 + \|\theta'^{(n)} - \widehat{\theta}^{(n)}\|_1)^{k+1} \\
&\leq (\|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|_1 + \|\theta_{\mathrm{MAP}}^{(n)} - \widehat{\theta}^{(n)}\|_1 + \|\theta'^{(n)} - \theta^{(n)}\|_1 + \|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|_1 + \|\theta_{\mathrm{MAP}}^{(n)} - \widehat{\theta}^{(n)}\|_1)^{k+1} \\
&\leq c(\underbrace{\|\theta'^{(n)} - \theta^{(n)}\| + \|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\| + \|\theta_{\mathrm{MAP}}^{(n)} - \widehat{\theta}^{(n)}\|}_{\in L^{k+2}})^{k+1}
\end{aligned}$$

for some $c > 0$, by the triangle inequality and norm equivalence. We thus have $\varphi(\theta^{(n)}, \theta'^{(n)}) \in L^{(k+2)/(k+1)}$ and

$$\mathbb{E}[\varphi(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}).$$

Likewise,

$$\sum_{i=1}^{m^{(n)}} \psi_i = \frac{1}{(k+1)!} \sum_{i=1}^{m^{(n)}} \overline{U}_{k+1,i} \in L^{k+2}.$$

Together this gives $\overline{\lambda}(\theta^{(n)}, \theta'^{(n)}) \in L^1$ by Hölder's inequality. Since in our setup $(\theta^{(n)}, \theta'^{(n)})$ is conditionally independent of all other randomness given $Y_{1:n}$, we thus have

$$\mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = \mathbb{E}[\varphi(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}]\mathbb{E}[\sum_{i=1}^{m^{(n)}} \psi_i|Y_{1:n}] = O_{P_0}(n^{(1-k)/2}). \tag{F.1}$$

□

Note that in the preceding result we could use weaker integrability assumptions on $\|\theta^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|$, $\|\theta^{(n)} - \theta'^{(n)}\|$, and $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|$ by using a stronger integrability assumption on $\overline{U}_{k+1,i}$. Most generally, for any $\epsilon \geq 0$ we could require each

$$\begin{aligned} \overline{U}_{k+1,i} &\in L^{(k+1+\epsilon)/\epsilon} \\ \|\theta^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|, \|\theta^{(n)} - \theta'^{(n)}\|, \|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\| &\in L^{k+1+\epsilon}. \end{aligned}$$

The case $\epsilon = 0$ would mean $\overline{U}_{k+1,i} \in L^\infty$.

Lemma F.2. *Suppose each U_i is twice continuously differentiable, each $\overline{U}_{2,i} \in L^3$, and $\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i} = O_{P_0}(n)$. If $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\| = O_{P_0}(1/\sqrt{n})$, then $\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\|$ is in $L^{3/2}$ and $O_{P_0}(\sqrt{n})$.*

Proof. By norm equivalence the Hessian satisfies

$$\|\nabla^2 U^{(n)}(\theta)\|_{\mathrm{op}} \leq c\|\nabla^2 U^{(n)}(\theta)\|_1 \leq c \sum_{i=1}^{m^{(n)}} \overline{U}_{2,i}$$

for some $c > 0$ (where $\|\cdot\|_1$ is understood to be applied as if $\nabla^2 U^{(n)}(\theta)$ were a vector), which means $\nabla U^{(n)}$ is $(c\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i})$-Lipschitz. Thus

$$\begin{aligned} \|\nabla U^{(n)}(\widehat{\theta}^{(n)})\| &= \|\nabla U^{(n)}(\widehat{\theta}^{(n)}) - \nabla U^{(n)}(\theta^{(n)}_{\mathrm{MAP}})\| \\ &\leq \underbrace{c(\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i})}_{\in L^3} \underbrace{\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|}_{\in L^{k+2} \subseteq L^3} \end{aligned}$$

since $k \geq 1$. By Cauchy-Schwarz we have therefore $\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\| \in L^{3/2}$.

Similarly, since $\widehat{\theta}^{(n)}$ and $\theta^{(n)}_{\mathrm{MAP}}$ are functions of $Y_{1:n}$,

$$\begin{aligned} \|\nabla U^{(n)}(\widehat{\theta}^{(n)})\| &= \mathbb{E}[\|\nabla U^{(n)}(\widehat{\theta}^{(n)}) - \nabla U^{(n)}(\theta^{(n)}_{\mathrm{MAP}})\||Y_{1:n}] \\ &\leq \mathbb{E}[c(\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i})\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\||Y_{1:n}] \\ &= \underbrace{c\,\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i}|Y_{1:n}]}_{=O_{P_0}(n)} \underbrace{\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|}_{=O_{P_0}(1/\sqrt{n})} \\ &= O_{P_0}(\sqrt{n}). \end{aligned}$$

□

Theorem F.2. *Suppose the assumptions of Theorem 3.1 hold, and additionally that for* $2 \leq \ell \leq k$, *each* $\overline{U}_{\ell,i} \in L^{\ell+1}$, *and* $\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{\ell,i} | Y_{1:n}] = O_{P_0}(n)$. *Then*

$$-\log(1 \wedge \frac{\widehat{\pi}_k^{(n)}(\theta'^{(n)})}{\widehat{\pi}_k^{(n)}(\theta^{(n)})}) = O_{P_0}(1)$$

for all $k \geq 1$.

Proof. It is useful to denote

$$\begin{aligned} U^{(n)}(\theta) &:= \sum_{i=1}^{m^{(n)}} U_i(\theta) \\ \widehat{U}_k^{(n)}(\theta) &:= \sum_{i=1}^{m^{(n)}} \widehat{U}_{k,i}(\theta) = -\log(\widehat{\pi}^{(n)}(\theta)). \end{aligned}$$

Observe that

$$0 \leq -\log(1 \wedge \frac{\widehat{\pi}^{(n)}(\theta'^{(n)})}{\widehat{\pi}^{(n)}(\theta^{(n)})}) \leq |\widehat{U}_k^{(n)}(\theta'^{(n)}) - \widehat{U}_k^{(n)}(\theta^{(n)})|. \tag{F.2}$$

Now,

$$\widehat{U}_k^{(n)}(\theta'^{(n)}) - \widehat{U}_k^{(n)}(\theta^{(n)}) = \langle \nabla U^{(n)}(\widehat{\theta}^{(n)}), \theta'^{(n)} - \theta^{(n)} \rangle + \sum_{2 \leq |\beta| \leq k} \frac{\partial^\beta U^{(n)}(\widehat{\theta}^{(n)})}{\beta!} ((\theta'^{(n)} - \widehat{\theta}^{(n)})^\beta - (\theta^{(n)} - \widehat{\theta}^{(n)})^\beta). \tag{F.3}$$

For the first term here, Cauchy-Schwarz gives

$$\begin{aligned} \mathbb{E}[|\langle \nabla U^{(n)}(\widehat{\theta}^{(n)}), \theta'^{(n)} - \theta^{(n)} \rangle| | Y_{1:n}] &\leq \mathbb{E}[\underbrace{\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\|}_{\in L^{3/2}} \underbrace{\|\theta'^{(n)} - \theta^{(n)}\|}_{\in L^{k+2} \subseteq L^3} | Y_{1:n}] \\ &= \underbrace{\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\|}_{=O_{P_0}(\sqrt{n})} \underbrace{\mathbb{E}[\|\theta'^{(n)} - \theta^{(n)}\| | Y_{1:n}]}_{=O_{P_0}(1/\sqrt{n})} \\ &= O_{P_0}(1). \end{aligned}$$

Integrability follows from Lemma F.2 and Hölder's inequality, and the asymptotic statements from conditional independence, Lemma F.2, and Lemma F.1. For the summation in (F.3), note that

$$|\partial^\beta U^{(n)}(\widehat{\theta}^{(n)})| \leq \sum_{i=1}^{m^{(n)}} |\partial^\beta U_i(\widehat{\theta}^{(n)})| \leq \sum_{i=1}^{m^{(n)}} \overline{U}_{|\beta|,i},$$

and that for some $c > 0$,

$$\begin{aligned} |(\theta'^{(n)} - \widehat{\theta}^{(n)})^\beta - (\theta^{(n)} - \widehat{\theta}^{(n)})^\beta| &\leq \|\theta'^{(n)} - \widehat{\theta}^{(n)}\|_\infty^{|\beta|} + \|\theta^{(n)} - \widehat{\theta}^{(n)}\|_\infty^{|\beta|} \\ &\leq c\|\theta'^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} + c\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} \end{aligned}$$

by norm equivalence. Thus, conditional on $Y_{1:n}$, the absolute value of the summation in (F.3) is bounded above by

$$\begin{aligned} &\sum_{2 \leq |\beta| \leq k} \frac{1}{\beta!} \mathbb{E}[\underbrace{(\sum_{i=1}^{m^{(n)}} \overline{U}_{|\beta|,i})}_{\in L^{|\beta|+1}} (c \underbrace{\|\theta'^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|}}_{\in L^{(|\beta|+1)/|\beta|}} + c \underbrace{\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|}}_{\in L^{(|\beta|+1)/|\beta|}}) | Y_{1:n}] \\ = &\sum_{2 \leq |\beta| \leq k} \frac{c}{\beta!} \underbrace{\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{|\beta|,i} | Y_{1:n}]}_{=O_{P_0}(n)} (\underbrace{\mathbb{E}[\|\theta'^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} | Y_{1:n}]}_{=O_{P_0}(n^{-|\beta|/2})} + \underbrace{\mathbb{E}[\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} | Y_{1:n}]}_{=O_{P_0}(n^{-|\beta|/2})}) \\ = &\ O_{P_0}(1). \end{aligned}$$

Again, integrability follows from Hölder's inequality. The second line holds since $\widehat{\theta}^{(n)} \equiv \widehat{\theta}^{(n)}(Y_{1:n})$ and since $(\theta^{(n)}, \theta'^{(n)})$ is conditionally independent of all other randomness given $Y_{1:n}$. Finally, the asymptotics follow from the law of large numbers and Lemma F.1 (noting that each $|\beta| \geq 2$).

Inspection of (F.3) now shows that (F.3) is $O_{P_0}(1)$ as required. □

F.1. Sufficient Conditions

We are interested in sufficient conditions that guarantee the convergence rate assumptions in Theorem 3.1 will hold. For simplicity we assume throughout that the likelihood of a data point $p(y|\theta)$ admits a density w.r.t. Lebesgue measure and that P_0 also admits a Lebesgue density denoted $p_0(y)$.

F.1.1. Concentration Around the Mode

We first consider the assumption

$$\mathbb{E}[\|\theta^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}).$$

Intuitively, this says that the distance of $\theta^{(n)}$ from the mode is $O(1/\sqrt{n})$, and hence connects directly with standard concentration results on Bayesian posteriors. To establish this rigorously, it is enough to show that for some $\theta^* \in \Theta$ both

$$\mathbb{E}[\|\theta^{(n)} - \theta^*\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}) \tag{F.4}$$

$$\mathbb{E}[\|\theta^{(n)}_{\mathrm{MAP}} - \theta^*\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}), \tag{F.5}$$

which entails the result by Lemma F.1 and the triangle inequality. Note that $\theta^{(n)}_{\mathrm{MAP}} \equiv \theta^{(n)}_{\mathrm{MAP}}(Y_{1:n})$ is deterministic function of the data, so that (F.5) may be written more simply as

$$\sqrt{n}(\theta^{(n)}_{\mathrm{MAP}} - \theta^*) = O_{P_0}(1). \tag{F.6}$$

We give sufficient conditions for (F.4) and (F.6) now.

By Proposition F.1 below, (F.4) holds as soon as we show that

$$\mathbb{E}[\|\sqrt{n}(\theta^{(n)} - \theta^*)\|^{k+1}\mathbb{I}(\|\sqrt{n}(\theta^{(n)} - \theta^*)\| > M_n)|Y_{1:n}] \xrightarrow{P_0} 0, \qquad \text{for all } M_n \to \infty. \tag{F.7}$$

This condition is a consequence of standard assumptions used to prove the Bernstein-von Mises theorem (BvM): in particular, it is (van der Vaart, 1998, (10.9)) when the model is well-specified (i.e. $p_0 = p(y|\theta_0)$ for some $\theta_0 \in \theta$), and (Kleijn & van der Vaart, 2012, (2.16)) in the misspecified case. In both cases

$$\theta^* = \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}(p_0(y) \parallel p(y|\theta)),$$

where $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ denotes the Kullback-Leibler divergence. The key assumption required for (F.7) is then the existence of certain test sequences $\phi_n \equiv \phi_n(Y_{1:n})$ with $0 \leq \phi_n \leq 1$ such that, whenever $\epsilon > 0$, both

$$\int \phi_n(y_{1:n}) \prod_{i=1}^n p_0(y_i) dy_{1:n} \to 0 \qquad \text{and} \qquad \sup_{\|\theta - \theta^*\| \geq \epsilon} \int (1 - \phi_n(y_{1:n})) \prod_{i=1}^n \frac{p(y_i|\theta)}{p(y_i|\theta^*)} p_0(y_i) dy_{1:n} \to 0, \tag{F.8}$$

Note that in the well-specified case these conditions say that ϕ_n is uniformly consistent for testing the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \|\theta - \theta_0\| \geq \epsilon$. Since ϕ_n may have arbitrary form, this requirement does not seem arduous. Sufficient conditions are given by (van der Vaart, 1998, Lemma 10.4, Lemma 10.6) for the well-specified case, and (Kleijn & van der Vaart, 2012, Theorem 3.2) for the misspecified case.

In addition to (F.8), we require in both the well-specified and misspecified cases that the prior $p(\theta)$ be continuous and positive at θ^* and satisfy

$$\int \|\theta\|^{k+1} p(\theta) d\theta < \infty.$$

There are additionally some mild smoothness and regularity conditions imposed on the likelihood, which are naturally stronger in the misspecified case than in the well-specified one. In the well-specified case we require $p(y|\theta)$ is differentiable

in quadratic mean at θ^* (van der Vaart, 1998, (7.1)). In the misspecified case the conditions are more complicated. We omit repeating these for brevity and instead refer the reader to the statements of Lemma 2.1 and Theorem 3.1 in (Kleijn & van der Vaart, 2012).

Lemma F.3. *Suppose a sequence of random variables X_n is $O_{\mathbb{P}}(M_n)$ for every sequence $M_n \to \infty$. Then $X_n = O_{\mathbb{P}}(1)$.*

Proof. Suppose $X_n \neq O_{\mathbb{P}}(1)$. Then, for some $\epsilon > 0$, for every $c > 0$ we have $\mathbb{P}(|X_n| > c) \geq \epsilon$ for infinitely many X_n. This allows us to choose a subsequence X_{n_k} such that $\mathbb{P}(|X_{n_k}| > k) \geq \epsilon$ for each $k \in \mathbb{Z}_{\geq 1}$. Let

$$M_n := \begin{cases} k & \text{if } n = n_k \text{ for some (necessarily unique) } k \\ n & \text{otherwise.} \end{cases}$$

Then $M_n \to \infty$ but $\mathbb{P}(|X_n| > M_n) \geq \epsilon$ occurs for infinitely many n and hence $X_n \neq O_{\mathbb{P}}(M_n)$. □

Proposition F.1. *Suppose that for some $\theta^* \in \Theta$ and $\ell \geq 0$,*

$$\mathbb{E}[\|\sqrt{n}(\theta^{(n)} - \theta^*)\|^\ell \mathbb{I}(\|\sqrt{n}(\theta^{(n)} - \theta^*)\| > M_n)|Y_{1:n}] \xrightarrow{P_0} 0$$

whenever $M_n \to \infty$. Then

$$\mathbb{E}[\|\theta^{(n)} - \theta^*\|^\ell |Y_{1:n}] = O_{P_0}(n^{-\ell/2}).$$

Proof. For $M_n \to \infty$, our assumption lets us write

$$\begin{aligned} n^{\ell/2}\mathbb{E}[\|\theta^{(n)} - \theta^*\|^\ell |Y_{1:n}] &= \mathbb{E}[\|\sqrt{n}(\theta^{(n)} - \theta^*)\|^\ell \mathbb{I}(\|\sqrt{n}(\theta^{(n)} - \theta^*)\| \leq M_n)|Y_{1:n}] + o_{P_0}(1) \\ &\leq M_n^\ell + o_{P_0}(1) \\ &= O_{P_0}(M_n^\ell). \end{aligned}$$

Since M_n was arbitrary, Lemma F.3 entails the left-hand side is $O_{P_0}(1)$, so that

$$\mathbb{E}[\|\theta^{(n)} - \theta^*\|^\ell |Y_{1:n}] = O_{P_0}(n^{-\ell/2}).$$

□

It remains to give conditions for (F.6). Our discussion here is fairly standard. Recall that for $\theta^{(n)}_{\mathrm{MLE}}$ the maximum likelihood estimator,

$$\sqrt{n}(\theta^{(n)}_{\mathrm{MLE}} - \theta^*) = O_{P_0}(1)$$

often holds under under mild smoothness assumptions. We show here that effectively those same assumptions are also sufficient to guarantee a similar result for $\theta^{(n)}_{\mathrm{MAP}}$.

In the following we define

$$\mathcal{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i|\theta).$$

Note that by definition

$$\theta^{(n)}_{\mathrm{MLE}} = \sup_{\theta \in \Theta} \mathcal{L}_n(\theta).$$

Our first result here shows that if both the MAP and the MLE are consistent and the prior is well-behaved, then the MAP is a *near maximiser* of $\mathcal{L}_n$ in the sense that (F.9). Combined with mild smoothness assumptions on the likelihood, (F.9) is a standard condition used to show results such as (F.6). See for instance (van der Vaart, 1998, Theorem 5.23) for a detailed statement.

Proposition F.2. *Suppose for some $\theta^* \in \Theta$ that $\theta^{(n)}_{\mathrm{MAP}}, \theta^{(n)}_{\mathrm{MLE}} \xrightarrow{P_0} \theta^*$ and that the prior $p(\theta)$ is continuous and positive at θ^*, then*

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) \geq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) - o_{P_0}(1/n). \tag{F.9}$$

Proof. Observe that by definition of the MAP,

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MLE}}) \leq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}).$$

We can rewrite this inequality as

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) \geq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log \frac{p(\theta^{(n)}_{\mathrm{MLE}})}{p(\theta^{(n)}_{\mathrm{MAP}})}.$$

The second term on the right-hand side is $o_{P_0}(1/n)$, since our assumption on the prior gives

$$\frac{p(\theta^{(n)}_{\mathrm{MLE}})}{p(\theta^{(n)}_{\mathrm{MAP}})} \xrightarrow{P_0} 1.$$

□

We next consider how to show that the MAP is indeed consistent, as the vast majority of such results in this area only consider the MLE. However, assuming the prior is not pathological, arguments for the consistency of the MLE ought to apply also for the MAP, since the MAP optimises the objective function

$$\mathcal{L}_n(\theta) + \frac{1}{n}\log p(\theta),$$

which is asymptotically equivalent to $\mathcal{L}_n(\theta)$ as $n \to \infty$ whenever $p(\theta) > 0$. By way of example, we show that (van der Vaart, 1998, Theorem 5.7), which can be used to show the consistency of the MLE, also applies to the MAP. For this, we assume that

$$\int |\log p(y|\theta^*)| p_0(y) dy < \infty, \tag{F.10}$$

and define

$$\mathcal{L}(\theta) := \int \log p(y|\theta) p_0(y) dy.$$

Proposition F.3. *Suppose that* (F.10) *holds, that*

$$\sup_{\theta \in \Theta} |\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| \xrightarrow{P_0} 0,$$

and that for some $\epsilon > 0$ *and* $\theta^* \in \Theta$

$$\sup_{\|\theta - \theta^*\| \geq \epsilon} \mathcal{L}(\theta) < \mathcal{L}(\theta^*). \tag{F.11}$$

Further, suppose the prior $p(\theta)$ *is continuous and positive at* θ^**, and that* $\sup_{\theta \in \Theta} p(\theta) < \infty$*. Then both* $\theta^{(n)}_{\mathrm{MLE}}, \theta^{(n)}_{\mathrm{MAP}} \xrightarrow{P_0} \theta^*$.

Proof. For each $\theta \in \Theta$ we have $\mathcal{L}_n(\theta) \xrightarrow{P_0} \mathcal{L}(\theta)$ as $n \to \infty$ by the law of large numbers, and thus $\theta^{(n)}_{\mathrm{MLE}} \xrightarrow{P_0} \theta^*$ by (van der Vaart, 1998, Theorem 5.7). Since $p(\theta)$ is continuous and positive at θ^*, this yields that

$$P_0(p(\theta^{(n)}_{\mathrm{MLE}}) > c) \to 1 \tag{F.12}$$

for some $c > 0$, as well as

$$\frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MLE}}) = O_{P_0}(1/n).$$

Now, by maximality

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MLE}}) \leq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}) \leq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}).$$

Observe that it implies that $p(\theta^{(n)}_{\mathrm{MLE}}) \leq p(\theta^{(n)}_{\mathrm{MAP}})$. Together with (F.12) and our boundedness assumption on the prior, this gives

$$\frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}) = O_{P_0}(1/n).$$

We can thus write

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) \geq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + O_{P_0}(1/n).$$

The result now follows from (van der Vaart, 1998, Theorem 5.7). □

Observe that by negating (F.11) and adding the constant $\int p_0(y) \log p_0(y) dy$ to both sides, we see it is equivalent to the perhaps more intuitive condition

$$\inf_{\|\theta - \theta^*\| \geq \epsilon} D_{\mathrm{KL}}(p_0(y) \parallel p(y|\theta)) > D_{\mathrm{KL}}(p_0(y) \parallel p(y|\theta^*)).$$

F.1.2. Scaling of the Proposal

We now consider the assumption

$$\mathbb{E}[\|\theta^{(n)} - \theta'^{(n)}\|^{k+1} | Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}). \tag{F.13}$$

Intuitively this holds if we scale our proposal like $1/\sqrt{n}$. We consider here proposals based on a noise distribution $\xi^{(n)} \overset{\mathrm{iid}}{\sim} \mathrm{Normal}(0, I_d)$, but generalisations are possible. We immediately obtain (F.13) for instance with the scaled random walk proposal (15), for which

$$\theta'^{(n)} = \theta^{(n)} + \frac{\sigma}{\sqrt{n}} \xi^{(n)}.$$

Similarly, the $\widehat{\pi}_1$-reversible proposal defined by (17) has

$$\theta'^{(n)} = \theta^{(n)} - \frac{1}{2n} \nabla U^{(n)}(\widehat{\theta}^{(n)}) + \frac{\sigma}{\sqrt{n}} \xi^{(n)},$$

with $\xi^{(n)} \overset{\mathrm{iid}}{\sim} \mathrm{Normal}(0, I_d)$. If the conditions of Lemma F.2 hold, then the second term is $O_{P_0}(1/\sqrt{n})$ and (F.13) follows.

More generally we can consider trying to match the covariance of our noise to the covariance of our target. Intuitively, under usual circumstances, $[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}$ is approximately proportional to the inverse observed Fisher information at θ^*, and hence preconditioning $\xi^{(n)}$ by $S^{(n)}$ such that

$$S^{(n)} {S^{(n)}}^\top = [\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}$$

matches our proposal to the characteristics of the target. Such an $S^{(n)}$ can be computed for instance via a Cholesky decomposition.

Under usual circumstances this achieves a correctly scaled proposal. In particular, if

$$\widehat{\theta}^{(n)} \xrightarrow{P_0} \theta^* \tag{F.14}$$

$$\frac{1}{n} \partial_j \partial_k U^{(n)}(\theta^*) \xrightarrow{P_0} \mathcal{I}_{j,k} \tag{F.15}$$

for some constants $\mathcal{I}_{j,k}$, then Proposition F.4 below entails $\|S^{(n)}\|_{\mathrm{op}} = O_{P_0}(1/\sqrt{n})$. Thus (F.13) holds for the preconditioned random walk proposal (16) for which

$$\theta'^{(n)} = \theta^{(n)} + S^{(n)} \xi^{(n)},$$

since

$$\|S^{(n)} \xi^{(n)}\| \leq \|S^{(n)}\|_{\mathrm{op}} \|\xi^{(n)}\| = O_{P_0}(1/\sqrt{n}). \tag{F.16}$$

The same is also true a pCN proposal. In this case

$$\theta'^{(n)} - \theta^{(n)} = (\sqrt{\rho} - 1)(\theta^{(n)} - \widehat{\theta}^{(n)}) + (\sqrt{\rho} - 1)([\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1} \nabla U^{(n)}(\widehat{\theta}^{(n)})) + \sqrt{1 - \rho} S^{(n)} \xi^{(n)}.$$

Note that here the first term satisfies

$$E[\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^3 | Y_{1:n}] = O_{P_0}(n^{-3/2}),$$

while the remaining two terms are $O_{P_0}(1/\sqrt{n})$ by Lemma F.2 and (F.16). This gives F.13 by Lemma F.1.

Condition (F.14) holds for instance under the assumptions of Theorem 3.1 and provided concentration around θ^* of the kind described in Section F.1.1 occurs. Condition (F.15) will also often hold in practice. For instance, if

$$U^{(n)}(\theta) = -\log p(\theta) - \sum_{i=1}^{n} \log p(Y_i|\theta),$$

and if the prior is positive at θ^*, then for all $1 \leq j, k \leq d$ the law of large numbers gives

$$\begin{aligned} \frac{1}{n}\partial_j\partial_k U^{(n)}(\theta^*) &= -\frac{1}{n}\partial_j\partial_k \log p(\theta^*) - \frac{1}{n}\sum_{i=1}^{n} \partial_j\partial_k \log p(Y_i|\theta^*) \\ &\xrightarrow{P_0} -\int \partial_j\partial_k \log p(y|\theta^*) p_0(y) dy \end{aligned}$$

when the derivatives and the integral exists. More generally our model may be specified conditional on i.i.d. covariates X_i so that

$$U^{(n)}(\theta) = -\log p(\theta) - \sum_{i=1}^{n} \log p(Y_i|\theta, X_i) + \log p(X_i),$$

in which case the same argument still applies. (Note that here abuse notation by considering our data $Y_i \equiv (X_i, Y_i)$, where the right-hand Y_i are response variables.)

Proposition F.4. *Suppose for some $\theta^* \in \Theta$ we have $\widehat{\theta}^{(n)} \xrightarrow{P_0} \theta^*$ and*

$$\frac{1}{n}\partial_j\partial_k U^{(n)}(\theta^*) \xrightarrow{P_0} \mathcal{I}_{jk}$$

for all $1 \leq j, k \leq d$. Suppose moreover that each $\overline{U}_{3,i} \in L^1$ and each $\nabla^2 U^{(n)}(\widehat{\theta}^{(n)}) \succ 0$. If $[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1} = S^{(n)} S^{(n)\top}$ for some $S^{(n)} \in \mathbb{R}^{d\times d}$, then

$$\|S^{(n)}\|_{\mathrm{op}} = O_{P_0}(1/\sqrt{n}).$$

Proof. Suppose $|\beta| = 2$. Note that since for each i and θ

$$\|\nabla\partial^\beta U_i(\theta)\| \leq c\|\nabla\partial^\beta U_i(\theta)\|_1 = c\sum_{j=1}^{d} |\partial_j\partial^\beta U_i(\theta)| \leq cd\overline{U}_{3,i},$$

for some $c > 0$ by norm equivalence, it follows that $\partial^\beta U_i$ is $cd\overline{U}_{3,i}$-Lipschitz. Consequently for each θ

$$\left|\frac{1}{n}\partial^\beta U^{(n)}(\theta) - \frac{1}{n}\partial^\beta U^{(n)}(\theta^*)\right| \leq \frac{1}{n}\sum_{i=1}^{m^{(n)}} |\partial^\beta U_i(\theta) - \partial^\beta U_i(\theta^*)| \leq \frac{1}{n}(\sum_{i=1}^{m^{(n)}} \overline{U}_{3,i}) cd\|\theta - \theta^*\|.$$

Thus given $K, \eta > 0$

$$\begin{aligned} \mathbb{P}\left(\sup_{\|\theta-\theta^*\|<K} \left|\frac{1}{n}\partial^\beta U^{(n)}(\theta) - \frac{1}{n}\partial^\beta U^{(n)}(\theta^*)\right| > \eta\right) &\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{m^{(n)}} \overline{U}_{3,i} > \eta c^{-1}d^{-1}K^{-1}\right), \\ &\leq \frac{\mathbb{E}[\overline{U}_{3,i}]}{\eta c^{-1}d^{-1}K^{-1}}, \end{aligned}$$

by Markov's inequality. It is clear that given any $\eta > 0$ the right-hand side can be made arbitrarily small by taking $K \to 0$, which yields $n^{-1}\partial^\beta U^{(n)}(\theta)$ is stochastic equicontinuous at θ^*, and consequently that

$$\frac{1}{n}\partial^\beta U^{(n)}(\widehat{\theta}^{(n)}) - \frac{1}{n}\partial^\beta U^{(n)}(\theta^*) \xrightarrow{P_0} 0,$$

see (Pollard, 2012, page 139).

Define the matrix $\mathcal{I} \in \mathbb{R}^{d\times d}$ by the constants $\mathcal{I}_{jk}$. We thus have

$$\frac{1}{n}\nabla^2 U^{(n)}(\widehat{\theta}^{(n)}) \xrightarrow{P_0} \mathcal{I}$$

since it converges element-wise. Thus by the continuous mapping theorem

$$n\|\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})^{-1}\|_{\text{op}} \xrightarrow{P_0} \|\mathcal{I}^{-1}\|_{\text{op}},$$

from which it follows that

$$\|[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}\|_{\text{op}} = O_{P_0}(1/n).$$

It is a standard result from linear algebra that

$$\|[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}\|_{\text{op}} = \|S^{(n)}\|^2_{\text{op}},$$

which gives the result. □

G. Applications

We give here the results of applying our method to a logistic regression and a robust linear regression example. In both cases we write our covariates as x_i and responses as y_i, and our target is the posterior

$$\pi(\theta) = p(\theta|x_{1:n}, y_{1:n}) \propto p(\theta)\prod_{i=1}^{n} p(y_i|\theta, x_i).$$

G.1. Logistic Regression

In this case we have $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$, and

$$p(y_i|\theta, x_i) = \text{Bernoulli}(y_i|\frac{1}{1+\exp(-\theta^\top x_i)}).$$

For simplicity we assume a flat prior $p(\theta) \equiv 1$, which allows factorising π like (8) with $m = n$ and $\widetilde{\pi}_i(\theta) = p(y_i|\theta, x_i)$. It is then easy to show that

$$U_i(\theta) = -\log\widetilde{\pi}_i(\theta) = \log(1+\exp(\theta^T x_i)) - y_i\theta^\top x_i.$$

We require upper bounds $\overline{U}_{k+1,i}$ of the form (12) for these terms. For this we let $\sigma(z) = 1/(1+\exp(-z))$ and note the identity $\sigma'(z) = \sigma(z)(1-\sigma(z))$, which entails

$$\partial_j\sigma(\theta^\top x_i) = -x_{ij}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2).$$

We then have

$$\begin{aligned}
\partial_j U_i(\theta) &= x_{ij}(\sigma(\theta^\top x_i) - y_i)\\
\partial_k\partial_j U_i(\theta) &= x_{ij}x_{ik}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2)\\
\partial_\ell\partial_k\partial_j U_i(\theta) &= x_{ij}x_{ik}(x_{i\ell}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2) - 2\sigma(\theta^\top x_i)x_{i\ell}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2))\\
&= x_{ij}x_{ik}x_{i\ell}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2)(1 - 2\sigma(\theta^\top x_i)).
\end{aligned}$$

It is possible to show that (whether $y_i = 0$ or $y_1 = 1$)

$$\begin{aligned}
\sup_{t\in\mathbb{R}}|\sigma(t) - y_i| &= 1\\
\sup_{t\in\mathbb{R}}|\sigma(t) - \sigma(t)^2| &= \frac{1}{4}\\
\sup_{t\in\mathbb{R}}|(\sigma(t) - \sigma(t)^2)(1 - 2\sigma(t))| &= \frac{1}{6\sqrt{3}}.
\end{aligned}$$

Thus setting

$$\begin{aligned}\overline{U}_{1,i} &:= \max_{1\le j\le d}|x_{ij}| \\ \overline{U}_{2,i} &:= \frac{1}{4}\max_{1\le j\le d}|x_{ij}|^2 \\ \overline{U}_{3,i} &:= \frac{1}{6\sqrt{3}}\max_{1\le j\le d}|x_{ij}|^3\end{aligned}$$

satisfies (12).

In Figure 1 we compare the histogram of the samples of the first coordinate θ_1 to the marginal of the Gaussian approximation. This is done for $n = 2048$, the smallest data size for which we saw a significant ESS improvement of SMH-2 over MH, and for larger n showing the convergence of the Gaussian approximation and the posterior.

In Figure 2 we demonstrate the performance of the algorithm where θ is of dimension 20. The results are qualitatively similar to the 10-dimensional case in that SMH-2 eventually performs better than MH as the number of data increases. However for the 20-dimensional model SMH-2 yields superior performance to MH around the point at which n exceeds 32768, whereas in the 10-dimensional model this happens for n exceeding 2048.

G.2. Robust Linear Regression

Here $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We use a flat prior $p(\theta) \equiv 1$, and the likelihood is given by

$$p(y_i|\theta, x_i) \quad = \quad \mathrm{Student}(y_i - \theta^\top x_i \mid \nu).$$

Here $\mathrm{Student}(\nu)$ denotes the Student-t distribution with ν degrees of freedom that the user will specify. This gives

$$U_i(\theta) = \frac{\nu+1}{2}\log\left(1 + \frac{(y_i - \theta^\top x_i)^2}{\nu}\right).$$

To derive bounds necessary for (12), let $\phi_i(\theta) := y_i - \theta^T x_i$ and note that $\partial_j \phi_i(\theta) = -x_{ij}$. Then we have

$$\begin{aligned}U_i(\theta) &= \frac{\nu+1}{2}\log\left(1+\frac{\phi_i(\theta)^2}{\nu}\right) \\ \partial_j U_i(\theta) &= -(\nu+1)x_{ij}\frac{\phi_i(\theta)}{\nu+\phi_i(\theta)^2} \\ \partial_k\partial_j U_i(\theta) &= -(\nu+1)x_{ij}\frac{-x_{ik}(\nu+\phi_i(\theta)^2)+2x_{ik}\phi_i(\theta)^2}{\nu+\phi_i(\theta)} \\ &= (\nu+1)x_{ij}x_{ik}\frac{\nu-\phi_i(\theta)^2}{(\nu+\phi_i(\theta)^2)^2} \\ \partial_\ell\partial_k\partial_j U_i(\theta) &= (\nu+1)x_{ij}x_{ik}\frac{2x_{i\ell}\phi_i(\theta)(\nu+\phi_i(\theta)^2)^2+4x_{i\ell}(\nu-\phi_i(\theta)^2)(\nu+\phi_i(\theta)^2)\phi_i(\theta)}{(\nu+\phi_i(\theta)^2)^4} \\ &= -2(\nu+1)x_{ij}x_{ik}x_{i\ell}\frac{\phi_i(\theta)(\phi_i(\theta)^2-3\nu)}{(\nu+\phi_i(\theta)^2)^3}\end{aligned}$$

In general,

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\frac{t}{\nu+t^2}\right| &= \frac{1}{2\sqrt{\nu}} \\ \sup_{t\in\mathbb{R}}\left|\frac{\nu-t^2}{(\nu+t^2)^2}\right| &= \frac{1}{\nu} \\ \sup_{t\in\mathbb{R}}\left|\frac{t(t^2-3\nu)}{(\nu+t^2)^3}\right| &= \frac{3+2\sqrt{2}}{8\nu^{3/2}},\end{aligned}$$

so setting

$$\begin{aligned}\overline{U}_{1,i} &:= \frac{\nu+1}{2\sqrt{\nu}} \max_{1\le j\le d} |x_{ij}| \\ \overline{U}_{2,i} &:= \frac{\nu+1}{\nu} \max_{1\le j\le d} |x_{ij}|^2 \\ \overline{U}_{3,i} &:= \frac{(\nu+1)(3+2\sqrt{2})}{4\nu^{3/2}} \max_{1\le j\le d} |x_{ij}|^3\end{aligned}$$

satisfies (12).

In Figure 3 we show effective sample size (ESS) per second for the robust linear regression model; this experiment mimics the conditions of Figure 2 in the main text, where we used a logistic regression model. The performance for this model is qualitatively similar to that for logistic regression. Figures 4 and 5 show the ESS and acceptance rate for pCN proposals as ρ is varied. These mimic Figures 3 and 4 in the main text. For these experiments we use synthetic data, taking an $n \times 10$ matrix X with elements drawn independently from a standard normal distribution, and simulate $y_i = \sum_j X_{ij} + \epsilon$ where ϵ itself is drawn from a standard normal distribution. We choose as the model parameter $\nu = 4.0$.

References

Andrieu, C., Lee, A., and Vihola, M. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872, 2018.

Banterle, M., Grazian, C., Lee, A., and Robert, C. P. Accelerating Metropolis–Hastings algorithms by delayed acceptance. *arXiv preprint arXiv:1503.00996*, 2015.

Fukui, K. and Todo, S. Order-n cluster Monte Carlo method for spin systems with long-range interactions. *Journal of Computational Physics*, 228(7):2629–2642, 2009.

Geyer, C. J. Markov chain Monte Carlo lecture notes. 1998. URL http://www.stat.umn.edu/geyer/f05/8931/n1998.pdf.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. URL http://dx.doi.org/.

Jones, G. L., Roberts, G. O., and Rosenthal, J. S. Convergence of conditional Metropolis–Hastings samplers. *Advances in Applied Probability*, 46(2):422445, 2014. doi: 10.1239/aap/1401369701.

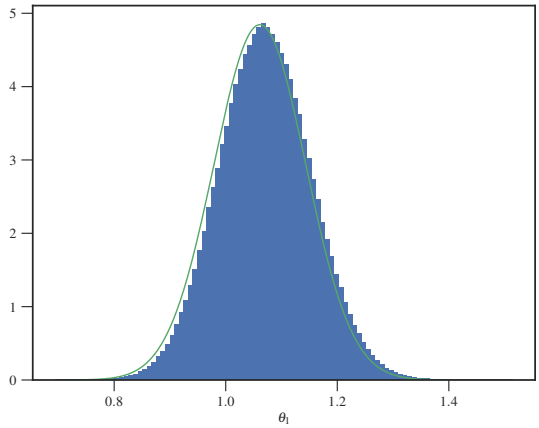
Kleijn, B. and van der Vaart, A. The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381, 2012. doi: 10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675.

Meyn, S. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0521731828, 9780521731829.

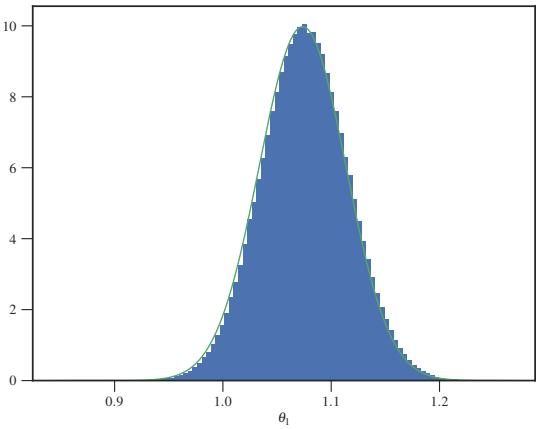
Pollard, D. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer New York, 2012. ISBN 9781461252542. URL https://books.google.co.uk/books?id=g5DbBwAAQBAJ.

Roberts, G. and Rosenthal, J. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997. doi: 10.1214/ECP.v2-981. URL https://doi.org/10.1214/ECP.v2-981.

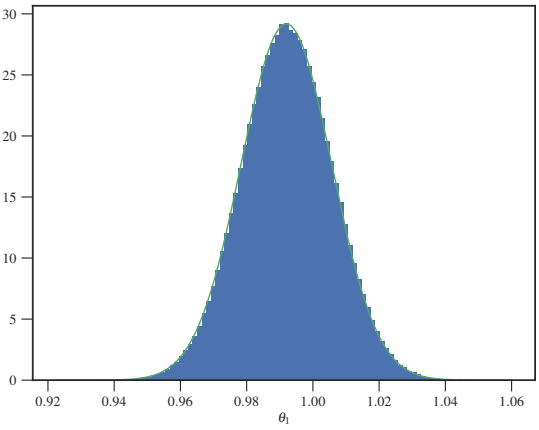
van der Vaart, A. W. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.



(a) $n = 2048$



(b) $n = 8192$



(c) $n = 65536$

Figure 1. Histogram of samples of first regression coefficient (θ_1) versus marginal of Gaussian approximation (green lines).

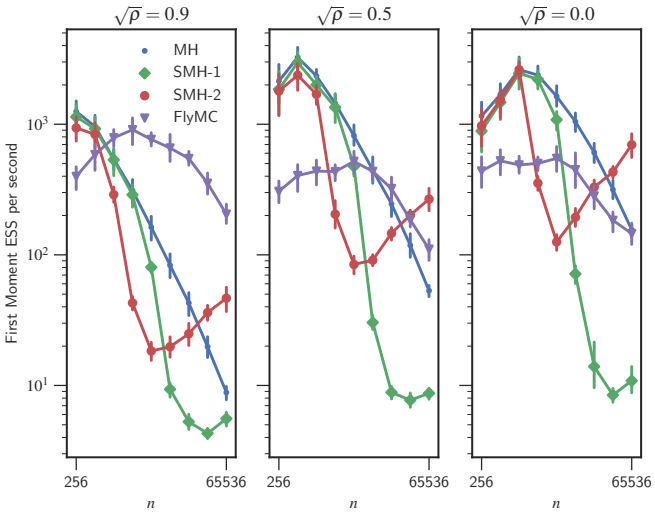


Figure 2. ESS of first regression coefficient for a logistic regression model of dimension 20.

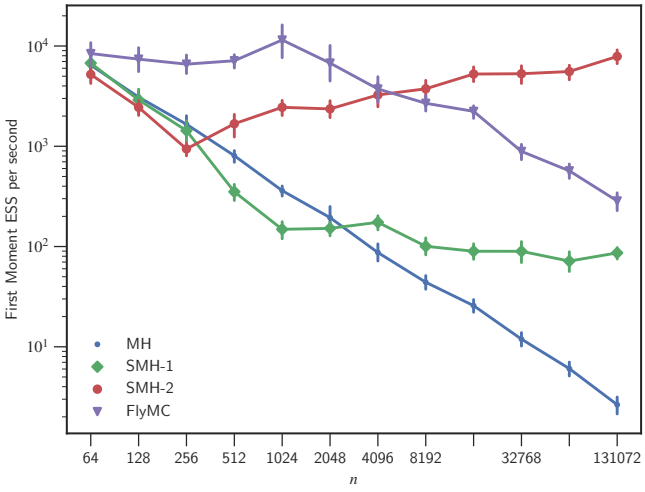


Figure 3. ESS for first regression coefficient of a robust linear regression posterior, scaled by execution time (higher is better).

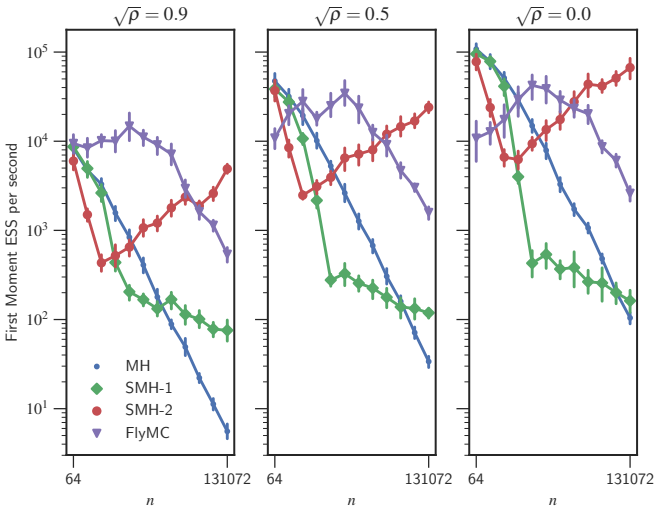


Figure 4. Effect of ρ on ESS for first regression coefficient of the robust linear regression model, scaled by execution time (higher is better).

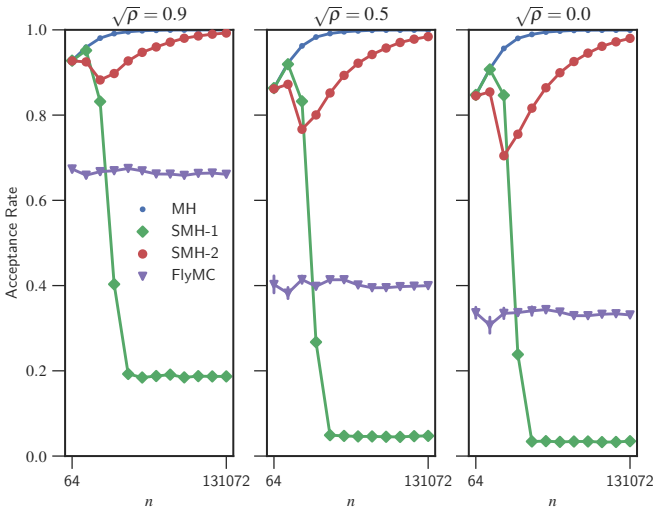


Figure 5. Acceptance rates for pCN proposals for the robust linear regression model.

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	Scalable Metropolis-Hastings for Exact Bayesian Inference with Large Datasets
Publication Status	☑Published ☐ Accepted for Publication ☐Submitted for Publication ☐Unpublished and unsubmitted work written in a manuscript style
Publication Details	Cornish, R., Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., Doucet, A. (2019). Scalable Metropolis-Hastings for Exact Bayesian Inference with Large Datasets. In *Proceedings of the 36th International Conference on Machine Learning,* Long Beach, California, PMLR 97, 2019.

Student Confirmation

Student Name:	Rob Cornish		
Contribution to the Paper	Developed the idea suggested by my supervisors. Performed initial experiments to obtain a proof of concept, while simultaneously uncovering ergodicity issues with a naive approach. Resolved the ergodicty issues via truncation method. Formalised the framework for the method, and formulated and proved all theorems. Wrote the initial draft of the paper.		
Signature	Robert Cornish	Date	17/01/2020

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: Prof. Arnaud Doucet		
Supervisor comments	I do certify that the candidate made a substantial contribution to the publication, and the description above is accurate	
Signature	Date	17/01/2020

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

3
On Nested Monte Carlo

On Nesting Monte Carlo Estimators

Tom Rainforth [1] **Robert Cornish** [1 2] **Hongseok Yang** [3] **Andrew Warrington** [2] **Frank Wood** [4]

Abstract

Many problems in machine learning and statistics involve nested expectations and thus do not permit conventional Monte Carlo (MC) estimation. For such problems, one must nest estimators, such that terms in an outer estimator themselves involve calculation of a separate, nested, estimation. We investigate the statistical implications of nesting MC estimators, including cases of multiple levels of nesting, and establish the conditions under which they converge. We derive corresponding rates of convergence and provide empirical evidence that these rates are observed in practice. We further establish a number of pitfalls that can arise from naïve nesting of MC estimators, provide guidelines about how these can be avoided, and lay out novel methods for reformulating certain classes of nested expectation problems into single expectations, leading to improved convergence rates. We demonstrate the applicability of our work by using our results to develop a new estimator for discrete Bayesian experimental design problems and derive error bounds for a class of variational objectives.

1 Introduction

Monte Carlo (MC) methods are used throughout the quantitative sciences. For example, they have become a ubiquitous means of carrying out approximate Bayesian inference (Doucet et al., 2001; Gilks et al., 1995). The convergence of MC estimation has been considered extensively in the literature (Durrett, 2010). However, the implications arising from the *nesting* of MC schemes, where terms in the integrand depend on the result of separate, nested, MC estimators, is generally less well known. This paper examines the convergence of such nested Monte Carlo (NMC) methods.

Nested expectations occur in wide variety of problems from portfolio risk management (Gordy and Juneja, 2010) to stochastic control (Belomestny et al., 2010). In particular, simulations of agents that reason about other agents often include nested expectations. Tackling such problems requires some form of nested estimation scheme like NMC.

A common class of nested expectations is doubly-intractable inference problems (Murray et al., 2006; Liang, 2010), where the likelihood is only known up to a parameter-dependent normalizing constant. This can occur, for example, when nesting probabilistic programs (Mantadelis and Janssens, 2011; Le et al., 2016). Some problems are even multiply-intractable, such that they require multiple levels of nesting to encode (Stuhlmüller and Goodman, 2014). Our results can be used to show that changes are required to the approaches currently employed by probabilistic programming systems to ensure consistent estimation for such problems (Rainforth, 2017; 2018).

The expected information gain used in Bayesian experimental design (Chaloner and Verdinelli, 1995) requires the calculation of an entropy of a marginal distribution and therefore the expectation of the logarithm of an expectation. By extension, any Kullback-Leibler divergence where one of the terms is a marginal distribution also involves a nested expectation. Hence, our results have important implications for relaxing mean-field assumptions, or using different bounds, in variational inference (Hoffman and Blei, 2015; Naesseth et al., 2017; Maddison et al., 2017) and deep generative models (Burda et al., 2015; Le et al., 2018).

Certain nested estimation problems can be tackled by pseudo-marginal methods (Beaumont, 2003; Andrieu and Roberts, 2009; Andrieu et al., 2010). These consider inference problems where the likelihood is intractable, but can be estimated unbiasedly. From a theoretical perspective, they reformulate the problem in an extended space with auxiliary variables that are used to represent the stochasticity in the likelihood computation, enabling the problem to be expressed as a single expectation.

Our work goes beyond this by considering cases in which a non-linear mapping is applied to the output of the inner expectation, (e.g. the logarithm in the experimental design example), prohibiting such reformulation. We demonstrate that the construction of consistent NMC algorithms is possible, establish convergence rates, and provide empirical evi-

[1]Department of Statistics, University of Oxford [2]Department of Engineering, University of Oxford [3]School of Computing, KAIST [4]Department of Computer Science, University of British Columbia. Correspondence to: Tom Rainforth <rainforth@stats.ox.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

dence that these rates are observed in practice. Our results show that whenever an outer estimator depends non-linearly on an inner estimator, then the number of samples used in *both* the inner and outer estimators must, in general, be driven to infinity for convergence. We extend our results to cases of repeated nesting and show that the optimal NMC convergence rate is $O(1/T^{\frac{2}{D+2}})$ where T is the total number of samples used in the estimator and D is the nesting depth (with $D = 0$ being conventional MC), whereas naïve approaches only achieve a rate of $O(1/T^{\frac{1}{D+1}})$. We further lay out methods for reformulating certain classes of nested expectation problems into a single expectation, allowing usage of conventional MC estimation schemes with superior convergence rates than naïve NMC. Finally, we use our results to make application-specific advancements in Bayesian experimental design and variational auto-encoders.

1.1 Related Work

Though the convergence of NMC has previously received little attention within the machine learning literature, a number of special cases having been investigated in other fields, sometimes under the name of *nested simulation* (Longstaff and Schwartz, 2001; Belomestny et al., 2010; Gordy and Juneja, 2010; Broadie et al., 2011). While most of this literature focuses on particular application-specific non-linear mappings, a convergence bound for a wider range of problems was shown by Hong and Juneja (2009) and recently revisited in the context of rare-event problems by Fort et al. (2017). The latter paper further considers the case where samples in the outer estimator originate from a Markov chain. Compared to this previous work, ours is the first to consider multiple levels of nesting, applies to a wider range of non-linear mappings, and provides more precise convergence rates. By introducing new results, outlining special cases, providing empirical assessment, and examining specific applications, we provide a unified investigation and practical guide nesting MC estimators in a machine learning context. We begin to realize the potential significance of this by using our theoretical results to make advancements in a number of specific application areas.

Another body of literature related to our work is in the study of the convergence of Markov chains with approximate transition kernels (Rudolf and Schweizer, 2015; Alquier et al., 2016; Medina-Aguayo et al., 2016). The analysis in this work is distinct, but complementary, to our own, focusing on the impact of a known bias on an MCMC chain, whereas our focus is more on quantifying this bias. Also related is the study of techniques for variance reduction, such as multilevel MC (Heinrich, 2001; Giles, 2008), and bias reduction, such as the multi-step Richardson-Romberg method (Pagés, 2007; Lemaire et al., 2017) and Russian roulette sampling (Lyne et al., 2015), many of which are applicable in a NMC context and can improve performance.

2 Problem Formulation

The key idea of MC is that the expectation of an arbitrary function $\lambda\colon \mathcal{Y} \to \mathcal{F} \subseteq \mathbb{R}$ under a probability distribution $p(y)$ for its input $y \in \mathcal{Y}$ can be approximated using:

$$I = \mathbb{E}_{y\sim p(y)}\left[\lambda(y)\right] \tag{1}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\lambda(y_n) \quad \text{where} \quad y_n \overset{i.i.d.}{\sim} p(y). \tag{2}$$

In this paper, we consider the case that λ is itself intractable, defined only in terms of a functional mapping of an expectation. Specifically, $\lambda(y) = f(y, \gamma(y))$ where we can evaluate $f\colon \mathcal{Y} \times \Phi \to \mathcal{F}$ exactly for a given y and $\gamma(y)$, but $\gamma(y)$ is the output of the following intractable expectation of another variable $z \in \mathcal{Z}$:

$$\text{either} \quad \gamma(y) = \mathbb{E}_{z\sim p(z|y)}\left[\phi(y,z)\right] \tag{3a}$$

$$\text{or} \quad \gamma(y) = \mathbb{E}_{z\sim p(z)}\left[\phi(y,z)\right] \tag{3b}$$

depending on the problem, with $\phi\colon \mathcal{Y} \times \mathcal{Z} \to \Phi$. All our results apply to both cases, but we will focus on (3a) for clarity. Estimating I involves computing an integral over z for each value of y in the outer integral. We refer to the approach of tackling both integrations using MC as *nested Monte Carlo* (NMC):

$$I = \mathbb{E}\left[f(y,\gamma(y))\right] \approx I_{N,M} = \frac{1}{N}\sum_{n=1}^{N} f(y_n, (\hat{\gamma}_M)_n) \tag{4a}$$

where $y_n \overset{i.i.d.}{\sim} p(y)$ and

$$(\hat{\gamma}_M)_n = \frac{1}{M}\sum_{m=1}^{M}\phi(y_n, z_{n,m}) \tag{4b}$$

where each $z_{n,m} \sim p(z|y_n)$ are independently sampled. In Section 3 we will build on this further by considering cases with multiple levels of nesting, where calculating $\phi(y,z)$ involves computation of an intractable (nested) expectation.

3 Convergence of Nested Monte Carlo

We now show that approximating $I \approx I_{N,M}$ is in principle possible, at least when f is well-behaved. In particular, we establish a convergence rate of the mean squared error of $I_{N,M}$ and prove a form of almost sure convergence to I. We further generalize our convergence rate to apply to the case of multiple levels of estimator nesting.

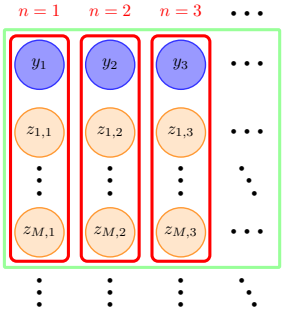


Figure 1. Informal convergence representation

Before providing a formal examination of the convergence of NMC, we first provide intuition about how we might expect to construct a convergent

NMC estimator. Consider the diagram shown in Figure 1, and suppose that we want our error to be less than some arbitrary ε. Assume that f is sufficiently smooth that we can choose M large enough to make $|I - \mathbb{E}[f(y_n, (\hat{\gamma}_M)_n)]| < \varepsilon$ (we will characterize the exact requirements for this later). For this fixed M, we have a standard MC estimator on an extended space $y, z_1, \ldots, z_M$ such that each sample constitutes one of the red boxes. As we take $N \to \infty$, i.e. taking all the samples in the green box, this estimator converges such that $I_{N,M} \to \mathbb{E}[f(y_n, (\hat{\gamma}_M)_n)]$ as $N \to \infty$ for fixed M. As we can make ε arbitrarily small, we can also achieve an arbitrarily small error.

More formally, convergence bounds for NMC have previously been shown by Hong and Juneja (2009). Under the assumptions that each $(\hat{\gamma}_M)_n$ is Gaussian distributed (which is often reasonable due to the central limit theorem) and that f is thrice differentiable other than at some finite number of points, they show that it is possible to achieve a converge rate of $O(1/N + 1/M^2)$. We now show that these assumptions can be relaxed to only requiring f to be Lipschitz continuous, at the expense of weakening the bound.

Theorem 1. *If f is Lipschitz continuous and $f(y_n, \gamma(y_n)), \phi(y_n, z_{n,m}) \in L^2$, the mean squared error of $I_{N,M}$ converges to 0 at rate $O(1/N + 1/M)$.*

Proof. The theorem follows as a special case of Theorem 3. For exposition, a more accessible proof for this particular result is also provided in Appendix A in the supplement. □

Inspection of the convergence rate above shows that, given a total number of samples $T = MN$, our bound is tightest when $N \propto M$, with a corresponding rate $O(1/\sqrt{T})$ (see Appendix G). When the additional assumptions of Hong and Juneja (2009) apply, this rate can be lowered to $O(1/T^{2/3})$ by setting $N \propto M^2$. We will later show that this faster convergence rate can be achieved whenever f is twice continuously differentiable, see also (Fort et al., 2017).

These convergence rates suggest that, for most f, it is necessary to increase not only the total number of samples, T, but also the number of samples used for each evaluation of the inner estimator, M, to achieve convergence. Further, as we show in Appendix B, the estimates produced by NMC are, in general, biased. This is perhaps easiest to see by noting that as $N \to \infty$, the variance of the estimator must tend to zero by the law of large numbers, but our bounds remain non-zero for any finite M, implying a bias.

3.1 Minimum Continuity Requirements

We next consider the question of what is the minimal requirement on f to ensures some form of convergence? For a given y_1, we have that $(\hat{\gamma}_M)_1 = \frac{1}{M}\sum_{m=1}^{M} \phi(y_1, z_{1,m}) \to \gamma(y_1)$ almost surely as $M \to \infty$, because the left-hand side is a MC estimator. If f is continuous around y_1, this also implies $f(y_1, (\hat{\gamma}_M)_1) \to f(y_1, \gamma(y_1))$. Our candidate requirement is that this holds in expectation, i.e. that it holds when we incorporate the effect of the outer estimator. More precisely, we define $(\epsilon_M)_n = |f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))|$ and require that $\mathbb{E}[(\epsilon_M)_1] \to 0$ as $M \to \infty$ (noting that $(\epsilon_M)_n$ are i.i.d. and so $\mathbb{E}[(\epsilon_M)_1] = \mathbb{E}[(\epsilon_M)_n], \forall n \in \mathbb{N}$). Informally, this "expected continuity" requirement is weaker than uniform continuity (and much weaker than Lipschitz continuity) as it allows (potentially infinitely many) discontinuities in f. More formally we have the following result.

Theorem 2. *For $n \in \mathbb{N}$, let*

$$(\epsilon_M)_n = |f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))|.$$

Assume that $\mathbb{E}[(\epsilon_M)_1] \to 0$ as $M \to \infty$. Let Ω be the sample space of our underlying probability space, so that $I_{\tau_\delta(M),M}$ forms a mapping from Ω to $\mathbb{R}$. Then, for every $\delta > 0$, there exists a measurable $A_\delta \subseteq \Omega$ with $\mathbb{P}(A_\delta) < \delta$, and a function $\tau_\delta : \mathbb{N} \to \mathbb{N}$ such that, for all $\omega \notin A_\delta$,

$$I_{\tau_\delta(M),M}(\omega) \overset{a.s.}{\to} I \quad as \quad M \to \infty.$$

Proof. See Appendix C. □

As well as providing proof of a different form of convergence to any existing results, this result is particularly important because many, if not most, functions are not Lipschitz continuous due to the their behavior in the limits. For example, even the function $f(y, \gamma(y)) = (\gamma(y))^2$ is not Lipschitz continuous because the derivative is unbounded as $|\gamma(y)| \to \infty$, whereas the vast majority of problems will satisfy our weaker requirement of $\mathbb{E}[(\epsilon_M)_1] \to 0$.

3.2 Repeated Nesting and Exact Bounds

We next consider the case of multiple levels of nesting. As previously explained, this case is particularly important for analyzing probabilistic programming languages. To formalize what we mean by arbitrary nesting, we first assume some fixed integral depth $D > 0$, and real-valued functions $f_0, \cdots, f_D$. We then define

$$\gamma_D\left(y^{(0:D-1)}\right) = \mathbb{E}\left[f_D\left(y^{(0:D)}\right)\middle| y^{(0:D-1)}\right] \quad \text{and}$$

$$\gamma_k(y^{(0:k-1)}) = \mathbb{E}\left[f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)\middle| y^{(0:k-1)}\right],$$

for $0 \leq k < D$, where $y^{(k)} \sim p\left(y^{(k)}|y^{(0:k-1)}\right)$. Note that our single nested case corresponds to the setting of $D = 1$, $f_0 = f$, $f_1 = \phi$, $y^{(0)} = y$, $y^{(1)} = z$, $\gamma_0 = I$, and $\gamma_1 = \gamma$. Our goal is to estimate $\gamma_0 = \mathbb{E}\left[f_0\left(y^{(0)}, \gamma_1\left(y^{(0)}\right)\right)\right]$. To do so we will use the following NMC scheme:

$$I_D\left(y^{(0:D-1)}\right) = \frac{1}{N_D}\sum_{n=1}^{N_D} f_D\left(y^{(0:D-1)}, y_n^{(D)}\right) \quad \text{and}$$

$$I_k\left(y^{(0:k-1)}\right) = \frac{1}{N_k}\sum_{n=1}^{N_k} f_k\left(y^{(0:k-1)}, y_n^{(k)}, I_{k+1}\left(y^{(0:k-1)}, y_n^{(k)}\right)\right)$$

for $0 \leq k \leq D-1$, where each $y_n^{(k)} \sim p\left(y^{(k)}|y^{(0:k-1)}\right)$ is drawn independently. Note that there are multiple values of $y_n^{(k)}$ for each possible $y^{(0:k-1)}$ and that $I_k\left(y^{(0:k-1)}\right)$ is still a random variable given $y^{(0:k-1)}$.

We are now ready to provide our general result for the convergence bounds that applies to cases of repeated nesting, provides constant factors (rather than just using big O notation), and shows how the bound can be improved if the additional assumption of continuous differentiability holds.

Theorem 3. *If $f_0, \cdots, f_D$ are all differentiable and Lipschitz continuous in their second input with Lipschitz constants*

$$K_k := \sup_{y^{(0:k)}} \left| \frac{\partial f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}} \right| < \infty,$$

for all $k \in 0, \ldots, D-1$ and if

$$\varsigma_k^2 := \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right] < \infty \quad \forall k \in 0, \ldots, D$$

then

$$\mathbb{E}\left[(I_0 - \gamma_0)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \sum_{k=1}^{D}\left(\prod_{\ell=0}^{k-1} K_\ell^2\right)\frac{\varsigma_k^2}{N_k} + O(\epsilon) \quad (5)$$

where $O(\epsilon)$ represents asymptotically dominated terms.

If $f_0, \cdots, f_D$ are twice continuously differentiable with second derivative bounds

$$C_k := \sup_{y^{(0:k)}} \left| \frac{\partial^2 f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^2} \right| < \infty$$

then this mean square error bound can be tightened to

$$\mathbb{E}\left[(I_0 - \gamma_0)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \left(\frac{C_0\varsigma_1^2}{2N_1} + \sum_{k=0}^{D-2}\left(\prod_{d=0}^{k} K_d\right)\frac{C_{k+1}\varsigma_{k+2}^2}{2N_{k+2}}\right)^2 + O(\epsilon). \quad (6)$$

For a single nesting, we can further characterize $O(\epsilon)$ giving

$$\mathbb{E}\left[(I_0 - \gamma_0)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \frac{4K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_0\varsigma_1}{N_0\sqrt{N_1}} + \frac{K_0^2\varsigma_1^2}{N_1} \quad (7)$$

$$\mathbb{E}\left[(I_0 - \gamma_0)^2\right] \leq \frac{\varsigma_0^2}{N_0} + \frac{C_0^2\varsigma_1^4}{4N_1^2}\left(1 + \frac{1}{N_0}\right) + \frac{K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_1}{N_0\sqrt{N_1}}\sqrt{\varsigma_0^2 + \frac{C_0^2\varsigma_1^4}{4N_1^2}} + O\left(\frac{1}{N_1^3}\right) \quad (8)$$

for when the continuous differentiability assumption does not hold and holds respectively.

Proof. See Appendix D. □

These results give a convergence rate of $O(\sum_{k=0}^{D} 1/N_k)$ when only Lipschitz continuity holds and $O(1/N_0 + (\sum_{k=1}^{D} 1/N_k)^2)$ when all the f_k are also twice continuously differentiable. As estimation requires drawing $O(T)$ samples where $T = \prod_{k=0}^{D} N_k$, the convergence rate will rapidly diminish with repeated nesting. More precisely, as shown in Appendix G, the optimal convergence rates are $O(1/T^{\frac{1}{D+1}})$ and $O(1/T^{\frac{2}{D+2}})$ respectively for the two cases, both of which imply that the rate diminishes exponentially with D.

4 Special Cases

We now outline some special cases where it is possible to achieve a convergence rate of $O(1/N)$ in the mean square error (MSE) as per conventional MC estimation. Establishing these cases is important because it identifies for which problems we can use conventional results, when we can achieve an improved convergence rate, and what precautions we must take to ensure this. We will focus on single nesting instances, but note that all results still apply to repeated nesting scenarios because they can be used to "collapse" layers and thereby reduce the depth of the nesting.

4.1 Linear f

Our first special case is that f is linear in its second argument, i.e. $f(y, \alpha v + \beta w) = \alpha f(y, v) + \beta f(y, w)$. Here the problem can be rearranged to a single expectation, a well-known result which forms the basis for pseudo-marginal, nested sequential MC (Naesseth et al., 2015), and certain ABC methods (Csilléry et al., 2010). Namely we have

$$\begin{aligned} I &= \mathbb{E}_{y\sim p(y)}\left[f\left(y, \mathbb{E}_{z\sim p(z|y)}\left[\phi(y,z)\right]\right)\right] \\ &= \mathbb{E}_{y\sim p(y)}\left[\mathbb{E}_{z\sim p(z|y)}\left[f(y, \phi(y,z))\right]\right] \\ &\approx \frac{1}{N}\sum_{n=1}^{N} f(y_n, \phi(y_n, z_n)) \end{aligned} \quad (9)$$

where $(y_n, z_n) \sim p(y)p(z|y)$ if $\gamma(y)$ is of the form of (3a) and $y_n \sim p(y)$ and $z_n \sim p(z)$ are independently drawn if $\gamma(y)$ is of the form of (3b).

4.2 Finite Possible Realizations of y

Our second case is if y must take one of finitely many values $y_1, \cdots, y_C$, then it is possible to use another approach to ensure the same convergence rate as standard MC. The key observation is to note that in this case we can convert the nested problem (2) into C separate non-nested problems

$$I = \sum_{c=1}^{C} P(y = y_c)\, f(y_c, \gamma(y_c)) \quad (10)$$

which can then be estimated using

$$I_N = \sum_{c=1}^{C} (\hat{P}_N)_c\, (\hat{f}_N)_c \quad \text{where} \quad (11)$$

$$P(y = y_c) \approx (\hat{P}_N)_c = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(y_n = y_c) \quad (12)$$

$$f(y_c, \gamma(y_c)) \approx (\hat{f}_N)_c = f\left(y_c, \frac{1}{N}\sum_{n=1}^{N}\phi(y_c, z_{n,c})\right) \quad (13)$$

with $y_n \overset{i.i.d.}{\sim} p(y)$ and $z_{n,c} \sim p(z|y_c)$ (or $z_{n,c} \sim p(z)$ if using the formulation in (3b)). Note the critical point that each $z_{n,c}$ is independent of y_n as each y_c is a constant. We can now show the following result which, though intuitively straightforward, requires care to formally prove.

Theorem 4. *If f is Lipschitz continuous, then the mean squared error of $I_N = \sum_{c=1}^{C}(\hat{P}_N)_c\,(\hat{f}_N)_c$ as an estimator for I as per* (10) *converges at rate $O(1/N)$.*

Proof. See Appendix E. □

4.3 Products of Expectations

We next consider the scenario, which occurs for many latent variables models and probabilistic programming problems, where $\gamma(y)$ is equal to the product of multiple expectations, rather than just a single expectation as per (3a). That is,

$$I = \mathbb{E}_{y\sim p(y)}\left[f\left(y, \prod_{\ell=1}^{L}\mathbb{E}_{z_\ell\sim p(z_\ell|y)}\left[\psi_\ell(y, z_\ell)\right]\right)\right]. \quad (14)$$

Because the z_ℓ will not in general be independent, we cannot trivially rearrange (14) to a standard nested estimation by moving the product within the expectation. Our insight is that the required rearrangement can instead be achieved by introducing new random variables $\{z'_\ell\}_{\ell=1:L}$ such that each $z'_\ell|y \sim p(z_\ell|y)$ and the z'_ℓ are independent of one another. This can be achieved by, for example, taking L independent samples from the joint $Z_\ell \overset{i.i.d.}{\sim} p(z_{1:L}|y)$ and using the ℓ^{th} such draw for the ℓ^{th} dimension of z', i.e. setting $z'_\ell = \{Z_\ell\}_\ell$. For every $y \in \mathcal{Y}$ we now have

$$\begin{aligned}\prod_{\ell=1}^{L}\mathbb{E}_{z_\ell\sim p(z_\ell|y)}[\psi_\ell(y, z_\ell)] &= \prod_{\ell=1}^{L}\mathbb{E}_{z'_\ell\sim p(z'_\ell|y)}[\psi_\ell(y, z'_\ell)] \\ &= \mathbb{E}_{\{z'_\ell\}_{\ell=1:L}\sim p(\{z'_\ell\}_{\ell=1:L}|y)}\left[\prod_{\ell=1}^{L}\psi_\ell(y, z'_\ell)\right] \quad (15)\end{aligned}$$

which is a single expectation on an extended space and shows that (14) fits the NMC formulation. Furthermore, we can now show that if f is linear, the MSE of the NMC estimator (14) converges at the standard MC rate $O(1/N)$, provided that M remains fixed.

Theorem 5. *Consider the NMC estimator*

$$I_N = \frac{1}{N}\sum_{n=1}^{N} f\left(y_n, \prod_{\ell=1}^{L}\frac{1}{M_\ell}\sum_{m=1}^{M_\ell}\psi_\ell(y_n, z'_{n,\ell,m})\right)$$

where each $y_n \in \mathcal{Y}$ and $z'_{n,\ell,m} \in \mathcal{Z}_\ell$ are independently drawn from $y_n \sim p(y)$ and $z'_{n,\ell,m}|y_n \sim p(z_\ell|y_n)$, respectively. If f is linear, the estimator converges almost surely to I, with a convergence rate of $O(1/N)$ in the mean square error for any fixed choice of $\{M_\ell\}_{\ell=1:L}$.

Proof. See Appendix F. □

As this result holds in the case $L = 1$, an important consequence is that whenever f is linear, the same convergence rate is achieved regardless of whether we reformulate the problem to a single expectation or not, provided that the number of samples used by the inner estimator is fixed.

4.4 Polynomial f

Perhaps surprisingly, whenever f is of the form

$$f(y, \gamma(y)) = g(y)\,\gamma(y)^\alpha \quad (16)$$

where $\alpha \in \mathbb{Z}_{\geq 0}$, then it is also possible to construct a standard MC estimator by building on the ideas introduced in Section 4.3 and those of (Goda, 2016). The key idea is

$$\left(\mathbb{E}\left[z\right]\right)^2 = \mathbb{E}\left[z\right]\mathbb{E}\left[z'\right] = \mathbb{E}\left[zz'\right] \quad (17)$$

where z and z' are i.i.d. Therefore, assuming appropriate integrability requirements, we can construct the following non-nested MC estimator:

$$\begin{aligned}\mathbb{E}\left[g(y)\,\gamma(y)^\alpha\right] &= \mathbb{E}\left[g(y)\prod_{\ell=1}^{\alpha}\mathbb{E}_{z_\ell\sim p(z|y)}\left[\phi(y, z_\ell)|y\right]\right] \\ &= \mathbb{E}\left[g(y)\prod_{\ell=1}^{\alpha}\phi(y, z_\ell)\right] \approx \frac{1}{N}\sum_{n=1}^{N} g(y_n)\prod_{\ell=1}^{\alpha}\phi(y_n, z_{n,\ell})\end{aligned}$$

where we independently draw each $z_{n,\ell}|y_n \sim p(z|y_n)$.

5 Empirical Verification

The convergence rates proven in Section 3 are only *upper bounds* on the worst-case performance. We will now examine whether these convergence rates are tight in practice, investigate what happens when our guidelines are not followed, and outline some applications of our results.

5.1 Simple Analytic Model

We start with the following analytically calculable problem

$$y \sim \text{Uniform}(-1, 1), \quad (18a)$$

$$z \sim \mathcal{N}(0, 1), \quad (18b)$$

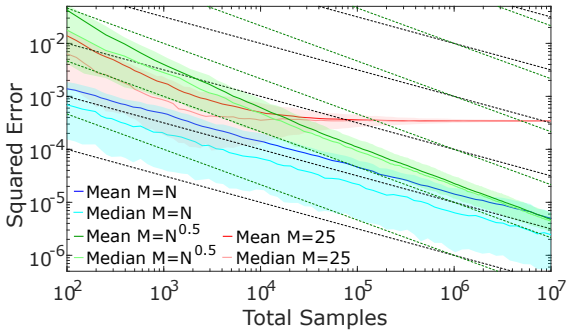
$$\phi(y, z) = \sqrt{2/\pi}\exp\left(-2(y - z)^2\right), \quad (18c)$$

$$f(y, \gamma(y)) = \log(\gamma(y)) = \log(\mathbb{E}_z[\phi(y, z)]). \quad (18d)$$

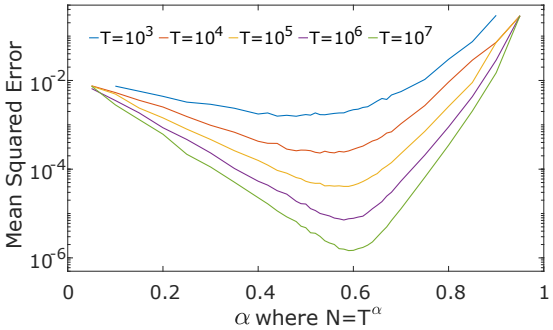
for which $I = \frac{1}{2}\log\left(\frac{2}{5\pi}\right) - \frac{2}{15}$. Figure 2a shows the corresponding empirical convergence obtained by applying (4) to (18) directly. It shows that, for this problem, the theoretical convergence rates from Theorem 3 are indeed realized. The figure also demonstrates the danger of not increasing M with N, showing that the NMC estimator converges to an incorrect solution when M is held constant. Figure 2b shows the effect of varying N and M for various fixed sample budgets T and demonstrates that the asymptotically optimal strategy can be suboptimal for finite budgets.

5.2 Planning Cancer Treatment

We now introduce a real-world example to show the applicability of NMC in a scenario where the solution is not



(a) Convergence with increasing sample budget.



(b) Final error for different T and N.

Figure 2. Empirical convergence of NMC for (18). [Left] convergence in total samples for different ways of setting M and N. Results are averaged over 1000 independent runs, while shaded regions give the 25%-75% quantiles. The theoretical convergence rates, namely $O(1/\sqrt{T})$ and $O(1/T^{2/3})$ for setting $N \propto M$ and $N \propto M^2$ respectively, are observed (see the dashed black and green lines respectively for reference). The fixed M case converges at the standard MC error rate of $O(1/T)$ but to a biased solution. [Right] final error for different total sample budgets as a function of α where $N = T^\alpha$ and $M = T^{1-\alpha}$ iterations are used for the outer and inner estimators respectively. This shows that even though $\alpha = \frac{2}{3}$ is the asymptotically optimal allocation strategy, this is not the optimal solution for finite T. Nonetheless, as T increases, the optimum value of α increases, starting around 0.5 for $T = 10^3$ and reaching around 0.6 for $T = 10^7$.

analytically tractable and conventional MC is insufficient. Consider a treatment center assessing a new policy for planning cancer treatments, subject to a budget. Clinicians must decide on a patient-by-patient basis whether to administer chemotherapy in the hope that their tumor will reduce in size sufficiently to be able to perform surgery at a later date. A treatment is considered to have been successful if the size of the tumor drops below a threshold value in a fixed time window. The clinicians have at their disposal a simulator for the evolution of tumors with time, parameterized by both observable values, y, such as tumor size, and unobservable values, z, such as the patient-specific response to treatment. Given a set of input parameters, the simulator deterministically returns a binary response $\phi(y, z) \in \{0, 1\}$, with 1 indicating a successful treatment. To estimate the probability of a successful treatment for a given patient, the clinician must calculate the expected success over these unobserved variables, namely $\mathbb{E}_{z\sim p(z|y)}[\phi(y, z)]$ where $p(z|y)$ represents a probabilistic model for the unobserved variables, which could, for example, be constructed based on empirical data. The clinician then decides whether to go ahead with the treatment for that patient based on whether the calculated probability of success exceeds a certain threshold T_{treat}.

The treatment center wishes to estimate the expected number of patients that will be treated for a given T_{treat} so that it can minimize this threshold without exceeding its budget. To do this, it calculates the expectation of the clinician's decisions to administer treatment, giving the complete nested expectation for calculating the number of treated patients as

$$I(T_{\text{treat}}) = \mathbb{E}\left[\mathbb{I}\left(\mathbb{E}_{z\sim p(z|y)}[\phi(y, z)] > T_{\text{treat}}\right)\right], \quad (19)$$

where the step function $\mathbb{I}(\cdot > T_{\text{treat}})$ imposes a non-linear mapping, preventing conventional MC estimation. Full details on ϕ, $p(y)$, and $p(z|y)$ are given in Appendix H.

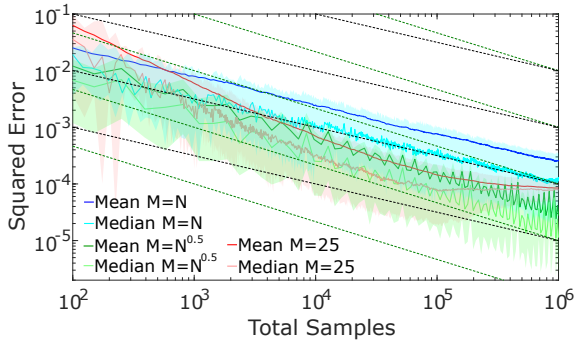


Figure 3. Convergence of NMC for cancer simulation. A ground truth estimate was calculated using a single run with $M = 10^5$ and $N = 10^5$. Experimental setup and conventions are as per Figure 2a and we again observe the expected convergence rates. When $M = \sqrt{N}$ an interesting fluctuation behavior is observed. Further testing suggests that this originates because the bias of the estimator depends in a fluctuating manner on the value of M as the binary output of $\phi(y, z)$ creates a quantization effect on the possible estimates for $\hat{\gamma}$. This effect is also observed for the $M = N$ case but is less pronounced.

To verify the convergence rate, we repeated the analysis from Section 5.1 for (19) at a fixed value of $T_{\text{treat}} = 0.35$. The results, shown in Figure 3, again verify the theoretical rates. By further testing different values of T_{treat}, we found $T_{\text{treat}} = 0.125$ to be optimal under the budget.

5.3 Repeated Nesting

We next consider some simple models with multiple levels of nesting, starting with

$$y^{(0)} \sim \text{Uniform}(0, 1), \quad y^{(1)} \sim \mathcal{N}(0, 1), \quad y^{(2)} \sim \mathcal{N}(0, 1),$$
$$f_0\left(y^{(0)}, \gamma_1\left(y^{(0)}\right)\right) = \log \gamma_1\left(y^{(0)}\right) \quad (20a)$$

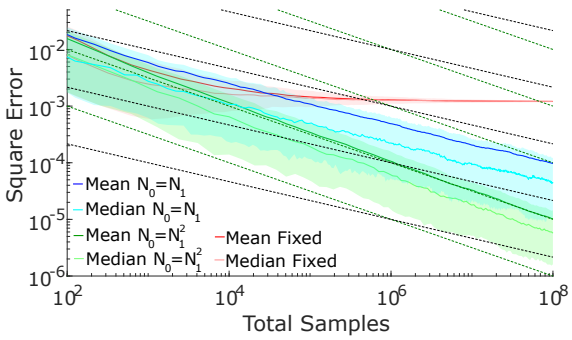


Figure 4. Empirical convergence of NMC to (20) for an increasing total sample budget $T = N_0 N_1 N_2$. Setup and conventions as per Figure 2a. Shown in red is the convergence with a fixed $N_2 = 5$ and $N_0 = N_1^2$, which we see gives a biased solution. Shown in blue is the convergence when setting $N_0 = N_1 = N_2$, which we see converges at the expected $O(T^{-1/3})$ rate. Shown in green is the convergence when setting $N_0 = N_1^2 = N_2^2$ which we see again gives the theoretical convergence rate, namely $O(T^{-1/2})$.

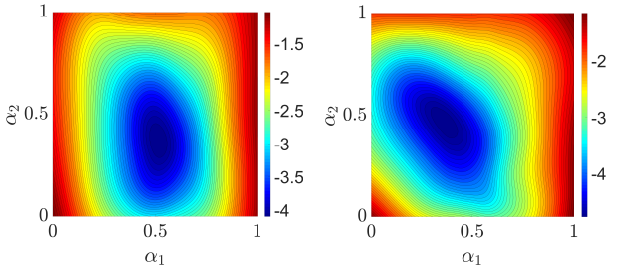
$$f_1\left(y^{(0:1)}, \gamma_2\left(y^{(0:1)}\right)\right) = \exp\left(-\frac{1}{2}\left(y^{(0)} - y^{(1)} - \log\gamma_2\left(y^{(0:1)}\right)\right)\right) \tag{20b}$$

$$f_2\left(y^{(0:2)}\right) = \exp\left(y^{(2)} - \frac{y^{(0)} + y^{(1)}}{2}\right) \tag{20c}$$

which has analytic solution $I = -3/32$. The convergence plot shown in Figure 4 demonstrates that the theoretically expected convergence behaviors are observed for different methods of setting N_0, N_1, and N_2.

We further investigated the empirical performance of different strategies for choosing N_0, N_1, N_2 under a finite fixed budget $T = N_0 N_1 N_2$. In particular, we looked to establish the optimal empirical setting under the fixed budget $T = 10^6$ for the model described in (20) and a slight variation where $y^{(0)}$ is replaced with $y^{(0)}/10$, for which the ground truth is now $I = 39/160$. Defining α_1 and α_2 such that $N_0 = T^{\alpha_1}$, $N_1 = T^{\alpha_2(1-\alpha_1)}$, and $N_2 = T^{(1-\alpha_1)(1-\alpha_2)}$, we ran a Bayesian optimization algorithm, namely BOPP (Rainforth et al., 2016), to optimize the log MSE, $\log_{10}\left(\mathbb{E}\left[(I_0(\alpha_1,\alpha_2) - \gamma_0)^2\right]\right)$, with respect to (α_1, α_2). For each tested (α_1, α_2), the MSE was estimated using 1000 independently generated samples of I_0 and we allowed a total of 200 such tests. We found respective optimal values for (α_1, α_2) of $(0.53, 0.36)$ and $(0.38, 0.45)$. By comparison, the asymptotically optimal setup suggested by our theoretical results is $(0.5, 0.5)$, showing that the finite budget optimal allocation can vary significantly from the asymptotically optimal solution and that it does so in a problem dependent manner.

As a byproduct, BOPP also produced Gaussian process approximations to the log MSE variations, as shown in Figure 5. We see that the two problems lead to distinct performance variations. Based on the (unshown) uncertainty estimates of these Gaussian processes, we believe these approximations are a close representation of the truth.



(a) Original (b) Modification

Figure 5. Contour plots of $\log_{10}\left(\mathbb{E}\left[(I_0 - \gamma_0)^2\right]\right)$ produced by BOPP for different allocations of the sample budget $T = 10^6$ for the problem shown in (20) and its modified variant.

6 Applications

6.1 Bayesian Experimental Design

In this section, we show how our results can be used to derive an improved estimator for the problem of Bayesian experimental design (BED) in the case where the experiment outputs are discrete. A summary of our approach is provided here, with full details provided in Appendix I.

Bayesian experimental design provides a framework for designing experiments in a manner that is optimal from an information-theoretic viewpoint (Chaloner and Verdinelli, 1995; Sebastiani and Wynn, 2000). Given a prior $p(\theta)$ on parameters θ and a corresponding likelihood $p(y|\theta, d)$ for experiment outcomes y given a design d, the Bayesian optimal design d^* is given by maximizing the mutual information between θ and y defined as follows

$$\bar{U}(d) = \int_{\mathcal{Y}}\int_{\Theta} p(y,\theta|d)\log\left(\frac{p(\theta|y,d)}{p(\theta)}\right)d\theta dy. \tag{21}$$

Estimating d^* is challenging as $p(\theta|y,d)$ is rarely known in closed-form. However, appropriate algebraic manipulation shows that (21) is consistently estimated by

$$\hat{U}_{\text{NMC}}(d) = \frac{1}{N}\sum_{n=1}^{N}\left[\log(p(y_n|\theta_{n,0},d)) - \log\left(\frac{1}{M}\sum_{m=1}^{M}p(y_n|\theta_{n,m},d)\right)\right] \tag{22}$$

where $\theta_{n,m} \sim p(\theta)$ for each $(m,n) \in \{0,\dots,M\} \times \{1,\dots,N\}$, and $y_n \sim p(y|\theta = \theta_{n,0}, d)$ for each $n \in \{1,\dots,N\}$. This naïve NMC estimator has been implicitly used by (Myung et al., 2013) amongst others and gives a convergence rate of $O(1/N + 1/M^2)$ as per Theorem 3.

When y can only take on finitely many realizations $y_1,\dots,y_c$, we use the ideas introduced in Section 4.2 to

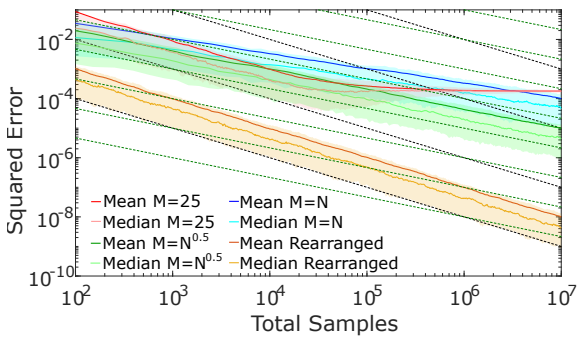


Figure 6. Convergence of NMC (i.e. (22)) and our reformulated estimator (23) for the BED problem. Experimental setup and conventions are as per Figure 2a, with a ground truth estimate made using a single run of the reformulated estimator with 10^{10} samples. We see that the theoretical convergence rates are observed, with the advantages of the reformulated estimator particularly pronounced.

derive the following improved estimator

$$\hat{U}_R(d) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} p(y_c|\theta_n, d)\log\left(p(y_c|\theta_n, d)\right) \qquad (23)$$
$$-\sum_{c=1}^{C}\left[\left(\frac{1}{N}\sum_{n=1}^{N} p(y_c|\theta_n, d)\right)\log\left(\frac{1}{N}\sum_{n=1}^{N} p(y_c|\theta_n, d)\right)\right]$$

where $\theta_n \sim p(\theta), \forall n \in \{1, \ldots, N\}$. As C is fixed, (23) converges at the standard MC error rate of $O(1/N)$. This constitutes a substantially faster convergence as (22) requires a total of MN samples compared to N for (23).

We finish by showing that the theoretical advantages of this reformulation also lead to empirical gains. For this we consider a model used in psychology experiments introduced by (Vincent, 2016), details of which are given in Appendix I. Figure 6 demonstrates that the theoretical convergence rates are observed while results given in Appendix I show that this leads to significant practical gains in estimating $\bar{U}(d)$.

6.2 Variational Autoencoders

To give another example of the applicability of our results, we now use Theorem 3 to directly derive a new result for the importance weighted autoencoder (IWAE) (Burda et al., 2015). Both the IWAE and the standard variational autoencoder (VAE) (Kingma and Welling, 2013) use lower bounds on the model evidence as objectives for train deep generative models and employ estimators of the form

$$I_{N,M} = \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{M}\sum_{m=1}^{M} w_{n,m}(\theta)\right) \qquad (24)$$

for some given θ upon which the random $w_{n,m}(\theta)$ depend. The IWAE sets $N = 1$ and the VAE $M = 1$. We can view (24) as a (biased) NMC estimator for the evidence $\log \mathbb{E}\left[w_{1,1}(\theta)\right]$, which is the target one actually wishes to optimize (for the generative network). We can now assess the MSE of this biased estimator using (8), noting that this is a special case where $\varsigma_0^2 = 0$, giving $\mathbb{E}\left[\left(I_{N,M} - I\right)^2\right] \leq \frac{C_0^2\varsigma_1^4}{4M^2}\left(1 + \frac{1}{N}\right) + \frac{K_0^2\varsigma_1^2}{NM} + \frac{C_0 K_0 \varsigma_1^3}{NM^{3/2}} + O(\frac{1}{M^3})$. For a fixed budget $T = NM$ this becomes $O\left(\frac{1}{M^2} + \frac{1}{T} + \frac{1}{T\sqrt{M}}\right)$. Given T is fixed, we thus see that the higher M is, the lower the error bound. Therefore, the lowest MSE is achieved by setting $N = 1$ and $M = T$, as is done by the IWAE. As we show in Rainforth et al. (2018), these results further carry over to the reparameterized derivative estimates $\nabla_\theta I_{N,M}$.

6.3 Nesting Probabilistic Programs

Probabilistic programming systems (PPSs) (Goodman et al., 2008; Wood et al., 2014) provide a strong motivation for the study of NMC methods because many PPSs allow for arbitrary nesting of models (or queries, as they are known in the PPS literature), such that it is easy to define and run nested inference problems, including cases with multiple layers of nesting (Stuhlmüller and Goodman, 2012; 2014). Though this ability to nest queries has started to be exploited in application-specific work (Ouyang et al., 2016; Le et al., 2016), the resulting nested inference problems fall outside the scope of conventional convergence proofs and so the statistical validity of the underlying inference engines has previously been an open question in the field.

As we show in Rainforth (2017; 2018), the results presented here can be brought to bear on assessing the relative correctness of the different ways PPSs allow model nesting. In particular, the correctness of sampling from the conditional distribution of one query within another follows from Theorem 3, but only if the computation for each call to the inner query increases the more times that query is called. This requirement is not satisfied by current systems. Meanwhile, Theorem 5 can be used to the assert that observing the output of one query inside another leads to convergence at the standard MC rate, provided that the computation of the inner query instead remains fixed.

7 Conclusions

We have introduced a formal framework for NMC estimation and shown that it can be used to yield a consistent estimator for problems that cannot be tackled with conventional MC alone. We have derived convergence rates and considered what minimal continuity assumptions are required for convergence. However, we have also highlighted a number of potential pitfalls for naïve application of NMC and provided guidelines for avoiding these, e.g. highlighting the importance of increasing the number of samples in both the inner and the outer estimators to ensure convergence. We have further introduced techniques for converting certain classes of NMC problems to conventional MC ones, providing improved convergence rates. Our work has implications throughout machine learning and we hope it will provide the foundations for exploring this plethora of applications.

Appendix A Proof of Theorem 1 - Simplified Convergence Rate

Theorem 1. *If f is Lipschitz continuous and $f(y_n, \gamma(y_n)), \phi(y_n, z_{n,m}) \in L^2$, the mean squared error of $I_{N,M}$ converges to 0 at rate $O(1/N + 1/M)$.*

Proof. Though the Theorem follows directly from Theorem 3, we also provide the following proof for this simplified case to provide a more accessible intuition behind the result. Note that the approach taken is distinct from the proof of Theorem 3.

Using Minkowski's inequality, we can bound the mean squared error of $I_{N,M}$ by

$$\mathbb{E}[(I - I_{N,M})^2] = \|I - I_{N,M}\|_2^2 \leq U^2 + V^2 + 2UV \leq 2\left(U^2 + V^2\right) \tag{25}$$

$$\text{where} \quad U = \left\| I - \frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n)) \right\|_2 \quad \text{and} \quad V = \left\| \frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n)) - I_{N,M} \right\|_2 .$$

We see immediately that $U = O\left(1/\sqrt{N}\right)$, since $\frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n))$ is a MC estimator for I, noting our assumption that $f(y_n, \gamma(y_n)) \in L^2$. For the second term,

$$\begin{aligned} V &= \left\| \frac{1}{N}\sum_{n=1}^{N} f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n)) \right\|_2 \\ &\leq \frac{1}{N}\sum_{n=1}^{N} \|f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))\|_2 \leq \frac{1}{N}\sum_{n=1}^{N} K \|(\hat{\gamma}_M)_n - \gamma(y_n)\|_2 \end{aligned}$$

where K is a fixed constant, again by Minkowski and using the assumption that f is Lipschitz. We can rewrite

$$\|(\hat{\gamma}_M)_n - \gamma(y_n)\|_2^2 = \mathbb{E}\left[\mathbb{E}\left[((\hat{\gamma}_M)_n - \gamma(y_n))^2 \middle| y_n\right]\right].$$

by the tower property of conditional expectation, and note that

$$\mathbb{E}\left[((\hat{\gamma}_M)_n - \gamma(y_n))^2 | y_n\right] = \mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} \phi(y_n, z_{n,m}) \middle| y_n\right) = \frac{1}{M}\mathrm{Var}\left(\phi(y_n, z_{n,1}) | y_n\right)$$

since each $z_{n,m}$ is i.i.d. and conditionally independent given y_n. As such

$$\|(\hat{\gamma}_M)_n - \gamma(y_n)\|_2^2 = \frac{1}{M}\mathbb{E}\left[\mathrm{Var}\left(\phi(y_n, z_{n,1}) | y_n\right)\right] = O(1/M),$$

noting that $\mathbb{E}\left[\mathrm{Var}\left(\phi(y_n, z_{n,1}) | y_n\right)\right]$ is a finite constant by our assumption that $\phi(y_n, z_{n,m}) \in L^2$. Consequently,

$$V \leq \frac{NK}{N} O\left(1/\sqrt{M}\right) = O\left(1/\sqrt{M}\right).$$

Substituting these bounds for U and V in (25) gives

$$\|I - I_{N,M}\|_2^2 \leq 2\left(O\left(1/\sqrt{N}\right)^2 + O\left(1/\sqrt{M}\right)^2\right) = O\left(1/N + 1/M\right)$$

as desired. □

Appendix B The Inevitable Bias of Nested Estimation

In this section we demonstrate formally that NMC schemes must produce biased estimates of $I(f)$ for certain functions f. In fact, our result applies more generally: we show that this holds for any MC scheme that makes use of imperfect estimates $\hat{\zeta}_n$ of $\gamma(y_n)$, either via a NMC procedure (e.g. $\hat{\zeta}_n = (\hat{\gamma}_M)_n$), or when these inner estimates are generated by some other methods such as a variational approximation (Blei et al., 2016) or Bayesian quadrature (O'Hagan, 1991).

Theorem 6. *Suppose that the random variables $\hat{\zeta}_n$ satisfy $\mathbb{P}(\hat{\zeta}_n \neq \gamma(y_n)) > 0$. Then we can choose f such that if $y_n \sim p(y)$, $\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] \neq I(f)$ for any N (including the limit $N \to \infty$).*

Proof. Take $f(y, w) = (\gamma(y) - w)^2$. Then $f(y, \gamma(y)) = 0$, so that $I(f) = 0$. On the other hand, $f(y_n, \hat{\zeta}_n) \geq 0$ since f is non-negative. Moreover, $f(y_n, \hat{\zeta}_n) > 0$ on the event $\{\hat{\zeta}_n \neq \gamma(y_n)\}$. Since we assumed this event has nonzero probability, it follows that $\mathbb{E}\left[f(y_n, \hat{\zeta}_n)\right] > 0$ and hence

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[f(y_n, \hat{\zeta}_n)\right] > 0 = I(f)$$

which gives the required result. □

It also follows from Jensen's inequality that *any* strictly convex or concave f entails a biased estimator when $\hat{\zeta}_n$ is unbiased but has non-zero variance given y_n, e.g. when $\hat{\zeta}_n$ is a MC estimate. More formally we have

Theorem 7. *Suppose that* $y_n \sim p(y)$ *and that each* $\hat{\zeta}_n$ *satisfies* $\mathbb{E}\left[\hat{\zeta}_n \middle| y_n\right] = \gamma(y_n)$. *Define* $\mathcal{A} \subseteq \mathcal{Y}$ *as* $\mathcal{A} = \left\{y \in \mathcal{Y} \;\middle|\; \mathrm{Var}\left(\hat{\zeta}_n \middle| y_n = y\right) > 0\right\}$ *and assume that* $\mathbb{P}(y_n \in \mathcal{A}) > 0$. *Then for any* f *that is strictly convex in its second argument,*

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] > I(f).$$

Similarly for any f *that is strictly concave in its second argument,*

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] < I(f).$$

Proof. We prove our claim for the case that f is strictly convex; our proof for the other concave case is symmetrical. We have

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(y_n, \hat{\zeta}_n)\right] = \mathbb{E}\left[f(y_1, \hat{\zeta}_1)\right] = \mathbb{E}\left[\mathbb{E}\left[f(y_1, \hat{\zeta}_1)\middle| y_1\right]\right] \geq \mathbb{E}\left[f\left(y_1, \mathbb{E}\left[\hat{\zeta}_1 \middle| y_1\right]\right)\right] = I(f)$$

where the $\geq$ is a result of Jensen's inequality on the inner expectation. Since f is strictly convex and therefore non-linear, equality holds if and only if $\hat{\zeta}_1$ is almost surely constant given y_1. This is violated whenever $y_1 \in \mathcal{A}$, which by assumption has a non-zero probability of occurring. Consequently, the inequality must be strict, giving the desired result. □

In addition to some special cases discussed in the Section 4, it may still be possible to develop unbiased estimation schemes for certain non-linear f using Russian roulette sampling (Lyne et al., 2015) or other debiasing techniques. However, these induce their own complications: for some problems the resultant estimates have infinite variance (Lyne et al., 2015) and as shown by (Jacob et al., 2015), there are no general purpose "f-factories" that produce both non-negative and unbiased estimates for non-constant, positive output functions $f : \mathbb{R} \to \mathbb{R}^+$, given unbiased estimates for the inputs.

Appendix C Proof of Theorem 2 - "Almost almost sure" convergence

Theorem 2. *For* $n \in \mathbb{N}$, *let*

$$(\epsilon_M)_n = \left|f(y_n, (\hat{\gamma}_M)_n) - f(y_n, \gamma(y_n))\right|.$$

Assume that $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$ *as* $M \to \infty$. *Let* Ω *be the sample space of our underlying probability space, so that* $I_{\tau_\delta(M),M}$ *forms a mapping from* Ω *to* $\mathbb{R}$. *Then, for every* $\delta > 0$, *there exists a measurable* $A_\delta \subseteq \Omega$ *with* $\mathbb{P}(A_\delta) < \delta$, *and a function* $\tau_\delta : \mathbb{N} \to \mathbb{N}$ *such that, for all* $\omega \notin A_\delta$,

$$I_{\tau_\delta(M),M}(\omega) \overset{a.s.}{\to} I \quad as \quad M \to \infty.$$

Proof. For all N, M, we have by the triangle inequality that

$$|I_{N,M} - I| \leq V_{N,M} + U_N, \quad \text{where}$$

$$V_{N,M} = \left|\frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n)) - I_{N,M}\right| \quad \text{and} \quad U_N = \left|I - \frac{1}{N}\sum_{n=1}^{N} f(y_n, \gamma(y_n))\right|.$$

A second application of the triangle inequality then allows us to write

$$V_{N,M} \leq \frac{1}{N}\sum_{n=1}^{N}(\epsilon_M)_n$$

where we recall that $(\epsilon_M)_n = |f(y_n, \gamma(y_n)) - f(y_n, \hat{\gamma}_n)|$. Now, for all fixed M, each $(\epsilon_M)_n$ is i.i.d. Furthermore, since $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$ as $M \to \infty$ by our assumption and $(\epsilon_M)_n$ is nonnegative, there exists some $L \in \mathbb{N}$ such that

$\mathbb{E}\left[|(\epsilon_M)_n|\right] < \infty$ for all $M \geq L$. Consequently, the strong law of large numbers means that as $N \to \infty$ then for all $M \geq L$

$$\frac{1}{N}\sum_{n=1}^{N}(\epsilon_M)_n \overset{a.s.}{\to} \mathbb{E}\left[(\epsilon_M)_1\right]. \tag{26}$$

For any fixed $\delta > 0$ then by repeatedly applying Egorov's theorem to each $M \geq L$, we can find a sequence of events

$$B_L, B_{L+1}, B_{L+2}, \ldots$$

such that for every $M \geq L$,

$$\mathbb{P}(B_M) < \frac{\delta}{4} \cdot \frac{1}{2^{M-L}}$$

and outside of B_M, the sequence $\frac{1}{N}\sum_{n=1}^{N}(\epsilon_M)_n$ converges *uniformly* to $\mathbb{E}\left[(\epsilon_M)_1\right]$. This uniform convergence (as opposed to just the piecewise convergence implied by (26)) now guarantees that we can define some function $\tau_\delta^1 : \mathbb{N} \to \mathbb{N}$ such that

$$\left|\frac{1}{M'}\sum_{n=1}^{M'}(\epsilon_M)_n(\omega) - \mathbb{E}\left[(\epsilon_M)_1\right]\right| < \frac{1}{M} \tag{27}$$

for all $M \geq L$, $M' \geq \tau_\delta^1(M)$, and $\omega \notin B_M$, remembering that ω is a point in our sample space. We further have that (27) holds for all $M \geq M_0$, $M' \geq \tau_\delta^1(M)$, and $\omega \notin B_\delta := \bigcup_{M \geq L} B_M$. Consequently, for all such M, M' and ω,

$$V_{M',M}(\omega) \leq \frac{1}{M'}\sum_{n=1}^{M'}(\epsilon_M)_n(\omega) < \frac{1}{M} + \mathbb{E}\left[(\epsilon_M)_1\right], \tag{28}$$

while we also have

$$\mathbb{P}(B_\delta) \leq \sum_{M \geq L}\mathbb{P}\left(B_M\right) < \sum_{M \geq L}\frac{\delta}{4} \cdot \frac{1}{2^{M-L}} = \frac{\delta}{2}. \tag{29}$$

To complete the proof, we must remove the dependence of U_N on N as well. This is straightforward once we observe that $U_N \overset{a.s.}{\to} 0$ as $N \to \infty$ by the strong law of large numbers. So, by Egorov's theorem again, there exists an event C_δ such that

$$\mathbb{P}(C_\delta) < \frac{\delta}{2} \tag{30}$$

and outside of C_δ, the sequence U_N converges uniformly to 0. This uniform convergence, in turn, ensures the existence of a function $\tau_\delta^2 : \mathbb{N} \to \mathbb{N}$ such that

$$U_{M'}(\omega) < \frac{1}{M} \tag{31}$$

for all $M \in \mathbb{N}$, $M' \geq \tau_\delta^2(M)$, and $\omega \notin C_\delta$.

We can now define $\tau_\delta(M) = \max(\tau_\delta^1(M), \tau_\delta^2(M))$, and $A_\delta = B_\delta \cup C_\delta$. By inequalities in (29) and (30),

$$\mathbb{P}(A_\delta) \leq \mathbb{P}(B_\delta) + \mathbb{P}(C_\delta) < \delta.$$

Also, by the inequalities in (28) and (31),

$$\left|I - I_{\tau_\delta(M),M}(\omega)\right| \leq V_{\tau_\delta(M),M}(\omega) + U_{\tau_\delta(M)}(\omega) \leq \frac{1}{M} + \frac{1}{M} + \mathbb{E}\left[(\epsilon_M)_1\right]$$

for all $M \geq L$ and $\omega \notin A_\delta$. Since $\mathbb{E}\left[(\epsilon_M)_1\right] \to 0$, we have here that $I_{\tau_\delta(M),M}(\omega) \to I$ as desired. □

Appendix D Proof of Theorem 3 - Convergence for Repeated Nesting

Theorem 3. *If $f_0, \cdots, f_D$ are all differentiable and Lipschitz continuous in their second input with Lipschitz constants*

$$K_k := \sup_{y^{(0:k)}}\left|\frac{\partial f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}}\right| < \infty,$$

for all $k \in 0, \ldots, D-1$ and if

$$\begin{aligned}\varsigma_k^2 &:= \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right] \\ &< \infty \quad \forall k \in 0, \ldots, D\end{aligned}$$

then

$$\mathbb{E}\left[\left(I_0-\gamma_0\right)^2\right] \leq \frac{\varsigma_0^2}{N_0}+\sum_{k=1}^{D}\left(\prod_{\ell=0}^{k-1}K_\ell^2\right)\frac{\varsigma_k^2}{N_k}+O(\epsilon) \tag{5}$$

where $O(\epsilon)$ represents asymptotically dominated terms.

If $f_0,\cdots,f_D$ are twice continuously differentiable with second derivative bounds

$$C_k := \sup_{y^{(0:k)}}\left|\frac{\partial^2 f_k\left(y^{(0:k)},\gamma_{k+1}(y^{(0:k)})\right)}{\partial\gamma_{k+1}^2}\right| < \infty$$

then this mean square error bound can be tightened to

$$\begin{aligned}\mathbb{E}\left[\left(I_0-\gamma_0\right)^2\right] &\leq \frac{\varsigma_0^2}{N_0}+\\ &\left(\frac{C_0\varsigma_1^2}{2N_1}+\sum_{k=0}^{D-2}\left(\prod_{d=0}^{k}K_d\right)\frac{C_{k+1}\varsigma_{k+2}^2}{2N_{k+2}}\right)^2+O(\epsilon).\end{aligned} \tag{6}$$

For a single nesting, we can further characterize $O(\epsilon)$ giving

$$\mathbb{E}\left[\left(I_0-\gamma_0\right)^2\right] \leq \frac{\varsigma_0^2}{N_0}+\frac{4K_0^2\varsigma_1^2}{N_0N_1}+\frac{2K_0\varsigma_0\varsigma_1}{N_0\sqrt{N_1}}+\frac{K_0^2\varsigma_1^2}{N_1} \tag{7}$$

$$\begin{aligned}\mathbb{E}\left[\left(I_0-\gamma_0\right)^2\right] &\leq \frac{\varsigma_0^2}{N_0}+\frac{C_0^2\varsigma_1^4}{4N_1^2}\left(1+\frac{1}{N_0}\right)\\ &+\frac{K_0^2\varsigma_1^2}{N_0N_1}+\frac{2K_0\varsigma_1}{N_0\sqrt{N_1}}\sqrt{\varsigma_0^2+\frac{C_0^2\varsigma_1^4}{4N_1^2}}+O\left(\frac{1}{N_1^3}\right)\end{aligned} \tag{8}$$

for when the continuous differentiability assumption does not hold and holds respectively.

Proof. As this is a long and involved proof, we start by defining a number of useful terms that will be used throughout. Unless otherwise stated, these definitions hold for all $k\in\{0,\ldots,D\}$. Note that most of these terms implicitly depend on the number of samples $N_0,N_1,\ldots,N_D$. However, s_k, $\zeta_{d,k}$, and ς_k do not and are thus constants for a particular problem.

$E_k\left(y^{(0:k-1)}\right)$ is the MSE of the estimator at depth k given $y^{(0:k-1)}$

$$E_k\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left(I_k\left(y^{(0:k-1)}\right)-\gamma_k\left(y^{(0:k-1)}\right)\right)^2\middle| y^{(0:k-1)}\right] \tag{32}$$

$\bar{f}_k\left(y^{(0:k-1)}\right)$ is the expected value of the estimate at depth k, or equivalently the expected function output using the estimate of the layer below

$$\begin{aligned}\bar{f}_k\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[I_k\left(y^{(0:k-1)}\right)\middle| y^{(0:k-1)}\right] \quad \forall k\in\{1,\ldots,D-1\}\\ &= \mathbb{E}\left[f_k\left(y^{(0:k)},I_{k+1}\left(y^{(0:k)}\right)\right)\middle| y^{(0:k-1)}\right]\end{aligned} \tag{33}$$

$v_k^2\left(y^{(0:k-1)}\right)$ is the variance of the estimator at depth k

$$\begin{aligned}v_k^2\left(y^{(0:k-1)}\right) &:= \text{Var}\left[I_k\left(y^{(0:k-1)}\right)\middle| y^{(0:k-1)}\right]\\ &= \mathbb{E}\left[\left(I_k\left(y^{(0:k-1)}\right)-\bar{f}_k\left(y^{(0:k-1)}\right)\right)^2\middle| y^{(0:k-1)}\right]\end{aligned} \tag{34}$$

$\beta_k\left(y^{(0:k-1)}\right)$ is the bias of the estimator at depth k

$$\begin{aligned}\beta_k\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[I_k\left(y^{(0:k-1)}\right)-\gamma_k\left(y^{(0:k-1)}\right)\middle| y^{(0:k-1)}\right]\\ &= \bar{f}_k\left(y^{(0:k-1)}\right)-\gamma_k\left(y^{(0:k-1)}\right)\\ &= \mathbb{E}\left[f_k\left(y^{(0:k)},I_{k+1}\left(y^{(0:k)}\right)\right)-f_k\left(y^{(0:k)},\gamma_{k+1}\left(y^{(0:k)}\right)\right)\middle| y^{(0:k-1)}\right]\end{aligned} \tag{35}$$

$s_k^2\left(y^{(0:k-1)}\right)$ is the variance at depth k of the true function output

$$s_k^2\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right] \tag{36}$$

$$s_D^2\left(y^{(0:D-1)}\right) := \mathbb{E}\left[\left(f_D\left(y^{(0:D)}\right) - \gamma_D\left(y^{(0:D)}\right)\right)^2 \middle| y^{(0:D-1)}\right] \tag{37}$$

$\zeta_{d,k}^2\left(y^{(0:k-1)}\right)$ is expectation of $s_d^2\left(y^{(0:d-1)}\right)$ over $y^{(k:d-1)}$

$$\begin{aligned}\zeta_{d,k}^2\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[s_d^2\left(y^{(0:d-1)}\right) \middle| y^{(0:k-1)}\right] \\ &= \mathbb{E}\left[\left(f_d\left(y^{(0:d)}, \gamma_{d+1}\left(y^{(0:d)}\right)\right) - \gamma_d\left(y^{(0:d-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right]\end{aligned} \tag{38}$$

ς_k^2 is expectation of $s_k^2\left(y^{(0:k-1)}\right)$ over all $y^{(0:k-1)}$

$$\varsigma_k^2 := \zeta_{k,0}^2 = \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2\right] \tag{39}$$

$A_k\left(y^{(0:k-1)}\right)$ is the MSE in the function output from using the estimate of the next layer, rather than the true value $\gamma_{k+1}\left(y^{(0:k)}\right)$, we fix $A_D := 0$

$$A_k\left(y^{(0:k-1)}\right) := \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) - f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)\right)^2 \middle| y^{(0:k-1)}\right] \tag{40}$$

$\sigma_k^2\left(y^{(0:k-1)}\right)$ is the variance in the function output from using the estimate of the next layer, we fix $\sigma_D^2\left(y^{0:D-1}\right) := s_D^2\left(y^{0:D-1}\right)$

$$\begin{aligned}\sigma_k^2\left(y^{(0:k-1)}\right) &:= \mathrm{Var}\left[f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) \middle| y^{(0:k-1)}\right] \\ &= \mathbb{E}\left[\left(f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right]\end{aligned} \tag{41}$$

$\omega_k\left(y^{(0:k-1)}\right)$ is the expectation over $y^{(k)}$ of the MSE for the next layer, we fix $\omega_D\left(y^{(0:D-1)}\right) := 0$

$$\begin{aligned}\omega_k\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right) \middle| y^{(0:k-1)}\right] \\ &= \mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^2 \middle| y^{(0:k-1)}\right]\end{aligned} \tag{42}$$

$\lambda_k\left(y^{(0:k-1)}\right)$ is the expectation over $y^{(k)}$ of the magnitude of the bias for the next layer, we fix $\lambda_D\left(y^{(0:D-1)}\right) := 0$ and note that $\lambda_{D-1}\left(y^{(0:D-2)}\right) := 0$ also as the innermost layer is an unbiased

$$\begin{aligned}\lambda_k\left(y^{(0:k-1)}\right) &:= \mathbb{E}\left[\left|\beta_{k+1}\left(y^{(0:k)}\right)\right| \middle| y^{(0:k-1)}\right] \\ &= \mathbb{E}\left[\left|\mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right) \middle| y^{(0:k)}\right]\right| \middle| y^{(0:k-1)}\right]\end{aligned} \tag{43}$$

Lipschitz Continuous Case

Given these definitions, we start by breaking the error down into a variance and bias term. Using the standard bias-variance decomposition we have

$$\begin{aligned}E_k\left(y^{(0:k-1)}\right) &= \mathbb{E}\left[\left(I_k\left(y^{(0:k-1)}\right) - \gamma_k\left(y^{(0:k-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right] \\ &= v_k^2\left(y^{(0:k-1)}\right) + \left(\beta_k\left(y^{(0:k-1)}\right)\right)^2\end{aligned} \tag{44}$$

It is immediately clear from its definition in (35) that the bias term $\left(\beta_k\left(y^{(0:k-1)}\right)\right)^2$ is independent of N_0. On the other hand, we will show later that the dominant components of the variance term for large $N_{0:D}$ depend only on N_0. We can thus think of increasing N_0 as reducing the variance of the estimator and increasing $N_{1:D}$ as reducing the bias.

We first consider the variance term

$$v_k^2\left(y^{(0:k-1)}\right) = \mathbb{E}\left[\left(\frac{1}{N_k}\sum_{n=1}^{N_k} f_k\left(y_n^{(0:k)}, I_{k+1}\left(y_n^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right]$$
$$= \frac{1}{N_k}\mathbb{E}\left[\left(f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right]$$

with the equality following because the $y_n^{(0:k)}$ being drawn i.i.d. and the expectation of each $f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right)$ equaling $\bar{f}_k\left(y^{(0:k-1)}\right)$ means that all the cross terms are zero. By the definition of σ_k^2 we now have

$$v_k^2\left(y^{(0:k-1)}\right) = \frac{\sigma_k^2\left(y^{(0:k-1)}\right)}{N_k}. \tag{45}$$

By using Minkowski's inequality and the definition of A_k it also follows that

$$\sigma_k\left(y^{(0:k-1)}\right) \leq \left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\left(f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) - \bar{f}_k\left(y^{(0:k-1)}\right)\right)^2 \middle| y^{(0:k-1)}\right]\right)^{\frac{1}{2}}. \tag{46}$$

Using a bias-variance decomposition on the second term above and noting that $s_k^2\left(y^{(0:k-1)}\right)$ and $\bar{f}_k\left(y^{(0:k-1)}\right) - \beta_k\left(y^{(0:k-1)}\right)$ are respectively the variance and expectation of $f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)$, we can rearrange the right-hand size of (46) to give

$$\sigma_k\left(y^{(0:k-1)}\right) \leq \left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} + \left(s_k^2\left(y^{(0:k-1)}\right) + \left(\beta_k\left(y^{(0:k-1)}\right)\right)^2\right)^{\frac{1}{2}}. \tag{47}$$

Here s_k^2 is independent of the number of samples used at any level of the estimate, while A_k and β_k^2 are independent of $N_d\ \forall d \leq k$. Now by Jensen's inequality, we have that

$$\left(\beta_k\left(y^{(0:k-1)}\right)\right)^2 \leq A_k\left(y^{(0:k-1)}\right) \tag{48}$$

noting that the only difference in the definition of $\left(\beta_k\left(y^{(0:k-1)}\right)\right)^2$ and $A_k\left(y^{(0:k-1)}\right)$ is whether the squaring occurs inside or outside the expectation. Therefore, presuming that A_k does not increase with $N_d\ \forall d > k$, neither will $\sigma_k^2\left(y^{(0:k-1)}\right)$, and so the variance term will converge to zero with rate $O(1/N_k)$. Further, if $A_k \to 0$ as $N_{k+1}, \ldots, N_D \to \infty$, then for a large number of inner samples $\sigma_k^2 \to s_k^2$ and thus we will have $v_k^2\left(y^{(0:k-1)}\right) \leq \frac{s_k^2}{N_k} + O(\epsilon)$ where $O(\epsilon)$ represents higher order terms that are dominated in the limit $N_k, \ldots, N_D \to \infty$. Provided this holds, we will also, therefore, have that

$$E_k\left(y^{(0:k-1)}\right) = \frac{\sigma_k^2\left(y^{(0:k-1)}\right)}{N_k} + \beta_k^2\left(y^{(0:k-1)}\right) = \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \beta_k^2\left(y^{(0:k-1)}\right) + O(\epsilon). \tag{49}$$

We now show that Lipschitz continuity is sufficient for $A_k \to 0$ and derive a concrete bound on the variance by bounding A_k. By definition of Lipschitz continuity, we have that

$$\left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} \leq \left(\mathbb{E}\left[K_k^2\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^2 \middle| y^{(0:k-1)}\right]\right)^{\frac{1}{2}}$$
$$= K_k\left(\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} \tag{50}$$

where we remember that $\omega_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\middle| y^{(0:k-1)}\right]$ is the expected MSE of the next level estimator. Once we also have an expression for the bias, we will thus be able to use this bound on A_k along with (44), (45), and (47) to inductively derive a bound on the error.

For the case where we only assume Lipschitz continuity then we will simply use the bound on the bias given by (48) leading to

$$E_k\left(y^{(0:k-1)}\right) \leq \frac{\sigma_k^2\left(y^{(0:k-1)}\right)}{N_k} + A_k\left(y^{(0:k-1)}\right) \tag{51}$$
$$\leq \frac{s_k^2\left(y^{(0:k-1)}\right) + 2A_k\left(y^{(0:k-1)}\right) + 2\left(A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}\left(s_k^2\left(y^{(0:k-1)}\right) + A_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}}{N_k} + A_k\left(y^{(0:k-1)}\right)$$

$$= \frac{s_k^2\left(y^{(0:k-1)}\right) + 2K_k^2\omega_k\left(y^{(0:k-1)}\right)}{N_k} + K_k^2\omega_k\left(y^{(0:k-1)}\right) + \frac{2K_k\left(\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}\left(s_k^2\left(y^{(0:k-1)}\right) + K_k^2\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}}}{N_k}$$

$$\leq \frac{s_k^2\left(y^{(0:k-1)}\right) + 4K_k^2\omega_k\left(y^{(0:k-1)}\right) + 2K_k\left(\omega_k\left(y^{(0:k-1)}\right)\right)^{\frac{1}{2}} s_k\left(y^{(0:k-1)}\right)}{N_k} + K_k^2\omega_k\left(y^{(0:k-1)}\right) \tag{52}$$

which fully defines a bound on conditional the variance of one layer given the mean squared error of the layer below. In particular as $\omega_D\left(y^{(0:D-1)}\right) = 0$ we now have

$$E_D\left(y^{(0:D-1)}\right) \leq \frac{s_D^2\left(y^{(0:D-1)}\right)}{N_D} = \frac{\mathbb{E}\left[\left(f_D\left(y^{(0:D)}\right) - \gamma_D\left(y^{(0:D)}\right)\right)^2 \middle| y^{(0:D-1)}\right]}{N_D}$$

which is the standard error for Monte Carlo convergence. We further have

$$\omega_{D-1}\left(y^{(0:D-2)}\right) = \mathbb{E}\left[E_D\left(y^{(0:D-1)}\right)\middle| y^{(0:D-2)}\right] = \frac{\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D}.$$

and thus

$$\begin{aligned} E_{D-1}\left(y^{(0:D-2)}\right) \leq & \frac{s_{D-1}^2\left(y^{(0:D-2)}\right)}{N_{D-1}} + \frac{4K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D N_{D-1}} \\ & + \frac{2K_{D-1}s_{D-1}\left(y^{(0:D-2)}\right)\zeta_{D,D-1}\left(y^{(0:D-2)}\right)}{N_{D-1}\sqrt{N_D}} + \frac{K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_D}. \end{aligned} \tag{53}$$

This leads to the following result for the single nesting case

$$E_0 \leq \frac{\varsigma_0^2}{N_0} + \frac{4K_0^2\varsigma_1^2}{N_0 N_1} + \frac{2K_0\varsigma_0\varsigma_1}{N_0\sqrt{N_1}} + \frac{K_0^2\varsigma_1^2}{N_1} \tag{54}$$

$\approx \frac{\varsigma_0^2}{N_0} + \frac{K_0^2\varsigma_1^2}{N_1} = O\left(\frac{1}{N_0} + \frac{1}{N_1}\right)$ where the approximation becomes exact as $N_0, N_1 \to \infty$. Note that there is no $O(\epsilon)$ term as this bound is exact in the finite sample case.

Things quickly get messy for double nesting and beyond so we will ignore non-dominant terms in the limit $N_0, \ldots, N_D \to \infty$ and resort to using $O(\epsilon)$ for these instead. We first note that removing dominated terms from (52) gives

$$E_k\left(y^{(0:k-1)}\right) \leq \frac{s_k^2}{N_k} + K_k^2\omega_k\left(y^{(0:k-1)}\right) + O(\epsilon) \tag{55}$$

as s_k^2 does not decrease with increasing $N_{k+1:D}$ whereas the other terms do. We therefore also have

$$\begin{aligned} \omega_k\left(y^{(0:k-1)}\right) &= \mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\middle| y^{(0:k-1)}\right] \\ &\leq \mathbb{E}\left[\frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}} + K_{k+1}^2\omega_{k+1}\left(y^{(0:k)}\right)\middle| y^{(0:k-1)}\right] + O(\epsilon) \end{aligned} \tag{56}$$

and therefore by recursively substituting (56) into itself we have

$$K_k^2\omega_k\left(y^{(0:k-1)}\right) \leq \sum_{d=k+1}^{D} \frac{\left(\prod_{\ell=k}^{d-1} K_\ell^2\right)\mathbb{E}\left[s_d^2\left(y^{(0:d-1)}\right)\middle| y^{(0:k-1)}\right]}{N_d} + O(\epsilon). \tag{57}$$

Now noting that $\zeta_{d,k}^2\left(y^{(0:k-1)}\right) = \mathbb{E}\left[s_d^2\left(y^{(0:d-1)}\right)\middle| y^{(0:k-1)}\right]$, substituting (57) back into (55) gives

$$E_k\left(y^{(0:k-1)}\right) = \frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k} + \sum_{d=k+1}^{D} \frac{\left(\prod_{\ell=k}^{d-1} K_\ell^2\right)\zeta_{d,k}^2\left(y^{(0:k-1)}\right)}{N_d} + O(\epsilon). \tag{58}$$

By definition we have that $\zeta_{0,0}^2 = s_0^2 = \varsigma_0^2$ and $\zeta_{d,0}^2 = \varsigma_d^2$ and as (58) holds in the case $k = 0$, the mean squared error of the overall estimator is as follows

$$\mathbb{E}\left[\left(I_0 - \gamma_0\right)^2\right] = E_0 \leq \frac{\varsigma_0^2}{N_0} + \sum_{k=1}^{D} \frac{\left(\prod_{\ell=0}^{k-1} K_\ell^2\right)\varsigma_k^2}{N_k} + O(\epsilon) \tag{59}$$

and we have the desired result for the Lipschitz case.

Twice Continuously Differentiable Case

We now revisit the bound for the bias in the twice continuously differentiable case to show that a tighter overall bound can be found. We first remember that

$$\beta_k\left(y^{(0:k-1)}\right) = \mathbb{E}\left[f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) - f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right)\middle| y^{(0:k-1)}\right].$$

Taylor's theorem implies that for any twice continuously differentiable f_k we can write

$$\begin{aligned} f_k\left(y^{(0:k)}, I_{k+1}\left(y^{(0:k)}\right)\right) &- f_k\left(y^{(0:k)}, \gamma_{k+1}\left(y^{(0:k)}\right)\right) \\ &= \frac{\partial f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}}\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right) \\ &+ \frac{1}{2}\frac{\partial f_k^2\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^2}\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^2 \\ &+ h_3\left(I_{k+1}\left(y^{(0:k)}\right)\right)\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^3 \end{aligned} \tag{60}$$

where $\lim_{x\to\gamma_{k+1}\left(y^{(0:k)}\right)} h_3(x) = 0$. Consequently, the last term is $O\left(\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^3\right)$ and so will diminish in magnitude faster than the first two terms provided that the derivatives are bounded, which is guaranteed by our assumptions. We will thus use $O(\epsilon)$ for this term and note that it is dominated in the limit.

Now defining

$$\delta_{\ell,k} = \mathbb{E}\left[\frac{\partial f_k^\ell\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^\ell}\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^\ell \middle| y^{(0:k-1)}\right]$$

then we have

$$\beta_k^2\left(y^{(0:k-1)}\right) = \delta_{1,k}^2 + \frac{\delta_{2,k}^2}{4} + \delta_{1,k}\delta_{2,k} + O(\epsilon).$$

By using the tower property we further have that

$$\begin{aligned} \delta_{\ell,k} &= \mathbb{E}\left[\mathbb{E}\left[\frac{\partial f_k^\ell\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^\ell}\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^\ell \middle| y^{(0:k)}\right] \middle| y^{(0:k-1)}\right] \\ &= \mathbb{E}\left[\frac{\partial f_k^\ell\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^\ell}\mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^\ell \middle| y^{(0:k)}\right] \middle| y^{(0:k-1)}\right] \\ &\leq \mathbb{E}\left[\left|\frac{\partial f_k^\ell\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^\ell}\right| \left|\mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^\ell \middle| y^{(0:k)}\right]\right| \middle| y^{(0:k-1)}\right] \\ &\leq \left(\sup_{y^{(0)}}\left|\frac{\partial^\ell f_k\left(y^{(0:k)}, \gamma_{k+1}(y^{(0:k)})\right)}{\partial \gamma_{k+1}^\ell}\right|\right)\mathbb{E}\left[\left|\mathbb{E}\left[\left(I_{k+1}\left(y^{(0:k)}\right) - \gamma_{k+1}\left(y^{(0:k)}\right)\right)^\ell \middle| y^{(0:k)}\right]\right| \middle| y^{(0:k-1)}\right]. \end{aligned}$$

Now for the particular cases of $\ell = 1$ and $\ell = 2$ then the derivative terms where defined in the theorem and the expectations correspond respectively to our definitions of λ_k and ω_k giving

$$\begin{aligned} \delta_{1,k} &\leq K_k \lambda_k\left(y^{(0:k-1)}\right) \\ \delta_{2,k} &\leq C_k \omega_k\left(y^{(0:k-1)}\right) \end{aligned}$$

and therefore

$$\begin{aligned} \beta_k^2\left(y^{(0:k-1)}\right) &\leq K_k^2\lambda_k^2\left(y^{(0:k-1)}\right) + \frac{C_k^2}{4}\omega_k^2\left(y^{(0:k-1)}\right) + K_k\, C_k\, \lambda_k\left(y^{(0:k-1)}\right)\omega_k\left(y^{(0:k-1)}\right) + O(\epsilon) \\ &= \left(K_k\lambda_k\left(y^{(0:k-1)}\right) + \frac{C_k}{2}\omega_k\left(y^{(0:k-1)}\right)\right)^2 + O(\epsilon). \end{aligned} \tag{61}$$

Remembering (49) we can recursively define the error bound in the same manner as the Lipschitz case. We can immediately see that, as $\beta_D = 0$ without any nesting, we recover the bound from the Lipschitz case for the inner most estimator as

expected. As the innermost estimator is unbiased we also have $\lambda_{D-1}\left(y^{(0:D-2)}\right)=0$ and so

$$\begin{aligned}\beta_{D-1}^2\left(y^{(0:D-2)}\right) &\leq \frac{C_{D-1}^2}{4}\omega_{D-1}^2\left(y^{(0:D-2)}\right)+O(\epsilon)\\ &\leq \frac{C_{D-1}^2}{4}\left(\mathbb{E}\left[\frac{s_D^2\left(y^{(0:D-1)}\right)}{N_D}\middle| y^{(0:D-2)}\right]\right)^2+O(\epsilon)\\ &= \frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}+O(\epsilon).\end{aligned}$$

Going back to our original bound on $\sigma_{D-1}^2\left(y^{(0:D-2)}\right)$ given in (47) and substituting for $\beta_{D-1}\left(y^{(0:D-2)}\right)$ we now have

$$\sigma_{D-1}\left(y^{(0:D-2)}\right)\leq\left(A_{D-1}\left(y^{(0:D-2)}\right)\right)^{\frac{1}{2}}+\left(s_{D-1}^2\left(y^{(0:D-2)}\right)+\frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}+O(\epsilon)\right)^{\frac{1}{2}}. \tag{62}$$

There does not appear to be tighter bound for $A_{D-1}\left(y^{(0:D-2)}\right)$ than in the Lipschitz continuous case and so using the same bound of $A_{D-1}\left(y^{(0:D-2)}\right)\leq K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)/N_{D-1}$ we have

$$\begin{aligned}E_{D-1}\left(y^{(0:D-2)}\right)\leq&\frac{\sigma_{D-1}^2\left(y^{(0:D-2)}\right)}{N_{D-1}}+\frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}+O(\epsilon)\\ \leq&\frac{s_{D-1}^2\left(y^{(0:D-2)}\right)}{N_{D-1}}+\frac{K_{D-1}^2\zeta_{D,D-1}^2\left(y^{(0:D-2)}\right)}{N_DN_{D-1}}+\frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}\left(1+\frac{1}{N_{D-1}}\right)\\ &+\frac{2K_{D-1}\zeta_{D,D-1}\left(y^{(0:D-2)}\right)}{N_{D-1}\sqrt{N_D}}\left(s_{D-1}\left(y^{(0:D-2)}\right)^2+\frac{C_{D-1}^2\,\zeta_{D,D-1}^4\left(y^{(0:D-2)}\right)}{4N_D^2}\right)^{\frac{1}{2}}+O(\epsilon).\end{aligned} \tag{63}$$

Therefore for the single nesting case, we now have

$$E_0\leq\frac{\varsigma_0^2}{N_0}+\frac{K_0^2\varsigma_1^2}{N_0N_1}+\frac{2K_0\varsigma_1}{N_0\sqrt{N_1}}\sqrt{\varsigma_0^2+\frac{C_0^2\varsigma_1^4}{4N_1^2}}+\frac{C_0^2\varsigma_1^4}{4N_1^2}\left(1+\frac{1}{N_0}\right)+O\left(\frac{1}{N_1^3}\right) \tag{64}$$

$\approx\frac{\varsigma_0^2}{N_0}+\frac{C_0^2\varsigma_1^4}{4N_1^2}=O\left(\frac{1}{N_0}+\frac{1}{N_1^2}\right)$ where again the approximation becomes tight when $N_0,N_1\to\infty$. Here we have used the fact that the only $O(\epsilon)$ term comes from the Taylor expansion and is equal to $O\left(\frac{1}{N_1^3}\right)$ because we have $\delta_{1,D-1}=0$ and therefore

$$\begin{aligned}O(\epsilon)=&O\left(\delta_{2,D-1}\delta_{3,D-1}+\delta_{2,D-1}\delta_{4,D-1}\right)\\ =&O\left(\delta_{2,D-1}\mathbb{E}\left[\left(I_1\left(y^{(0)}\right)-\gamma_1\left(y^{(0)}\right)\right)^3\middle|y^{(0)}\right]\right)+O\left(\delta_{2,D-1}\mathbb{E}\left[\left(I_1\left(y^{(0)}\right)-\gamma_1\left(y^{(0)}\right)\right)^4\middle|y^{(0)}\right]\right)\\ =&O\left(\frac{1}{N_1}\mathbb{E}\left[\left(\frac{1}{N_1}\sum_{n=1}^{N_1}f_1\left(y_n^{(0:1)}\right)-\mathbb{E}\left[f_1\left(y^{(0:1)}\right)\middle|y^{(0)}\right]\right)^3\middle|y^{(0)}\right]\right)\\ &+O\left(\frac{1}{N_1}\mathbb{E}\left[\left(\frac{1}{N_1}\sum_{n=1}^{N_1}f_1\left(y_n^{(0:1)}\right)-\mathbb{E}\left[f_1\left(y^{(0:1)}\right)\middle|y^{(0)}\right]\right)^4\middle|y^{(0)}\right]\right)\end{aligned}$$

now noting that the $y_n^{(0:1)}$ are i.i.d., and that $\mathbb{E}\left[f_1\left(y_1^{(0:1)}\right)-\mathbb{E}\left[f_1\left(y^{(0:1)}\right)\middle|y^{(0)}\right]\middle|y^{(0)}\right]=0$ such many of the cross terms when expanding the brackets are zero, we have

$$\begin{aligned}=&O\left(\frac{1}{N_1^4}\sum_{n=1}^{N_1}\mathbb{E}\left[\left(f_1\left(y_1^{(0:1)}\right)-\mathbb{E}\left[f_1\left(y^{(0:1)}\right)\middle|y^{(0)}\right]\right)^3\middle|y^{(0)}\right]\right)\\ &+O\left(\frac{1}{N_1^5}\sum_{n=1}^{N_1}\mathbb{E}\left[\left(f_1\left(y_1^{(0:1)}\right)-\mathbb{E}\left[f_1\left(y^{(0:1)}\right)\middle|y^{(0)}\right]\right)^4\middle|y^{(0)}\right]\right)\\ &+O\left(\frac{3}{N_1^5}\sum_{n=1}^{N_1}\sum_{m=1,m\neq n}^{N_1}\left(\mathbb{E}\left[\left(f_1\left(y_1^{(0:1)}\right)-\mathbb{E}\left[f_1\left(y^{(0:1)}\right)\middle|y^{(0)}\right]\right)^2\middle|y^{(0)}\right]\right)^2\right)\end{aligned}$$

$$=O\left(\frac{1}{N_1^3}\right)+O\left(\frac{1}{N_1^4}\right)+O\left(\frac{1}{N_1^3}\right)=O\left(\frac{1}{N_1^3}\right)$$

as required.

Returning to calculating the bound for the repeated nesting case then by substituting (61) into (49) we have more generally

$$E_k\left(y^{(0:k-1)}\right)\leq\frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k}+\left(K_k\lambda_k\left(y^{(0:k-1)}\right)+\frac{C_k}{2}\omega_k\left(y^{(0:k-1)}\right)\right)^2+O(\epsilon). \tag{65}$$

Now remembering that $\omega_k\left(y^{(0:k-1)}\right)=\mathbb{E}\left[E_{k+1}\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]$ from (49) we have

$$\begin{aligned}\omega_k\left(y^{(0:k-1)}\right)&=\mathbb{E}\left[\frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}}+\beta_{k+1}^2\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]+O(\epsilon)\\&=\frac{\zeta_{k+1,k}^2}{N_{k+1}}+\mathbb{E}\left[\beta_{k+1}^2\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]+O(\epsilon).\end{aligned} \tag{66}$$

We also have that except at $k=D-1$ and $k=D$ (for which both λ_k and β_{k+1} are zero), then

$$\lambda_k\left(y^{(0:k-1)}\right)=\mathbb{E}\left[\left|\beta_{k+1}\left(y^{(0:k)}\right)\right|\middle|y^{(0:k)}\right]\gg\mathbb{E}\left[\beta_{k+1}^2\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]$$

for sufficiently large $N_{k+1},\ldots,N_D$. This means that when we substitute (66) into (65), the second term in (66) becomes dominated giving

$$E_k\left(y^{(0:k-1)}\right)\leq\frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k}+\left(K_k\lambda_k\left(y^{(0:k-1)}\right)+\frac{C_k\zeta_{k+1,k}^2}{2N_{k+1}}\right)^2+O(\epsilon). \tag{67}$$

Now as $\beta_{k+1}^2\left(y^{(0:k)}\right)=E_{k+1}\left(y^{(0:k)}\right)-\frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}}$ we have

$$\lambda_k\left(y^{(0:k-1)}\right)=\mathbb{E}\left[\sqrt{E_{k+1}\left(y^{(0:k)}\right)-\frac{s_{k+1}^2\left(y^{(0:k)}\right)}{N_{k+1}}}\middle|y^{(0:k-1)}\right]+O(\epsilon)$$

and substituting in (67) gives

$$\begin{aligned}\lambda_k\left(y^{(0:k-1)}\right)&\leq\mathbb{E}\left[K_{k+1}\lambda_{k+1}\left(y^{(0:k)}\right)+\frac{C_{k+1}\zeta_{k+2,k+1}^2}{2N_{k+2}}\middle|y^{(0:k-1)}\right]+O(\epsilon)\\&=\frac{C_{k+1}\zeta_{k+2,k}^2}{2N_{k+2}}+K_{k+1}\mathbb{E}\left[\lambda_{k+1}\left(y^{(0:k)}\right)\middle|y^{(0:k-1)}\right]+O(\epsilon)\\&\leq\frac{C_{k+1}\zeta_{k+2,k}^2}{2N_{k+2}}+\sum_{d=k+1}^{D-2}\mathbb{E}\left[\left(\prod_{\ell=k+1}^{d}K_\ell\right)\frac{C_{d+1}\zeta_{d+2,d}^2}{2N_{d+2}}\middle|y^{(0:k-1)}\right]+O(\epsilon)\\&\leq\frac{C_{k+1}\zeta_{k+2,k}^2}{2N_{k+2}}+\sum_{d=k+1}^{D-2}\left(\prod_{\ell=k+1}^{d}K_\ell\right)\frac{C_{d+1}\zeta_{d+2,k}^2}{2N_{d+2}}+O(\epsilon)\end{aligned}$$

and thus

$$E_k\left(y^{(0:k-1)}\right)\leq\frac{s_k^2\left(y^{(0:k-1)}\right)}{N_k}+\frac{1}{4}\left(\frac{C_k\zeta_{k+1,k}^2}{N_{k+1}}+\sum_{d=k}^{D-2}\left(\prod_{\ell=k}^{d}K_\ell\right)\frac{C_{d+1}\zeta_{d+2,k}^2}{N_{d+2}}\right)^2+O(\epsilon).$$

and therefore

$$\mathbb{E}\left[\left(I_0-\gamma_0\right)^2\right]=E_0\leq\frac{\varsigma_0^2}{N_0}+\frac{1}{4}\left(\frac{C_0\varsigma_1^2}{N_1}+\sum_{k=0}^{D-2}\left(\prod_{d=0}^{k}K_d\right)\frac{C_{k+1}\varsigma_{k+2}^2}{N_{k+2}}\right)^2+O(\epsilon)$$

as required and we are done.

□

Appendix E Proof of Theorem 4 - Convergence Rate for Finite Realisations of y

Theorem 4. *If f is Lipschitz continuous, then the mean squared error of $I_N = \sum_{c=1}^{C} (\hat{P}_N)_c \, (\hat{f}_N)_c$ as an estimator for I as per* (10) *converges at rate $O(1/N)$.*

Proof. Denote $P_c = P(y = y_c)$ and $f_c = f(y_c, \gamma(y_c))$ noting that as the y_c are fixed values, so are P_c and f_c. Then, Minkowski's inequality allows us to bound the mean squared error as

$$\mathbb{E}\left[(I_N - I)^2\right] = \|I_N - I\|_2^2 \leq \left(\sum_{c=1}^{C} W_c\right)^2 \quad \text{where} \quad W_c := \left\| (\hat{P}_N)_c \, (\hat{f}_N)_c - P_c \, f_c \right\|_2 .$$

Moreover, again by Minkowski, we have $W_c \leq U_c + V_c$ where

$$U_c = \left\| (\hat{P}_N)_c \, (\hat{f}_N)_c - (\hat{P}_N)_c \, f_c \right\|_2 , \quad V_c = \left\| (\hat{P}_N)_c \, f_c - P_c \, f_c \right\|_2 .$$

Factoring out $(\hat{P}_N)_c$ in U_c and noting that each y_n and $z_{n,c}$ are sampled independently gives

$$U_c = \sqrt{\mathbb{E}\left[(\hat{P}_N)_c^2 \, \left((\hat{f}_N)_c - f_c\right)^2\right]} = \sqrt{\mathbb{E}\left[(\hat{P}_N)_c^2\right]} \sqrt{\mathbb{E}\left[\left((\hat{f}_N)_c - f_c\right)^2\right]}.$$

Using Minkowski's inequality, we may write the first right-hand term as

$$\sqrt{\mathbb{E}\left[(\hat{P}_N)_c^2\right]} = \left\| (\hat{P}_N)_c \right\|_2 \leq \frac{1}{N} \sum_{n=1}^{N} \| \mathbb{1}(y_n = y_c) \|_2 = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[\mathbb{1}(y_n = y_c)^2\right] = \frac{1}{N} \sum_{n=1}^{N} P_c = P_c.$$

For the second term, note that by Lipschitz continuity, we have for some constant $K > 0$

$$\sqrt{\mathbb{E}\left[\left((\hat{f}_N)_c - f_c\right)^2\right]} = \left\| (\hat{f}_N)_c - f_c \right\|_2 \leq K \left\| \frac{1}{N} \sum_{n=1}^{N} \phi(y_c, z_{n,c}) - \gamma(y_c) \right\|_2 = K \cdot O(1/\sqrt{N}) = O(1/\sqrt{N}),$$

since $\frac{1}{N} \sum_{n=1}^{N} \phi(y_c, z_{n,c})$ is a Monte Carlo estimator for $\gamma(y_c)$. Altogether then, we have that

$$U_c = P_c \cdot O(1/\sqrt{N}) = O(1/\sqrt{N}).$$

We can also factor out f_c in V_c to obtain

$$V_c = |f_c| \cdot \left\| (\hat{P}_N)_c - P_c \right\|_2 = |f_c| \cdot O(1/\sqrt{N}) = O(1/\sqrt{N}),$$

since $(\hat{P}_N)_c$ is a Monte Carlo estimator for P_c. Now by noting that $(A+B)^2 \leq 2(A^2 + B^2)$ for any $A, B \in \mathbb{R}$, an inductive argument shows that

$$\left(\sum_{\ell=1}^{L} A_\ell\right)^2 \leq 2^{\lceil \log_2 L \rceil} \sum_{\ell=1}^{L} A_\ell^2$$

for all $A_1, \cdots, A_L \in \mathbb{R}$. We can now show that our asymptotic bounds for U_c and V_c entail that our overall mean squared error satisfies

$$\begin{aligned} \mathbb{E}\left[(I_N - I)^2\right] &\leq 2^{\lceil \log_2 C \rceil} \sum_{c=1}^{C} W_c^2 \leq 2^{\lceil \log_2 C \rceil} \sum_{c=1}^{C} (U_c + V_c)^2 \leq 2^{\lceil \log_2 C \rceil + 1} \sum_{c=1}^{C} U_c^2 + V_c^2 \\ &= 2^{\lceil \log_2 C \rceil + 1} \sum_{c=1}^{C} O(1/N) + O(1/N) = O(1/N), \end{aligned}$$

as desired. □

Appendix F Proof for Theorem 5 - Products of Expectations

Theorem 5. *Consider the NMC estimator*

$$I_N = \frac{1}{N} \sum_{n=1}^{N} f\left(y_n, \prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \psi_\ell(y_n, z'_{n,\ell,m})\right)$$

where each $y_n \in \mathcal{Y}$ and $z'_{n,\ell,m} \in \mathcal{Z}_\ell$ are independently drawn from $y_n \sim p(y)$ and $z'_{n,\ell,m} | y_n \sim p(z_\ell | y_n)$, respectively. If f is linear, the estimator converges almost surely to I, with a convergence rate of $O(1/N)$ in the mean square error for any

fixed choice of $\{M_\ell\}_{\ell=1:L}$.

Proof. Consider fixed sizes of nested sample sets, $\{M_\ell\}_{\ell=1:L}$. For each $y \in \mathcal{Y}$ and

$$x = \{\{z'_{\ell,m}\}_{m=1:M_\ell}\}_{\ell=1:L} \in \mathcal{X} = \mathcal{Z}_1^{M_1} \otimes \cdots \otimes \mathcal{Z}_L^{M_L},$$

define

$$\eta(y, x) = f\left(y, \prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m_\ell=1}^{M_\ell} \psi_\ell(y, z'_{\ell,m})\right).$$

Now, $I_N = \frac{1}{N}\sum_{n=1}^{N} \eta(y_n, x_n)$ is a standard MC estimator on the space $\mathcal{Y} \otimes \mathcal{X}$. Thus, $I_N \overset{a.s.}{\to} \mathbb{E}[I_N]$ with convergence properties and rate as per standard MC. We finish the proof by showing that $\mathbb{E}[I_N] = I$ when f is linear:

$$\mathbb{E}[I_N] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f\left(y_n, \prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \psi_\ell(y_n, z'_{n,\ell,m})\right)\right] = \mathbb{E}\left[\mathbb{E}\left[f\left(y_1, \prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \psi_\ell(y_1, z'_{1,\ell,m})\right)\middle| y_1\right]\right],$$

now using the linearity of f

$$= \mathbb{E}\left[f\left(y_1, \mathbb{E}\left[\prod_{\ell=1}^{L} \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \psi_\ell(y_1, z'_{1,\ell,m})\middle| y_1\right]\right)\right],$$

and using the fact that terms for different ℓ are by construction independent

$$= \mathbb{E}\left[f\left(y_1, \prod_{\ell=1}^{L} \mathbb{E}\left[\frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \psi_\ell(y_1, z'_{1,\ell,m})\middle| y_1\right]\right)\right] = \mathbb{E}\left[f\left(y_1, \prod_{\ell=1}^{L} \mathbb{E}\left[\psi_\ell(y_1, z'_{1,\ell,1})\middle| y_1\right]\right)\right] = I,$$

as required. □

Appendix G Optimizing the Convergence Rates

We have shown that the mean squared error converges at a rate

$$O\left(\sum_{k=0}^{D} \frac{1}{N_k}\right) \quad \text{or} \quad O\left(\frac{1}{N_0} + \left(\sum_{k=1}^{D} \frac{1}{N_k}\right)^2\right)$$

depending on the smoothness assumptions that can be made about f. Here we show that given a sample budget for the inner most estimator $T = \prod_{k=0}^{D} N_k$, then these bounds are optimized by setting $N_0 \propto N_1 \propto \cdots \propto N_D$ and $N_0 \propto N_1^2 \propto \cdots \propto N_D^2$ respectively for the two cases and that this gives bounds of $O\left(1/T^{\frac{1}{D+1}}\right)$ and $O\left(1/T^{\frac{2}{D+2}}\right)$respectively. For the single nested case, this gives bounds of $O(1/\sqrt{T})$ and $O(1/T^{2/3})$ respectively.

We start by explaining why T is an appropriate measure of the overall computational cost. First note that for each sample of $y^{(0:k)}$, the NMC estimator requires N_k samples of $y^{(k+1)}$. Thus there are N_0 samples of the outermost level, $N_0 \times N_1$ of the next level, and $T = \prod_{k=0}^{D} N_k$ samples of the innermost level, regardless of the setup. In other words, each individual estimate of the innermost level uses N_D samples and we generate $\prod_{k=0}^{D-1} N_k = T/N_D$ of these estimates because we need to generate one estimate for each sample of the layer above. Thus what we can vary for a fixed T is whether we use more estimates each using fewer samples, or fewer estimates each using more samples.

Now the total cost of generating I_0 scales with sum the costs of each individual layer, namely

$$\text{Cost} = \sum_{k=0}^{D} c_k \prod_{\ell=0}^{k} N_\ell$$

where c_k is the per sample cost local computations made at the k^{th} layer (i.e. sampling $y^{(0:k)}$ and evaluating f_k for given inputs), which is independent of the N_k. For large N_D, we see that the dominant cost is that of the inner most layer, namely $c_T \prod_{\ell=0}^{D} N_\ell = c_T T$, and we asymptotically spend 100% of our time dealing with the innermost estimator. To give intuition to this, think about writing the estimator as a hierarchy of nested for loops; as the length of the loops increases we spend an increasing proportion of our time inside the innermost loop. Consequently, in the asymptotic regime, our computational cost is $O(T)$ and we can use T is an appropriate measure of the overall computational cost.

To derive the optimal rates, we first consider the single nested case where $D = 1$, $N_0 = N$, and $N_1 = M$. Consider setting

$N = \tau(M)$ then $T = \tau(M) \cdot M$ and our bounds become $O(R)$, where

$$R = 1/\tau(M) + 1/M \quad \text{and} \quad R = 1/\tau(M) + 1/M^2.$$

for the two cases respectively.

In this first case supposing $\tau(M) = O(M)$ easily gives

$$T = M\tau(M) = O\left(M^2\right)$$

and as such

$$R = O\left(\frac{1}{M}\right) = O\left(\frac{1}{\sqrt{T}}\right) \tag{68}$$

as $M \to \infty$. In contrast, consider the case that $\tau(M) \gg M$ as $M \to \infty$. We then have $\frac{1}{\sqrt{M}} \gg \frac{1}{\sqrt{\tau(M)}}$ as $M \to \infty$, so that

$$R = O\left(\frac{1}{M}\right) \gg \frac{1}{\sqrt{M}}\frac{1}{\sqrt{\tau(M)}} = \frac{1}{\sqrt{T}}.$$

Comparing with (68), we observe that, for the same total budget of samples T, this choice of τ provides a strictly weaker convergence guarantee than in the previous case. When $M \gg \tau(M)$ also then we have $\frac{1}{\sqrt{\tau(M)}} \gg \frac{1}{\sqrt{M}}$ as $M \to \infty$ and so

$$R = O\left(\frac{1}{\tau(M)}\right) \gg \frac{1}{\sqrt{M}}\frac{1}{\sqrt{\tau(M)}} = \frac{1}{\sqrt{T}}$$

which is again a weaker bound. We thus see that the $O(1/N + 1/M)$ bound is optimized when $N \propto M$, giving a convergence rate of $O(1/\sqrt{T})$.

In the second case suppose that $\tau(M) = O(M^2)$ as $M \to \infty$. This now gives

$$T = M\tau(M) = O\left(M^3\right)$$

and therefore

$$R = O\left(\frac{1}{M^2}\right) = O\left(\frac{1}{T^{2/3}}\right)$$

as $M \to \infty$. Now considering the cases $\tau(M) \gg M^2$ leads to $\frac{1}{M^{4/3}} \gg \frac{1}{\tau(M)^{2/3}}$ and thus

$$R = O\left(\frac{1}{M^2}\right) \gg \frac{1}{M^{2/3}}\frac{1}{\tau(M)^{2/3}} = \frac{1}{T^{2/3}}.$$

Similarly, if $\tau(M) \ll M^2$ then $\frac{1}{\tau(M)^{1/3}} \gg \frac{1}{M^{2/3}}$ and thus

$$R = O\left(\frac{1}{\tau(M)}\right) \gg \frac{1}{M^{2/3}}\frac{1}{\tau(M)^{2/3}} = \frac{1}{T^{2/3}}.$$

Both of these cases lead to weaker bounds and so we see that the $O(1/N + 1/M^2)$ bound is tightest when $N \propto M^2$, giving a convergence rate of $O(1/T^{2/3})$.

We now consider the repeated nesting case without twice continuous differentiability such that our bound is $O\left(\sum_{k=0}^{D} \frac{1}{N_k}\right)$. Here we can immediately see that $N_0 \propto N_1 \propto \cdots \propto N_D$ leads to $N_k \propto T^{\frac{1}{D+1}}$ and thus $O\left(1/T^{\frac{1}{D+1}}\right)$ convergence. If we were to set any $N_k \ll T^{\frac{1}{D+1}}$ then this term would dominate the sum and lead to a worse converge. Thus the result from the single nested case trivially extends to the multiple nested case, giving the required result.

Finally considering repeated nesting for the bound $O\left(\frac{1}{N_0} + \left(\sum_{k=1}^{D} \frac{1}{N_k}\right)^2\right)$ then we have from the previous result that $N_1 \propto N_2 \propto \cdots \propto N_D$ is required for optimality as otherwise one of the terms in the summation dominates the other terms. If we now define $M = \prod_{k=1}^{D} N_k = T/N_0$ then we get a convergence rate of $O(1/N_0 + 1/M^2)$ which is identical to the single nesting case for this tighter bound. We, therefore, have that the optimal configuration must be $N_0 \propto N_1^2 \propto \cdots \propto N_D^2$ giving a bound of $O\left(1/T^{\frac{2}{D+2}}\right)$ as it gives $N_0 \propto T^{\frac{2}{D+2}}$.

Appendix H Additional details pertaining to cancer simulator

In this section, we elucidate some more details about the cancer simulator described in the manuscript, provide more rigorous mathematical definitions for the relevant terms using the same nomenclature, and also include more results figures.

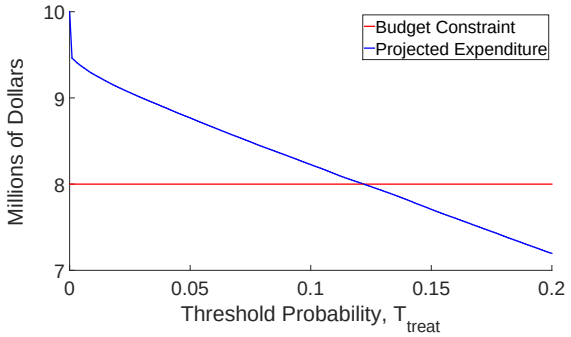


Figure 7. Projected expenditure (proportional to $I_{N,M}$) evaluated at different values of T_{treat}. The budget constraint is shown by the horizontal red line. The optimal value of T_{treat} is found by the intersection and occurs at $T_{\text{treat}} = 12.5\%$. Evaluated was carried out at 100 $T_{\text{treat}1}$. Only the bottom 20% is pictured as this is the operating range for most treatment centers.

H.1 Simulator details

We define $I(T_{\text{treat}})$ to be the expected proportion of patients who receive treatment. A particular patient is represented by $y \in \mathcal{R}^d$. Specifically, y consists of only a single real number ($d = 1$) representing the size of the tumor upon discovery. Initial tumor size is drawn from a scaled Rayleigh distribution. The outcome of the simulator is then $\phi(y, z) \in \{0, 1\}$, and is the binary outcome of whether that particular patient and sample of unobserved parameters yield an expected tumor size below the threshold, T_{opp}, after a fixed time duration, t_{max}. The simulator is a pair of coupled, parameterized differential equations for the action of an anti-tumor treatment such as chemotherapy, as described in Enderling and Chaplain (2014):

$$\frac{dc}{dt} = -\lambda c \log\left(\frac{c}{K}\right) - \xi c \tag{69}$$

$$\frac{dK}{dt} = \phi c - \psi K c^{2/3}, \tag{70}$$

where $c(t, x) \in \mathcal{R}_+$ represents tumor size, with initial size y_n. Similarly, $K(t, x) \in \mathcal{R}_+$ represents the notion of a carrying capacity, with the initial carrying capacity, $K(0, z)$, set to a known constant K_0. The magnitude of the patient response to an anti-tumor treatment (such as chemotherapy) is represented by $\xi \in [0, 1]$, drawn from a beta distribution. $\{\lambda, \psi, \phi\} \in \mathcal{R}^3_+$ represent the parameters of the simulator. We also define $x_{n,m} = \{\lambda, \psi, \phi\, K_0, \xi\}$ and $z_{n,m} = \{x_{n,m}, T_{\text{opp}}, t_{\text{max}}\}$, where all but ξ are set to constant values. Expanding this to condition all values on y_n is trivial given domain knowledge. Alternatively, they could also be drawn at random, but not be conditioned on y_n. Such relations are omitted here for simplicity.

We can now fully define ϕ as:

$$\phi(y_n, z_{n,m}) = \mathbb{1}(c(t_{\text{max}}, x_{n,m}) < T_{\text{opp}}). \tag{71}$$

Taking the expectation of ϕ over M different realizations of z yields the estimate $(\hat{\gamma}_M)_n$. This value is the probability that treatment will be successful for a particular patient, marginalizing over possible unobserved dynamics. This is the point at which clinician decides whether initiate the treatment plan. This decision is represented $f(y_n, (\hat{\gamma}_M)_n) \in [0, 1]$ as:

$$f(y_n, (\hat{\gamma}_M)_n) = \mathbb{1}((\hat{\gamma}_M)_n > T_{\text{treat}}) \tag{72}$$

where T_{treat} is the minimum probability of success required for that patient to receive the treatment, and again, could be conditioned on y also. Taking the expectation of f over patients yields the expected frequency with which the treatment will be delivered, given a value of T_{treat}. The hospital wishes to estimate the value T_{treat} that maximizes the number of patients treated, while only treating those patients with the highest probability of success, and (in expectation) staying within the budgetary constraint.

The model is completed by the definition of the following distributions and parameters.

$$K_0 = 10^8, \quad \phi = 0.001, \quad \psi = 0.05, \quad \lambda = 0.5, \quad \xi \sim \text{Beta}(5, 2),$$
$$c_0 \sim 1000 * \text{Rayleigh}(10), \quad T_{\text{opp}} = 2000, \quad T_{\text{treat}} = 0.35, \quad t_{\text{max}} = 250, \quad t_{\text{step}} = 0.01$$

H.2 Budget result

In the example outlined above, the treatment center is not actually attempting to evaluate the value of I, but to find the optimal value of T_{treat} subject to a budgetary constraint. A simplistic way of evaluating the optimal value is to perform a dense search over different values of the parameter, each time evaluating the estimated expenditure, and select the best performing value.

Figure 7 shows the variation of predicted expenditure against the threshold probability, as well as the budget constraint. The intersection of these curves is the optimal setting of T_{opp}, here evaluated to be 12.5%. From the blue line, it is clear that the relationship between expenditure and treatment probability is non-linear, especially at the extrema of the distribution, and hence the use of NMC was necessarily for evaluating the optimal value.

Appendix I Bayesian Experimental Design

Bayesian experimental design provides a framework for designing experiments in a manner that is optimal from an information-theoretic viewpoint (Chaloner and Verdinelli, 1995; Sebastiani and Wynn, 2000). By minimizing the entropy in the posterior distribution of the parameters of interest, one can maximize the information gathered by the experiment.

Let the parameters of interest be denoted by $\theta \in \Theta$ for which we define a prior distribution $p(\theta)$. Let the probability of achieving outcome $y \in \mathcal{Y}$, given parameters θ and a design $d \in \mathcal{D}$, be defined by likelihood model $p(y|\theta, d)$. Under our model, the outcome of the experiment given a chosen d is distributed according to

$$p(y|d) = \int_\Theta p(y, \theta|d) d\theta = \int_\Theta p(y|\theta, d) p(\theta) d\theta. \tag{73}$$

where we have used the fact that $p(\theta) = p(\theta|d)$ because θ is independent of the design. Our aim is to choose the optimal design d under some criterion. We, therefore, define a utility function, $U(y, d)$, representing the utility of choosing a design d and getting a response y. Typically our aim is to maximize information gathered from the experiment, and so we set $U(y, d)$ to be the gain in Shannon information between the prior and the posterior:

$$U(y, d) = \int_\Theta p(\theta|y, d) \log(p(\theta|y, d)) d\theta - \int_\Theta p(\theta) \log(p(\theta)) d\theta \tag{74}$$

However, we are still uncertain about the outcome. Thus, we use the expectation of $U(y, d)$ with respect to $p(y|d)$ as our target:

$$\begin{aligned}
\bar{U}(d) &= \int_\mathcal{Y} U(y, d) p(y|d) dy \\
&= \int_\mathcal{Y} \int_\Theta p(y, \theta|d) \log(p(\theta|y, d)) d\theta dy - \int_\Theta p(\theta) \log(p(\theta)) d\theta \\
&= \int_\mathcal{Y} \int_\Theta p(y, \theta|d) \log\left(\frac{p(\theta|y, d)}{p(\theta)}\right) d\theta dy.
\end{aligned} \tag{75}$$

noting that this corresponds to the mutual information between the parameters θ and the observations y. The Bayesian-optimal design is then given by

$$d^* = \underset{d \in \mathcal{D}}{\operatorname{argmax}} \bar{U}(d). \tag{76}$$

Finding d^* is challenging because the posterior $p(\theta|y, d)$ is rarely known in closed form. To solve the problem, we proceed by rearranging (75) using Bayes' rule (remembering that $p(\theta) = p(\theta|d)$):

$$\begin{aligned}
\bar{U}(d) &= \int_\mathcal{Y} \int_\Theta p(y, \theta|d) \log\left(\frac{p(\theta|y, d)}{p(\theta)}\right) d\theta dy \\
&= \int_\mathcal{Y} \int_\Theta p(y, \theta|d) \log\left(\frac{p(y|\theta, d)}{p(y|d)}\right) d\theta dy \\
&= \int_\mathcal{Y} \int_\Theta p(y, \theta|d) \log(p(y|\theta, d)) d\theta dy - \int_\mathcal{Y} p(y|d) \log(p(y|d)) dy.
\end{aligned} \tag{77}$$

The first of these terms can now be evaluated using standard MC approaches as the integrand is analytic. In contrast, the second term is not directly amenable to standard MC estimation as the marginal $p(y|d)$ represents an expectation and taking its logarithm represents a non-linear functional mapping.

To derive an estimator, we will now consider these terms separately. Starting with the first term,

$$\bar{U}_1(d) = \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log(p(y|\theta, d)) d\theta dy \approx \frac{1}{N} \sum_{n=1}^{N} \log(p(y_n|\theta_n, d)) \tag{78}$$

where $\theta_n \sim p(\theta)$ and $y_n \sim p(y|\theta = \theta_n, d)$. We note that evaluating (78) involves both sampling from $p(y|\theta, d)$ and directly evaluating it point-wise. The latter of these cannot be avoided, but in the scenario where we do not have direct access to a sampler for $p(y|\theta, d)$, we can use the standard importance sampling trick, sampling instead $y_n \sim q(y|\theta = \theta_n, d)$ and weighting the samples in (78) by $w_n = \frac{p(y_n|\theta_n,d)}{q(y_n|\theta_n,d)}$.

Now considering the second term we have

$$\bar{U}_2(d) = \int_{\mathcal{Y}} p(y|d) \log(p(y|d)) dy \approx \frac{1}{N} \sum_{n=1}^{N} \log \left(\frac{1}{M} \sum_{m=1}^{M} p(y_n|\theta_{n,m}, d) \right) \tag{79}$$

where $\theta_{n,m} \sim p(\theta)$ and $y_n \sim p(y|d)$. Here we can sample the latter by first sampling an otherwise unused $\theta_{n,0} \sim p(\theta)$ and then sampling $y_n \sim p(y|\theta_{n,0}, d)$. Again we can use importance sampling if we do not have direct access to a sampler for $p(y|\theta_{n,0}, d)$.

Putting (78) and (79) together (and renaming θ_n from (78) as $\theta_{n,0}$ for notational consistency with (79)) we now have the following complete estimator given in the main paper and implicitly used by (Myung et al., 2013) amongst others

$$\bar{U}(d) \approx \frac{1}{N} \sum_{n=1}^{N} \left[\log(p(y_n|\theta_{n,0}, d)) - \log \left(\frac{1}{M} \sum_{m=1}^{M} p(y_n|\theta_{n,m}, d) \right) \right] \tag{80}$$

where $\theta_{n,m} \sim p(\theta)\ \forall m \in 0 : M,\ n \in 1 : N$ and $y_n \sim p(y|\theta = \theta_{n,0}, d)\ \forall n \in 1 : N$.

We now show that if y can only take on one of C possible values $(y_1, \ldots, y_C)$, we can achieve significant improvements in the convergence rate by using a similar to that introduced in Section 3.2 to convert to single MC estimator:

$$\begin{aligned}
\bar{U}(d) &= \int_{\mathcal{Y}} \int_{\Theta} p(y, \theta|d) \log(p(y|\theta, d)) d\theta dy - \int_{\mathcal{Y}} p(y|d) \log(p(y|d)) dy \\
&= \int_{\Theta} \left[\sum_{c=1}^{C} p(\theta) p(y_c|\theta, d) \log(p(y_c|\theta, d)) \right] d\theta - \sum_{c=1}^{C} p(y_c|d) \log(p(y_c|d)) \\
&\approx \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} p(y_c|\theta_n, d) \log\left(p(y_c|\theta_n, d)\right) - \sum_{c=1}^{C} \left[\left(\frac{1}{N} \sum_{n=1}^{N} p(y_c|\theta_n, d) \right) \log \left(\frac{1}{N} \sum_{n=1}^{N} p(y_c|\theta_n, d) \right) \right]
\end{aligned} \tag{81}$$

where $\theta_n \sim p(\theta) \quad \forall n \in 1, \ldots, N$. As C is a fixed constant, the MSE for first term clearly converges at the standard MC error rate of $O(1/N)$. Similarly each $\hat{P}_N(y_c|d) = \frac{1}{N} \sum_{n=1}^{N} p(y_c|\theta_n, d)$ term also converges at a rate $O(1/N)$ to $p(y_c|d)$. Now noting that $\hat{P}_N(y_c|d) \leq 1$ and that $f(x) = x \log x$ is Lipschitz continuous in the range $(0, 1]$, each $\hat{P}_N(y_c|d) \log\left(\hat{P}_N(y_c|d)\right)$ term must also converge at the MC error rate if $p(y_c|d) > 0\ \forall c = 1, \ldots, C$. Finally if we assume that when $p(y_c|d) = 0$ then $\hat{P}_N(y_c|d) = 0$ almost surely for sufficiently large N, then the second term also converges at the MC error when $p(y_c|d) = 0$. We now have a finite sum of terms which each convergence to $\bar{U}(d)$ with MC MSE rate $O(1/N)$, and so the overall estimator (81) must also converge at this rate. This compares to $O(1/T^{2/3})$ for (80) (assuming we take $N \propto M^2$), noting that generating T samples for (80) has the same cost up to a constant factor as generating N for (81). To the best of our knowledge, this is the first introduction of this superior estimator in the literature.

We finish by showing that the theoretical advantages of this reformulation also leads to empirical gains in the estimation of $\bar{U}(d)$. For this, we consider a model used in psychology experiments for delay discounting introduced by (Vincent, 2016; Vincent and Rainforth, 2017). Our experiment comprises of asking questions of the form *"Would you prefer £A now, or £B in D days?"* and we wish to choose the question variables $d = \{A, B, D\}$ in the manner that will give the most incisive questions. The target participant is presumed to have parameters $\theta = \{k, \alpha\}$ and the following response model

$$y \sim \text{Bernoulli}\left(0.01 + 0.98 \cdot \Phi\left(\frac{1}{\alpha}\left(\frac{B}{1 + e^k D} - A\right)\right)\right) \tag{82}$$

where $y = 1$ indicates choosing the delayed response and Φ represents the cumulative normal distribution. As more questions are asked, the distribution over the parameters θ is updated, such that the most optimal question to ask at a particular time depends on the previous questions and responses. For the sake of brevity, when comparing the performance

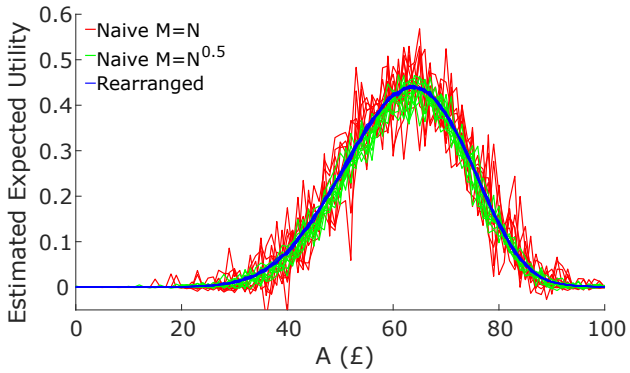


Figure 8. Estimated expected utilities $\bar{U}(d)$for different values of one of the design parameters $A \in \{1, 2, \ldots, 100\}$ given a fixed total sample budget of $T = 10^4$. Here the lines correspond to 10 independent runs, showing that the variance of (80) is far higher than (81).

of (80) and (81) we will neglect the problem of how best to optimize the design, and consider only the problem of evaluating $\bar{U}(d)$. We will further consider the case where $B = 100$ and $D = 50$ are fixed and we are only choosing the delayed value A. We presume the following distribution on the parameters

$$k \sim \mathcal{N}(-4.5, 0.5^2)$$
$$\alpha \sim \Gamma(2, 2).$$

We first consider convergence in the estimate of $\bar{U}(d)$ for the case $A = 70$ for our suggested method (81) and the naïve solution (80), the results of which are shown in Figure 2a in the main paper. Here we see that the convergence rates of the two methods are both as expected and that our suggested method offers significant empirical performance improvements.

We next consider setting a total sample budget $T = 10^4$ and look at the variation in the estimated values of $\bar{U}(d)$ for different values of A for the two methods as shown in Figure 8. This shows that the improvement in MSE leads to clearly visible improvements in the characterization of $\bar{U}(d)$ that will translate to improvements in seeking the optimum.

Acknowledgements

Tom Rainforth'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. However, the majority of this work was undertaken while he was in the Department of Engineering Science, University of Oxford, and was supported by a BP industrial grant. Robert Cornish is supported by an NVIDIA scholarship. Hongseok Yang is supported by an Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.R0190-16-2011, Development of Vulnerability Discovery Technologies for IoT Software Security). Frank Wood is supported under DARPA PPAML through the U.S. AFRL under Cooperative Agreement FA8750-14-2-0006, Sub Award number 61160290-111668.

References

P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47, 2016.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2010.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3): 1139–1160, 2003.

D. Belomestny, A. Kolodko, and J. Schoenmakers. Regression methods for stochastic control problems and their convergence analysis. *SIAM Journal on Control and Optimization*, 2010.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

M. Broadie, Y. Du, and C. C. Moallemi. Efficient risk estimation via nested sequential simulation. *Management Science*, 2011.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 1995.

K. Csilléry, M. G. Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

A. Doucet, N. De Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

R. Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

H. Enderling and M. A. Chaplain. Mathematical modeling of tumor growth and treatment. *Current Pharmaceutical Design*, 20(30):4934–4940, 2014. ISSN 1381-6128/1873-4286.

G. Fort, E. Gobet, and E. Moulines. MCMC design-based non-parametric regression for rare-event. application to nested risk computations. *Monte Carlo Methods Appl*, 2017.

M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *MCMC in practice*. CRC press, 1995.

T. Goda. Computing the variance of a conditional expectation via non-nested Monte Carlo. *Operations Research Letters*, 2016.

N. Goodman, V. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. *UAI*, 2008.

M. B. Gordy and S. Juneja. Nested simulation in portfolio risk measurement. *Management Science*, 2010.

S. Heinrich. Multilevel Monte Carlo methods. *LSSC*, 1:58–67, 2001.

M. Hoffman and D. Blei. Stochastic structured variational inference. In *AISTATS*, 2015.

L. J. Hong and S. Juneja. Estimating the mean of a non-linear function of conditional expectation. In *Winter Simulation Conference*, 2009.

P. E. Jacob, A. H. Thiery, et al. On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

T. A. Le, A. G. Baydin, and F. Wood. Nested compiled inference for hierarchical reinforcement learning. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

T. A. Le, M. Igl, T. Rainforth, T. Jin, and F. Wood. Auto-encoding sequential Monte Carlo. In *ICLR*, 2018.

V. Lemaire, G. Pagès, et al. Multilevel Richardson–Romberg extrapolation. *Bernoulli*, 23(4A):2643–2692, 2017.

F. Liang. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

F. A. Longstaff and E. S. Schwartz. Valuing American options by simulation: a simple least-squares approach. *Review of Financial studies*, 2001.

A.-M. Lyne, M. Girolami, Y. Atchade, H. Strathmann, D. Simpson, et al. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015.

C. J. Maddison, D. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. W. Teh. Filtering variational objectives. *arXiv preprint arXiv:1705.09279*, 2017.

T. Mantadelis and G. Janssens. Nesting probabilistic inference. *arXiv preprint arXiv:1112.3785*, 2011.

F. J. Medina-Aguayo, A. Lee, and G. O. Roberts. Stability of noisy Metropolis–Hastings. *Statistics and Computing*, 26(6): 1187–1211, 2016.

I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2006.

J. I. Myung, D. R. Cavagnaro, and M. A. Pitt. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3):53–67, 2013.

C. A. Naesseth, F. Lindsten, and T. Schön. Nested sequential Monte Carlo methods. In *ICML*, 2015.

C. A. Naesseth, S. W. Linderman, R. Ranganath, and D. M. Blei. Variational sequential Monte Carlo. *arXiv preprint arXiv:1705.11140*, 2017.

A. O'Hagan. Bayes–Hermite quadrature. *Journal of statistical planning and inference*, 1991.

L. Ouyang, M. H. Tessler, D. Ly, and N. Goodman. Practical optimal experiment design with probabilistic programs. *arXiv preprint arXiv:1608.05046*, 2016.

G. Pagés. Multi-step Richardson–Romberg extrapolation: remarks on variance control and complexity. *Monte Carlo Methods and Applications*, 13(1):37, 2007.

T. Rainforth. *Automating Inference, Learning, and Design using Probabilistic Programming*. PhD thesis, 2017.

T. Rainforth. Nesting probabilistic programs. In *UAI*, 2018.

T. Rainforth, T. A. Le, J.-W. van de Meent, M. A. Osborne, and F. Wood. Bayesian optimization for probabilistic programs. In *NIPS*, 2016.

T. Rainforth, A. R. Kosiorek, T. A. Le, C. J. Maddison, M. Igl, F. Wood, and Y. W. Teh. Tighter variational bounds are not necessarily better. In *ICML*, 2018.

D. Rudolf and N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *arXiv preprint arXiv:1503.04123*, 2015.

P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

A. Stuhlmüller and N. D. Goodman. A dynamic programming algorithm for inference in recursive probabilistic programs. In *Second Statistical Relational AI workshop at UAI 2012 (StaRAI-12)*, 2012.

A. Stuhlmüller and N. D. Goodman. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99, 2014.

B. T. Vincent. Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. *Behavior research methods*, 48(4):1608–1620, 2016.

B. T. Vincent and T. Rainforth. The DARC toolbox: automated, flexible, and efficient delayed and risky choice experiments using Bayesian adaptive design. *PsyArXiv*, 2017.

F. Wood, J. W. van de Meent, and V. Mansinghka. A new approach to probabilistic programming inference. In *AISTATS*, pages 2–46, 2014.

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	On Nesting Monte Carlo Estimators
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Rainforth, T., Cornish, R., Yang, H., Warrington, A., Wood, F. (2018). On Nesting Monte Carlo Estimators. In *Proceedings of the 35th International Conference on Machine Learning,* Stockholm, Sweden, PMLR 80, 2018.

Student Confirmation

Student Name:	Rob Cornish		
Contribution to the Paper	Formalised some initial ideas about nested estimation made by Tom Rainforth (first author) into precise mathematical statements. Formulated and proved rigorously all results (except Theorem 5) of the paper, some which were subsequently developed further collaboratively with Tom Rainforth. Co-wrote the initial paper draft along with Tom Rainforth.		
Signature	Robert Cornish	Date	17/01/2020

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: Prof. Arnaud Doucet			
Supervisor comments	I do certify that the candidate made a substantial contribution to the publication, and the description above is accurate.		
Signature		Date	17/01/2020

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

4

Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows

Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows

Rob Cornish [1] **Anthony Caterini** [1] **George Deligiannidis** [1 2] **Arnaud Doucet** [1]

Abstract

We show that normalising flows become pathological when used to model targets whose supports have complicated topologies. In this scenario, we prove that a flow must become arbitrarily numerically noninvertible in order to approximate the target closely. This result has implications for all flow-based models, and especially *residual flows* (ResFlows), which explicitly control the Lipschitz constant of the bijection used. To address this, we propose *continuously indexed flows* (CIFs), which replace the single bijection used by normalising flows with a continuously indexed family of bijections, and which can intuitively "clean up" mass that would otherwise be misplaced by a single bijection. We show theoretically that CIFs are not subject to the same topological limitations as normalising flows, and obtain better empirical performance on a variety of models and benchmarks.

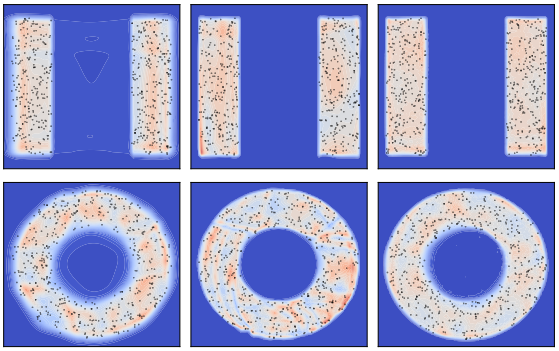


Figure 1: Densities learned by a 10-layer ResFlow (left), 100-layer ResFlow (middle), and 10-layer CIF-ResFlow (right) for two datasets (samples shown in black) that are not homeomorphic to the Gaussian prior. The 10-layer ResFlow visibly leaks mass outside of the support of the target due to its small bi-Lipschitz constant. The larger ResFlow improves on this, but still achieves smaller average log probability than the CIF-ResFlow, as is apparent from the greater homogeneity of the right-hand densities.

1 Introduction

Normalising flows (Rezende & Mohamed, 2015) have become popular methods for density estimation (Dinh et al., 2017; Papamakarios et al., 2017; Kingma & Dhariwal, 2018; Chen et al., 2019). These methods model an unknown target distribution $P_X^\star$ on a data space $\mathcal{X} \subseteq \mathbb{R}^d$ as the marginal of X obtained by the generative process

$$Z \sim P_Z, \quad X := f(Z), \tag{1}$$

where P_Z is a *prior* distribution on a space $\mathcal{Z} \subseteq \mathbb{R}^d$, and $f : \mathcal{X} \to \mathcal{Z}$ is a bijection. The use of a bijection means the density of X can be computed analytically by the change-of-variables formula, and the parameters of f can be learned by maximum likelihood using i.i.d. samples from $P_X^\star$.

To be effective, a normalising flow model must specify an expressive family of bijections with tractable Jacobians. Affine coupling layers (Dinh et al., 2015; 2017), autoregressive maps (Germain et al., 2015; Papamakarios et al., 2017), invertible linear transformations (Kingma & Dhariwal, 2018), ODE-based maps (Grathwohl et al., 2019), and invertible ResNet blocks (Behrmann et al., 2019; Chen et al., 2019) are all examples of such bijections that can be composed to produce expressive flows. These models have demonstrated significant promise in their ability to model complex datasets and to synthesise realistic data.

In all these cases, f and f^{-1} are both continuous. It follows that f is a *homeomorphism*, and therefore preserves the topology of its domain (Runde, 2007, Definition 3.3.10). As Dupont et al. (2019) and Dinh et al. (2019) mention, this seems intuitively problematic when P_Z and $P_X^\star$ are supported on domains with distinct topologies, which occurs for example when the supports differ in their number of connected components or "holes", or when they are "knotted" differently. This seems inevitable in practice, as P_Z is usually quite simple (e.g. a Gaussian) while $P_X^\star$ is very complicated (e.g. a distribution over images).

As our first contribution, we make precise the consequences

[1]University of Oxford, Oxford, United Kingdom [2]The Alan Turing Institute, London, United Kingdom. Correspondence to: Rob Cornish <rob.cornish@eng.ox.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

of using a topologically misspecified prior. We confirm that in this case it is indeed impossible to recover the target perfectly if f is a homeomorphism. Moreover, in Theorem 2.1 we prove that, in order to *approximate* such a target arbitrarily well, we must have $\mathrm{BiLip}\, f \to \infty$, where $\mathrm{BiLip}\, f$ denotes the *bi-Lipschitz constant* of f defined as the infimum over $M \in [1, \infty]$ such that

$$M^{-1}\|z - z'\| \leq \|f(z) - f(z')\| \leq M\|z - z'\| \qquad (2)$$

for all $z, z' \in \mathcal{Z}$. Theorem 2.1 applies essentially regardless of the training objective, and has implications for the case that P_Z and $P_X^\star$ both have full support but are heavily concentrated on regions that are not homeomorphic. Since $\mathrm{BiLip}\, f$ is a natural measure of the "invertibility" of f (Behrmann et al., 2020), this result shows that the goal of designing neural networks with well-conditioned inverses is fundamentally at odds with the goal of designing neural networks that can approximate complicated densities.

Theorem 2.1 also has immediate implications for *residual flows* (ResFlows) (Behrmann et al., 2019; Chen et al., 2019), which have recently achieved state-of-the-art performance on several large-scale density estimation tasks. Unlike models based on triangular maps (Jaini et al., 2019), ResFlows have the attractive feature that the structure of their Jacobians is unconstrained, which may explain their greater expressiveness. However, as part of the construction, the bi-Lipschitz constant of f is bounded, and so these models must be composed many times in order to achieve overall the large bi-Lipschitz constant required for a complex $P_X^\star$.[1]

To address this problem we introduce *continuously indexed flows* (CIFs), which generalise (1) by replacing the single bijection f with an indexed family of bijections $\{F(\cdot; u)\}_{u \in \mathcal{U}}$, where the index set $\mathcal{U}$ is continuous. Intuitively, CIFs allow mass that would be erroneously placed by a single bijection to be rerouted into a more optimal location. We show that CIFs can learn the support of a given $P_X^\star$ exactly regardless of the topology of the prior, and without the bi-Lipschitz constant of any $F(\cdot; u)$ necessarily becoming infinite. CIFs do not specify the form of F, and can be used in conjunction with any standard normalising flow architecture directly.

Our use of a continuous index overcomes several limitations associated with alternative approaches based on a discrete index (Dinh et al., 2019; Duan, 2019), which suffer either from a discontinuous loss landscape or an intractable computational complexity. However, as a consequence, we sacrifice the ability to compute the likelihood of our model analytically. To address this, we propose a variational approximation that exploits the bijective structure of the model and is suitable for training large-scale models in practice. We empirically evaluate CIFs applied to ResFlows, neural spline flows (NSFs) (Durkan et al., 2019), masked autoregressive flows (MAFs) (Papamakarios et al., 2017), and RealNVPs (Dinh et al., 2017), obtaining improved performance in all cases. We observe a particular benefit for ResFlows: with a 10-layer CIF-ResFlow we surpass the performance of a 100-layer baseline ResFlow and achieve state-of-the-art results on several benchmark datasets.

[1] Chen et al. (2019) report using 100-200 layers to learn even simple 2D densities.

2 Bi-Lipschitz Constraints on Pushforwards

Normalising flows fall into a larger class of density estimators based on *pushforwards*. Given a prior measure P_Z on $\mathcal{Z}$ and a mapping $f : \mathcal{Z} \to \mathcal{X}$, these models are defined as

$$P_X := f \# P_Z,$$

where the right-hand side denotes a distribution with $f\#P_Z(B) := P_Z(f^{-1}(B))$ for Borel $B \subseteq \mathcal{X}$. Normalising flows take f to be bijective, which under sufficient regularity yields a closed-form expression for the density[2] of P_X (Billingsley, 2008, Theorem 17.2).

Intuitively, the pushforward map f *transports* the mass allocated by P_Z into $\mathcal{X}$-space, thereby defining P_X based on where each unit of mass ends up. This imposes a global constraint on f if P_X is to match perfectly a given target $P_X^\star$. In particular, denote by $\operatorname{supp} P_Z$ the *support* of P_Z. While the precise definition of the support involves topological formalities (see Section B.1 in the Supplement), intuitively this set defines the region of $\mathcal{Z}$ to which P_Z assigns mass. It is then straightforward to show that $P_X = P_X^\star$ only if

$$\operatorname{supp} P_X^\star = \overline{f(\operatorname{supp} P_Z)}, \qquad (3)$$

where $\overline{A}$ denotes the closure of A in $\mathcal{X}$.[3]

The constraint (3) is especially onerous for normalising flows because of their bijectivity. In practice, f and f^{-1} are invariably both continuous, and so f is a *homeomorphism*. Consequently, for these models (3) entails[4]

$$\operatorname{supp} P_X = \operatorname{supp} P_X^\star \text{ only if } \operatorname{supp} P_Z \cong \operatorname{supp} P_X^\star, \quad (4)$$

where $\mathcal{A} \cong \mathcal{B}$ means that $\mathcal{A}$ and $\mathcal{B}$ are *homeomorphic*, i.e. isomorphic as topological spaces (Runde, 2007, Definition 3.3.10). This means that $\operatorname{supp} P_Z$ and $\operatorname{supp} P_X^\star$ must exactly share *all* topological properties, including number of connected components, number of "holes", the way they are "knotted", etc., in order to learn the target perfectly. Condition (4) therefore suggests that normalising flows are not optimally suited to the task of learning complex real-world densities, where such topological mismatch seems inevitable.

[2] Throughout, by "density" we mean Lebesgue density. We will write densities using lowercase, e.g. p_X for the measure P_X.

[3] See Proposition B.3 in the Supplement for a proof.

[4] Note that $\overline{f(\operatorname{supp} P_Z)} = f(\operatorname{supp} P_Z)$ here since $\operatorname{supp} P_Z$ is closed by Proposition B.2 in the Supplement.

However, (4) only rules out the limiting case $P_X = P_X^\star$. In practice it is likely enough to have $P_X \approx P_X^\star$, and it is therefore relevant to consider the implications of a topologically misspecified prior in this case also. Intuitively, this seems to require f become *almost* nonbijective as P_X approaches $P_X^\star$, but it is not immediately clear what this means, or whether this must occur for all models. Likewise, in practice it might be reasonable to assume the density of $P_X^\star$ is everywhere strictly positive. In this case, even if $P_X^\star$ is *concentrated* on some very complicated set, the constraint (4) would trivially be met if P_Z is Gaussian, for example. Nevertheless, it seems that infinitesimal regions of mass should not significantly change the behaviour required of f, and we would therefore like to extend (4) to apply here also.

The bi-Lipschitz constant (2) naturally quantifies the "invertibility" of f. Behrmann et al. (2020) recently showed a relationship between the bi-Lipschitz constant and the *numerical* invertibility of f. If f is injective and differentiable,

$$\operatorname{BiLip} f = \max\left(\sup_{z\in\mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}},\ \sup_{x\in f(\mathcal{Z})} \|\mathrm{D}f^{-1}(x)\|_{\mathrm{op}}\right),$$

where $\mathrm{D}g(y)$ is the Jacobian of g at y and $\|\cdot\|_{\mathrm{op}}$ is the operator norm. A large bi-Lipschitz constant thus means f or f^{-1} "jumps" somewhere in its domain. More generally, if f is not injective, then $\operatorname{BiLip} f = \infty$, while if $\operatorname{BiLip} f < \infty$, then f is a homeomorphism from $\mathcal{Z}$ to $f(\mathcal{Z})$.[5]

The following theorem shows that if the supports of P_Z and $P_X^\star$ are not homeomorphic, then the bi-Lipschitz constant of f must grow arbitrarily large in order to approximate $P_X^\star$. Here $\xrightarrow{\mathcal{D}}$ denotes weak convergence.

Theorem 2.1. *Suppose P_Z and $P_X^\star$ are probability measures on $\mathbb{R}^{d_\mathcal{Z}}$ and $\mathbb{R}^{d_\mathcal{X}}$ respectively, and that $\operatorname{supp} P_Z \ncong \operatorname{supp} P_X^\star$. Then for any sequence of measurable $f_n : \mathbb{R}^{d_\mathcal{Z}} \to \mathbb{R}^{d_\mathcal{X}}$, we can have $f_n \# P_Z \xrightarrow{\mathcal{D}} P_X^\star$ only if*

$$\lim_{n\to\infty} \operatorname{BiLip} f_n = \infty.$$

Weak convergence is implied by the minimisation of all standard statistical divergences used to train generative models, including the KL and Jensen-Shannon divergences and the Wasserstein metric (Arjovsky et al., 2017, Theorem 2). Thus, Theorem 2.1 states that these quantities can vanish only if the bi-Lipschitz constant of the learned mapping becomes arbitrarily large. Likewise, note that we do not assume $d_\mathcal{Z} = d_\mathcal{X}$ so that this result also applies to injective flow models (Kumar et al., 2019), as well as other pushforward-based models such as GANs (Goodfellow et al., 2014).[6]

[5] See Section B.2 in the Supplement for proofs.

[6] However, the implications for GANs seem less problematic since a GAN generator is not usually assumed to be bijective.

Theorem 2.1 also applies when $\operatorname{supp} P_Z$ is *almost* not homeomorphic to $\operatorname{supp} P_X^\star$, as is made precise by the following corollary. Here ρ denotes any metric for the weak topology; see Chapter 6 of Villani (2008) for standard examples.

Corollary 2.2. *Suppose P_Z is a probability measure on $\mathbb{R}^{d_\mathcal{Z}}$, $P_X^\star$ and P_0 are probability measures on $\mathbb{R}^{d_\mathcal{X}}$, and that $\operatorname{supp} P_0 \ncong \operatorname{supp} P_Z$. Then there exists nonincreasing $M : [0,\infty) \to [1,\infty]$ with $M(\epsilon) \to \infty$ as $\epsilon \to 0$ such that $\operatorname{BiLip} f \geq M(\epsilon)$ if $\min\left(\rho(f\#P_Z, P_X^\star), \rho(P_X^\star, P_0)\right) \leq \epsilon$.*

In other words, if the target is close to a probability measure with non-homeomorphic support to that of the prior P_Z (i.e. $\rho(P_X^\star, P_0)$ is small), and if the model is a good approximation of the target (i.e. $\rho(f\#P_Z, P_X^\star)$ is small), then the Bi-Lipschitz constant of f must be large.

Proofs of these results are in Section B.3 of the Supplement.

2.1 Practical Implications

The results of this section indicate a limitation of existing flow-based density models. This is most direct for *residual flows* (ResFlows) (Behrmann et al., 2019; Chen et al., 2019), which take $f = f_1 \circ \cdots \circ f_L$ with each layer of the form

$$f_\ell^{-1}(x) = x + g_\ell(x), \qquad \operatorname{Lip} g_\ell \leq \kappa < 1. \tag{5}$$

Here Lip denotes the Lipschitz constant, which is bounded by a fixed constant κ throughout training. The Lipschitz constraint is enforced by spectral normalisation (Miyato et al., 2018; Gouk et al., 2018) and ensures each f_ℓ is bijective. However, it also follows (Behrmann et al., 2019, Lemma 2) that

$$\operatorname{BiLip} f \leq \max(1+\kappa, (1-\kappa)^{-1})^L < \infty, \tag{6}$$

and Theorem 2.1 thus restricts how well a ResFlow can approximate $P_X^\star$ with non-homeomorphic support to P_Z. Figure 1 illustrates this in practice for simple 2-D examples.

It is possible to relax (6) by taking $\kappa \to 1$. However, this can have a detrimental effect on the variance of the Russian roulette estimator (Kahn, 1955) used by Chen et al. (2019) to compute the Jacobian, and in Section B.4 of the Supplement we give a simple example in which the variance is in fact infinite. Alternatively, we can also loosen the bound (6) by taking $L \to \infty$, and Figure 1 shows that this does indeed lead to better performance. However, greater depth means greater computational cost. In the next section we describe an alternative approach that allows relaxing the bi-Lipschitz constraint of Theorem 2.1 without modifying either κ or L, and thus avoids these potential issues.

Unlike ResFlows, most normalising flows used in practice have an unconstrained bi-Lipschitz constant (Behrmann et al., 2020). As as result, Theorem 2.1 does not prevent these models from approximating non-homeomorphic targets arbitrarily well, and indeed several architectures have

been proposed that can in principle do so (Huang et al., 2018; Jaini et al., 2019). Nevertheless, the constraint (4) shows that these models still face an underlying limitation in practice, and suggests we may improve performance more generally by relaxing the requirement of bijectivity. We verify empirically in Section 5 that, in addition to ResFlows, our proposed method also yields benefits for flows without an explicit bi-Lipschitz constraint.

Finally, Theorem 2.1 has implications for the numerical stability of normalising flows. It was recently pointed out by Behrmann et al. (2020) that, while having a well-defined mathematical inverse, many common flows can become *numerically* noninvertible over the course of training, leading to low-quality reconstructions and calling into question the accuracy of density values output by the change-of-variables formula. Behrmann et al. (2020) suggest explicitly constraining $\mathrm{BiLip}\ f$ in order to avoid this problem. Theorem 2.1 shows that this involves a fundamental tradeoff against expressivity: if greater numerical stability is required of our normalising flow, then we must necessarily reduce the set of targets we can represent arbitrarily well.

3 Continuously Indexed Flows

In this section we propose *continuously indexed flows* (CIFs) for relaxing the bijectivity of standard normalising flows. We begin by defining the model we consider, and then detail our suggested training and inference procedures. In the next section we discuss advantages over related approaches.

3.1 Model Specification

CIFs are obtained by replacing the single bijection f used by normalising flows with an indexed family $\{F(\cdot; u)\}_{u \in \mathcal{U}}$, where $\mathcal{U} \subseteq \mathbb{R}^{d_\mathcal{U}}$ is our index set and each $F(\cdot; u) : \mathcal{Z} \to \mathcal{X}$ is a bijection. We then define the model P_X as the marginal of X obtained from the following generative process:

$$Z \sim P_Z, \quad U \sim P_{U|Z}(\cdot|Z), \quad X := F(Z; U). \tag{7}$$

Like (1), we assume a prior P_Z on $\mathcal{Z}$, but now also require conditional distributions $P_{U|Z}(\cdot|z)$ on $\mathcal{U}$ for each $z \in \mathcal{Z}$.

We can increase the complexity of (7) by taking P_Z itself to have the same form. This is directly analogous to the standard practice of composing simple bijections to obtain a richer class of normalising flows. In our context, stacking L layers of (7) corresponds to the generative process

$$Z_0 \sim P_{Z_0},\ U_\ell \sim P_{U_\ell|Z_{\ell-1}}(\cdot|Z_{\ell-1}),\ Z_\ell := F_\ell(Z_{\ell-1}; U_\ell), \tag{8}$$

where $\ell \in \{1, \dots, L\}$. We then take P_X to be the marginal of $X := Z_L$. We have found this construction to improve significantly the expressiveness of our models and make extensive use of it in our experiments below. Note that this corresponds to an instance of (7) where, defining $F^\ell(\cdot; u_1, \dots, u_\ell) := F_\ell(\cdot; u_\ell) \circ \cdots \circ F_1(\cdot; u_1)$, we take $Z = Z_0$, $U = (U_1, \dots, U_L)$, $P_{U|Z}(\mathrm{d}u|z) = \prod_\ell P_{U_\ell|Z_{\ell-1}}(\mathrm{d}u_\ell|F^\ell(z; u_1, \dots, u_\ell))$, and $F = F^L$. We use this to streamline some of the discussion below.

Previous works, most notably RAD (Dinh et al., 2019), have considered related models with a discrete index set $\mathcal{U}$. We instead consider a *continuous* index. In particular, our $\mathcal{U}$ will be an open subset of $\mathbb{R}^{d_\mathcal{U}}$, with each $P_{U|Z}(\cdot|z)$ having a density $p_{U|Z}(\cdot|z)$. A continuous index confers various advantages that we describe in Section 4. The choice also requires a distinct approach to training and inference that we describe in Section 3.2.

We require choices of $p_{U|Z}$ and F for each layer of our model. Straightforward possibilities are

$$F(z; u) = f\left(e^{-s(u)} \odot z - t(u)\right) \tag{9}$$

$$p_{U|Z}(\cdot|z) = \mathrm{Normal}(\mu^p(z), \Sigma^p(z)) \tag{10}$$

for any bijection f (e.g. a ResFlow step) and appropriately defined neural networks s, t, μ^p, and Σ^p.[7] Here the exponential of a vector is meant elementwise, and $\odot$ denotes elementwise multiplication. Note that (9) may be used with all existing normalising flow implementations out-of-the-box. These choices yielded strong empirical results despite their simplicity, but more sophisticated alternatives are certainly possible and may bring improvements in some applications.

3.2 Training and Inference

Heuristically,[8] (7) yields the joint "density"

$$p_{X,U,Z}(x, u, z) := p_Z(z)\, p_{U|Z}(u|z)\, \delta(x - F(z; u)),$$

where p_Z is the density of P_Z and δ is the Dirac delta. If F is sufficiently regular, we can marginalise out the dependence on z by making the change of variable $z = F^{-1}(x'; u)$, which means $\mathrm{d}z = |\det \mathrm{D}F^{-1}(x'; u)|\, \mathrm{d}x'$.[9] This yields a proper density for (X, U) by integrating over x':

$$p_{X,U}(x, u) := p_Z(F^{-1}(x; u)) \times p_{U|Z}(u|F^{-1}(x; u))\, |\det \mathrm{D}F^{-1}(x; u)|. \tag{11}$$

For an L-layered model, an extension of this argument also gives the following joint density for each $(Z_\ell, U_{1:\ell})$:

$$p_{Z_\ell,U_{1:\ell}}(z_\ell, u_{1:\ell}) := p_{Z_{\ell-1},U_{1:\ell-1}}(F_\ell^{-1}(z_\ell; u_\ell), u_{1:\ell-1}) \times p_{U_\ell|Z_{\ell-1}}(u_\ell|F_\ell^{-1}(z_\ell; u_\ell))|\det \mathrm{D}F_\ell^{-1}(z_\ell; u_\ell)|. \tag{12}$$

[7] Note this requires $\mathcal{Z} = \mathcal{X} = \mathbb{R}^d$ and $\mathcal{U} = \mathbb{R}^{d_\mathcal{U}}$, i.e. these domains are not strict subsets. We assume this in all our experiments.

[8] We make this rigorous in Section B.5 of the Supplement.

[9] Here $\mathrm{D}F(z; u)$ denotes the Jacobian with respect to z only.

Taking $X := Z_L$ as before we obtain $p_{X,U_{1:L}}$ and hence a density for P_X via

$$p_X(x) := \int p_{X,U_{1:L}}(x, u_{1:L})\, \mathrm{d}u_{1:L}. \tag{13}$$

Since $\mathcal{U}$ is continuous, this is not analytically tractable. To facilitate likelihood-based training and inference, we make use of a variational scheme that we describe now.

Assuming an L-layered model (8), we introduce an approximate posterior density $q_{U_{1:L}|X} \approx p_{U_{1:L}|X}$ and consider the evidence lower bound (ELBO) of $\log p_X(x)$:

$$\mathcal{L}(x) := \mathbb{E}_{u_{1:L} \sim q_{U_{1:L}|X}(\cdot|x)} \left[\log \frac{p_{X,U_{1:L}}(x, u_{1:L})}{q_{U_{1:L}|X}(u_{1:L}|x)} \right]. \tag{14}$$

It is a standard result that $\mathcal{L}(x) \leq \log p_X(x)$ with equality if and only if $q_{U_{1:L}|X}$ is the exact posterior $p_{U_{1:L}|X}$. This allows learning an approximation to $P_X^\star$ by maximising $\sum_{i=1}^n \mathcal{L}(x_i)$ jointly in $p_{X,U_{1:L}}$ and $q_{U_{1:L}|X}$, where we assume a dataset of n i.i.d. samples $x_i \sim P_X^\star$.

We now consider how to parametrise an effective $q_{U_{1:L}|X}$. Standard approaches to designing inference networks for variational autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014; Rezende & Mohamed, 2015; Kingma et al., 2016), while mathematically valid, would not exploit the conditional independencies induced by the bijective structure of (8). We therefore propose a novel inference network that is specifically targeted towards our model, which we compare with existing VAE approaches in Section 4.3. In particular, our $q_{U_{1:L}|X}$ has the following form:

$$q_{U_{1:L}|X}(u_{1:L}|x) := \prod_{\ell=1}^{L} q_{U_\ell|Z_\ell}(u_\ell|z_\ell), \tag{15}$$

with $z_L := x$ and $z_\ell := F_{\ell+1}^{-1}(z_{\ell+1}; u_{\ell+1})$ for $\ell \in \{1, \ldots, L-1\}$, and $q_{U_\ell|Z_\ell}$ can be any parameterised conditional density. We show in Section B.6 of the Supplement that the posterior $p_{U_{1:L}|X}$ factors in the same way as (15), so that we do not lose any generality. Observe also that this scheme shares parameters between $q_{U_{1:L}|X}$ and $p_{X,U_{1:L}}$ in a natural way, since the same F_ℓ are used in both.

We assume each $q_{U_\ell|Z_\ell}$ can be suitably reparametrised (Kingma & Welling, 2014; Rezende et al., 2014) so that, for some function H_ℓ and some density η_ℓ that does not depend on the parameters of $q_{U_{1:L}|Z_\ell}$ and $p_{X,U_{1:L}}$, we have $H_\ell(\epsilon_\ell, z_\ell) \sim q_{U_\ell|Z_\ell}(\cdot|z_\ell)$ when $\epsilon_\ell \sim \eta_\ell$. We can then obtain unbiased estimates of $\mathcal{L}(x)$ using Algorithm 1, which corresponds to a single-sample approximation to the expectation in (14). It is straightforward to see that Algorithm 1 has $\Theta(L)$ complexity. Differentiating through this procedure allows maximising $\sum_{i=1}^n \mathcal{L}(x_i)$ via stochastic gradient descent. At test time, we can also estimate $\log p_X(x)$ directly using importance sampling as described by Rezende et al. (2014, (40)). In particular, letting $\hat{\mathcal{L}}^{(1)}, \ldots, \hat{\mathcal{L}}^{(m)}$ denote the result of separate calls to ELBO(x), we have

$$m^{-1} \operatorname{LogSumExp}(\hat{\mathcal{L}}^{(1)}, \ldots, \hat{\mathcal{L}}^{(m)}) \to \log p_X(x) \tag{16}$$

almost surely as $m \to \infty$.

Algorithm 1 Unbiased estimation of $\mathcal{L}(x)$

function ELBO(x)
  $z_L \leftarrow x$
  $\Delta \leftarrow 0$
  for $\ell = L, \ldots, 1$ **do**
    $\epsilon \sim \eta_\ell$
    $u \leftarrow H_\ell(\epsilon, z_\ell)$
    $z_{\ell-1} \leftarrow F_\ell^{-1}(z_\ell; u)$
    $\Delta \leftarrow \Delta + \log p_{U_\ell|Z_{\ell-1}}(u|z_{\ell-1}) - \log q_{U_\ell|Z_\ell}(u|z_\ell)$
      $+ \log |\det \mathrm{D} F_\ell^{-1}(z_\ell; u)|$
  return $\Delta + \log p_{Z_0}(z_0)$

In all our experiments we used

$$q_{U_\ell|Z_\ell}(\cdot|z_\ell) = \mathrm{Normal}(\mu_\ell^q(z_\ell), \Sigma_\ell^q(z_\ell)) \tag{17}$$

for appropriate neural networks μ_ℓ^q and Σ_ℓ^q, which is immediately reparameterisable as described e.g. by Kingma & Welling (2014). We found this gave good enough performance that we did not require alternatives such as IAF (Kingma et al., 2016), but such options may also be useful.

Finally, Algorithm 1 requires an expression for $\log |\det \mathrm{D} F_\ell^{-1}(z_\ell; u_\ell)|$. For (9) this is

$$\log \left| \det \mathrm{D} f_\ell^{-1} \left(e^{s_\ell(u_\ell)} \odot (z_\ell + t_\ell(u_\ell)) \right) \right| + \sum_{i=1}^{d} [s_\ell(u_\ell)]_i,$$

where $[x]_i$ denotes the i^{th} dimension of x.

4 Comparison with Related Models

4.1 Comparison with Normalising Flows

We now compare CIFs with normalising flows, and in particular describe how CIFs relax the constraints of bijectivity identified in Section 2.

4.1.1 Advantages

Observe that (7) generalises normalising flows: if $F(\cdot; u)$ does not depend on u, then we obtain (1). Moreover, training with the ELBO in this case does not reduce performance compared with training a flow directly, as the following result shows. Here the components of our model F_θ, $p_{U|Z}^\theta$, and $q_{U|X}^\theta$ are parameterised by $\theta \in \Theta$, and for a given choice of parameters θ we will denote by P_X^θ and $\mathcal{L}^\theta$ the corresponding distribution and ELBO (14) respectively.

Proposition 4.1. *Suppose there exists $\phi \in \Theta$ such that, for some bijection $f : \mathcal{Z} \to \mathcal{X}$, $F_\phi(\cdot; u) = f(\cdot)$ for all $u \in \mathcal{U}$. Likewise, suppose $p^\phi_{U|Z}$ and $q^\phi_{U|X}$ are such that, for some density r on $\mathcal{U}$, $p^\phi_{U|Z}(\cdot|z) = q^\phi_{U|X}(\cdot|x) = r(\cdot)$ for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. If $\mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\theta(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)]$, then*

$$D_{\mathrm{KL}}\big(P^\star_X \,\big\|\, P^\theta_X\big) \leq D_{\mathrm{KL}}(P^\star_X \,\|\, f\#P_Z).$$

Simply stated, in the limit of infinite data, optimising the ELBO will yield at least as performant a model (as measured by the KL) as any normalising flow our model family can express. The proof is in Section B.7 of the Supplement. In practice, our choices (9), (10), and (17) can easily realise the conditions of Proposition 4.1 by zeroing out the output weights of the neural networks (other than f) involved. Thus, for a given f, we have reason to expect a comparative or better performing model (as measured by average log-likelihood) when trained as a CIF rather than as a normalising flow.

We expect this will in fact lead to *improved* performance because, intuitively, $P_{U|Z}$ can reroute z that would otherwise map outside of $\operatorname{supp} P^\star_X$. To illustrate, fix f in (9) and choose some $z \in \mathcal{Z}$. If $f(z) \in \operatorname{supp} P^\star_X$, then setting $F(z; u) = f(z)$ for all $u \in \mathcal{U}$ as described above ensures $F(z; U) \in \operatorname{supp} P^\star_X$ when $U \sim P_{U|Z}(\cdot|z)$. If conversely $f(z) \notin \operatorname{supp} P^\star_X$, then we *still* have $F(z; U) \in \operatorname{supp} P^\star_X$ almost surely if $P_{U|Z}(\cdot|z)$ is supported on $\{u \in \mathcal{U} : F(z; u) \in \operatorname{supp} P^\star_X\}$. Of course, if f is too simple, then $P_{U|Z}$ must heuristically become very complex in order to obtain this behaviour. This would seem to make inference harder, leading to a looser ELBO (14) and thus overall worse performance after training. We therefore expect CIFs to work well for f that, like the 10-layer ResFlow in Figure 1, can learn a close approximation to the support of the target but "leak" some mass outside of it due to (4) or Theorem 2.1. A CIF can then use $P_{U|Z}$ to "clean up" these small extraneous regions of mass.

We provide empirical support for this argument in Section 5. We also summarise our discussion above with the following precise result. Here ∂A denotes the boundary of a set A.

Proposition 4.2. *If $P^\star_X(\partial \operatorname{supp} P^\star_X) = 0$ and $(z, u) \mapsto F(z; u)$ is jointly continuous with*

$$\overline{F(\operatorname{supp} P_Z \times \mathcal{U})} \supseteq \operatorname{supp} P^\star_X, \tag{18}$$

then there exists $P_{U|Z}$ such that $\operatorname{supp} P_X = \operatorname{supp} P^\star_X$ if and only if, for all $z \in \operatorname{supp} P_Z$, there exists $u \in \mathcal{U}$ with

$$F(z; u) \in \operatorname{supp} P^\star_X. \tag{19}$$

The assumptions here are fairly minimal: the boundary condition ensures $P^\star_X$ is not pathological, and if (18) does not hold, then $D_{\mathrm{KL}}(P^\star_X \,\|\, P_X) = \infty$ for *every* $P_{U|Z}$.[10]

[10] See Proposition B.1 and Proposition B.3 in the Supplement.

Additionally, the following result gives a sufficient condition under which it is possible to learn the target exactly.

Proposition 4.3. *If $F(z; \cdot) : \mathcal{U} \to \mathcal{X}$ is surjective for each $z \in \mathcal{Z}$, then there exists $P_{U|Z}$ such that $P_X = P^\star_X$.*

See Section B.8 of the Supplement for proofs. These results do not require $\operatorname{supp} P_Z \cong \operatorname{supp} P^\star_X$, thereby showing CIFs relax the constraint (4) for standard normalising flows.

Of course, in practice, our parameterisation (9) does not necessarily ensure that F will satisfy these conditions, and our parameterisation (10) may not be expressive enough to instantiate the $P_{U|Z}$ that is required. However, these results show that CIFs provide at least a *mechanism* for correcting a topologically misspecified prior. When F and $P_{U|Z}$ are sufficiently expressive, we can expect that they will learn to approximate these conditions over the course of training if doing so produces a better density estimate. We therefore anticipate CIFs will improve performance for ResFlows, where Theorem 2.1 applies, and may have benefits more generally, since all flows are ultimately constrained by (4).

4.1.2 Disadvantages

On the other hand, CIFs introduce additional overhead compared with regular normalising flows. It therefore remains to show we obtain better performance on a fixed computational budget, which requires using a smaller model. Empirically this holds for the models and datasets we consider in Section 5, but there are likely cases where it does not, particularly if the topologies of the target and prior are similar.

Likewise, CIFs sacrifice the exactness of normalising flows. We do not see this as a significant problem for the task of density estimation, since the importance sampling estimator (16) means that at test time we can obtain arbitrary accuracy by taking m to be large. However, the lack of a closed-form density does limit the use of CIFs in some downstream tasks. In particular, CIFs cannot immediately be plugged in to a variational approximation in the manner of Rezende & Mohamed (2015), since this requires exact likelihoods. However, it may be possible to use CIFs in the context of an extended-space variational framework along the lines of Agakov & Barber (2004), and we leave this for future work.

4.2 Comparison with Discretely Indexed Models

Similar models to CIFs have been proposed that use a discrete index space. In the context of Bayesian inference, Duan (2019) proposes a single-layer ($L = 1$) model consisting of (7) with $\mathcal{U} = \{1, \dots, I\}$ and $F(\cdot; i) = f_i$ for separate normalising flows $f_1, \dots, f_I$. A special case of this framework is given by *deep Gaussian mixture models* (Van den Oord & Schrauwen, 2014; van den Oord & Dambre, 2015), which corresponds to using invertible linear transformations for each f_i. In this case, (13) becomes a summation that can

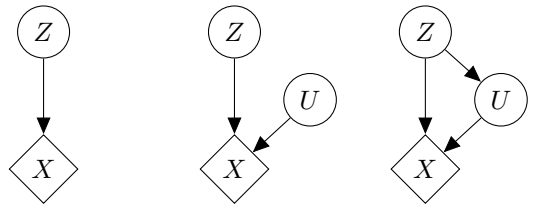


Figure 2: Comparison of related generative models. Circular nodes are random and diamond nodes are deterministic. CIFs generalise both normalising flows and VAEs as shown.

be computed analytically. However, this quickly becomes intractable as L grows larger, since the cost to compute this is seen to be $\Theta(I^L)$. Unlike for a continuous u, this cannot easily be reduced to $\Theta(L)$ using a variational approximation as in Section 3.2, since a discrete $q_{U|X}$ is not amenable to the reparameterisation trick. In addition, the use of separate bijections also means that the number of parameters of the model grows as I increases. In contrast, a continuous index allows a natural mechanism for sharing parameters across different $F(\cdot; u)$ as in (9).

Prior to Duan (2019), Dinh et al. (2019) proposed RAD as a means to mitigate the $\Theta(I^L)$ cost of naïvely stacking discrete layers. RAD partitions $\mathcal{X}$ into I disjoint subsets $B_1, \ldots, B_I$ and defines bijections $f_i : \mathcal{Z} \to B_i$ for each i. The model is then taken to be the marginal of X in

$$Z \sim P_Z, \quad U \sim P_{U|Z}(\cdot|Z), \quad X := f_U(Z),$$

where each $P_{U|Z}(\cdot|z)$ is a discrete distribution on $\{1, \ldots, I\}$. Note that this is not an instance of our model (7), since we require each $F(\cdot; u)$ to be surjective onto $\mathcal{X}$. The use of partitioning means that (13) is a summation with only a single term, which reduces the cost for L layers to $\Theta(L)$. However, partitioning also makes p_X discontinuous. This leads to a very difficult optimisation problem and Dinh et al. (2019) only report results for simple 2-D densities. Additionally, partitioning requires ad-hoc architectural changes to existing normalising flows, and does not directly address the increasing parameter cost as I grows large.

4.3 Comparison with Variational Autoencoders

CIFs also generalise a broad family of variational autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014). Recall that VAEs take

$$p_X(x) := \int p_U(u) p_{X|U}(x|u) \, \mathrm{d}u \tag{20}$$

for some choices of densities p_U and $p_{X|U}$.[11] For instance, a mean-field Gaussian observation density has

$$p_{X|U}(\cdot|u) := \mathrm{Normal}\left(t(u), \mathrm{diag}\left(e^{s(u)}\right)\right),$$

where $t, s : \mathcal{U} \to \mathcal{X}$, and $\mathrm{diag}(v)$ denotes the matrix with diagonal $v \in \mathbb{R}^d$ and zeros elsewhere. If P_Z is a standard Gaussian, if each $P_{U|Z}(\cdot|z)$ has independent density p_U, and if F is (9) with f the identity, then it follows that (7) has marginal density (20) (modulo the signs of s and t).[12]

More generally, every VAE model (20) with each $p_{X|U}(\cdot|u)$ strictly positive corresponds to an instance of (7) where U is sampled independently of Z. To see this, let p_Z be any strictly positive density on $\mathcal{Z}$, and let each $F(\cdot; u)$ be the Knothe-Rosenblatt coupling (Villani, 2008) of p_Z and $p_{X|U}(\cdot|u)$. By construction each $F(\cdot; u)$ is invertible and gives $F(Z; u) \sim p_{X|U}(\cdot|u)$ when $Z \sim p_Z$. As a result, (7) again yields X with a marginal density defined by (20). Consequently, CIFs generalise the VAE framework by adding an additional edge in the graphical model as shown in Figure 2.

On the other hand, CIFs differ from VAEs in the way they are composed. Whereas CIFs stack by taking p_Z to be a CIF, VAEs are typically stacked by taking p_U to be a VAE (Rezende et al., 2014; Kingma et al., 2014; Burda et al., 2016; Sønderby et al., 2016). This has implications for the design of the inference network $q_{U_{1:L}|X}$. In particular, a hierarchical VAE obtained in this way is *Markovian*, so that

$$p_{U_{1:L}|X}(x, u_{1:L}) = p_{U_L|X}(u_L|x) \prod_{\ell=1}^{L} p_{U_\ell|U_{\ell-1}}(u_\ell|u_{\ell-1})$$

where L is the number of layers. This directly allows specifying $q_{U_{1:L}|X}$ to be of the same form without any loss of generality (Kingma et al., 2014; Burda et al., 2016; Sønderby et al., 2016). Conversely, CIFs do not factor in this way, which motivates our alternative approach in Section 3.2.

Note finally that CIFs should not be conflated with the large class of methods that use normalising flows to improve the *inference* procedure in VAEs (Rezende & Mohamed, 2015; Kingma et al., 2016; van den Berg et al., 2018). These approaches are orthogonal to ours and indeed may be useful for improving our own inference procedure by replacing (17) with a more expressive model.

4.4 Other Related Work

Additional related methods have been proposed. Within a classification context, Dupont et al. (2019) identify topological problems related to ODE-based mappings (Chen et al.,

[11] Note that this notation is nonstandard for VAEs in order to align with the rest of the paper. Here our U corresponds to z as used by Kingma & Welling (2014).

[12] Here Z corresponds to ϵ as used by Kingma & Welling (2014).

2018), which like normalising flows are homeomorphisms and hence preserve the topology of their input. To avoid this, Dupont et al. (2019) propose augmenting the data by appending auxiliary dimensions and learning a new mapping on this space. In contrast, CIFs may be understood as augmenting not the data but instead the *model* by considering a family of individual bijections on the *original* space.

In addition, Ho et al. (2019) use a variational scheme to improve on the standard dequantisation method proposed by Theis et al. (2016) for modelling image datasets with normalising flows. This approach is potentially complementary to CIFs, but we do not make use of it in our experiments.

5 Experiments

We evaluated the performance of CIFs on several problems of varying difficulty, including synthetic 2-D data, several tabular datasets, and three image datasets. In all cases we took $\mathcal{Z} = \mathcal{X} = \mathbb{R}^d$ with d the dimension of the dataset. We used the stacked architecture (8) with the prior P_{Z_0} a Gaussian. At each layer, F had form (9) with f a primitive flow step from a baseline architecture (e.g. a single residual block for ResFlow). Each $p_{U|Z}$ and $q_{U|X}$ had form (10) and (17) respectively. We provide an overview of our results for the tabular and image datasets here. Full experimental details, including additional 2-D figures along the lines of Figure 1, are in Section C of the Supplement. See `github.com/jrmcornish/cif` for our code.

5.1 Tabular Datasets

We tested the performance of CIFs on the tabular datasets used by Papamakarios et al. (2017). For each dataset, we trained 10 and 100-layer baseline fully connected ResFlows, and corresponding 10-layer CIF-ResFlows. The CIF-ResFlows had roughly 1.5-4.5% more parameters (depending on the dimension of the dataset) than the otherwise identical 10-layer ResFlows, and roughly 10% of the parameters of the 100-layer ResFlows. Table 1 reports the average log-probability of the test set that we obtained for each model. Observe that in all cases CIF-ResFlows significantly outperform both baseline models. Moreover, for all but GAS, the CIF-ResFlows achieve state-of-the-art performance based on the results reported by Durkan et al. (2019, Table 1). This is particularly noticeable for POWER and BSDS300, where CIF-ResFlow improves on the best results of Durkan et al. (2019) by 0.94 and 2.77 nats respectively.

We additionally tried using *masked autoregressive flows* (MAFs) (Papamakarios et al., 2017) and *neural spline flows* (NSFs) (Durkan et al., 2019) for f. In each case, we closely match the experimental settings of the baselines and augment using CIFs, controlling for the number of parameters used by the CIF extensions. Table 1 reports the average log-probability across the test set for each experiment. Here, CIF-NSF-1 is a CIF with the same number of parameters as the baseline, and CIF-NSF-2 is a model using a baseline configuration for f (but having more parameters overall). We see that CIF-MAFs consistently outperform MAFs across datasets; CIF-NSFs do not improve upon NSFs as dramatically, although we still notice improvements and would expect to improve further with more hyperparameter tuning. Lastly it is important to notice that MAFs and NSFs do not restrict the Lipschitz constant of f. These results show that CIFs can yield benefits for normalising flows even if Theorem 2.1 is not directly a limitation.

Finally, for ablation purposes we tried taking f to be the identity. We obtained consistently worse performance than for CIF-ResFlows and CIF-MAF in this case, which aligns with our conjecture in Section 4.1.1 that a performant CIF requires an expressive base flow f. Details and results are given in Section C.1.4 of the Supplement.

5.2 Image Datasets

We also considered CIFs applied to the MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky & Hinton, 2009) datasets. Following our tabular experiments, we trained a multi-scale convolutional ResFlow and a corresponding CIF-ResFlow, as well as a larger baseline ResFlow to account for the additional parameters and depth introduced by our method. Note that these models were significantly smaller than those used by Chen et al. (2019): e.g. for CIFAR10, the ResFlow used by Chen et al. (2019) had 25M parameters, while our two baseline ResFlows and our CIF-ResFlow had 2.4M, 6.2M, and 5.6M parameters respectively. We likewise considered RealNVPs with the same multi-scale convolutional architecture used by Dinh et al. (2017) for their CIFAR-10 experiments. For these runs we trained baseline RealNVPs, corresponding CIF-RealNVPs, and larger baseline RealNVPs with more depth and parameters.

The results are given in Table 2 and Table 3. Observe CIFs outperformed the baseline models for all datasets, which shows that our approach can scale to high dimensions. For the CIF-ResFlows, we also obtained better performance than Chen et al. (2019) on MNIST and better performance than Glow (Kingma & Dhariwal, 2018) on CIFAR10, despite using a much smaller model. Samples from all models are shown in Section C.2 of the Supplement.

6 Conclusion and Future Work

The constraint (4) shows that normalising flows are unable to exactly model targets whose topology differs from that

[13] Only one seed was used per run due to computational limitations. However, the results were not cherry-picked.

Table 1: Mean ± standard error (over 3 seeds) of average test set log-likelihood (in nats). Higher is better. Best performing runs for each group are shown in bold. A ⋆ indicates state-of-the-art performance according to Durkan et al. (2019, Table 1).

	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
RESFLOW ($L = 10$)	-2.73 ± 0.03	4.16 ± 0.08	-20.68 ± 0.02	-14.2 ± 0.10	123.51 ± 0.09
RESFLOW ($L = 100$)	0.48 ± 0.00	10.57 ± 0.17	-16.67 ± 0.05	-11.16 ± 0.04	148.05 ± 0.61
CIF-RESFLOW ($L = 10$)	$\mathbf{1.60 \pm 0.21}^\star$	$\mathbf{12.12 \pm 0.10}$	$\mathbf{-13.74 \pm 0.03}^\star$	$\mathbf{-8.10 \pm 0.04}^\star$	$\mathbf{160.50 \pm 0.08}^\star$
MAF	0.19 ± 0.02	9.23 ± 0.07	-18.33 ± 0.10	-10.98 ± 0.03	156.13 ± 0.00
CIF-MAF	$\mathbf{0.48 \pm 0.01}$	$\mathbf{12.02 \pm 0.10}$	$\mathbf{-16.63 \pm 0.09}$	$\mathbf{-9.93 \pm 0.04}$	$\mathbf{156.67 \pm 0.02}$
NSF	$\mathbf{0.69 \pm 0.00}$	13.01 ± 0.02	-14.30 ± 0.05	-10.68 ± 0.06	$\mathbf{157.59 \pm 0.02}$
CIF-NSF-1	$\mathbf{0.68 \pm 0.01}$	12.94 ± 0.01	$\mathbf{-13.83 \pm 0.10}$	$\mathbf{-9.93 \pm 0.06}$	$\mathbf{157.60 \pm 0.02}$
CIF-NSF-2	$\mathbf{0.69 \pm 0.00}$	$\mathbf{13.08 \pm 0.00}$	-14.18 ± 0.09	-10.80 ± 0.01	157.56 ± 0.02

Table 2: Average test bits per dimension.[13] Lower is better.

	MNIST	CIFAR-10
RESFLOW (SMALL)	1.074	3.474
RESFLOW (BIG)	1.018	3.422
CIF-RESFLOW	**0.922**	**3.334**

Table 3: Mean ± standard error of average test set bits per dimension over 3 random seeds. Lower is better.

	FASHION-MNIST	CIFAR-10
REALNVP (SMALL)	2.944 ± 0.003	3.565 ± 0.001
REALNVP (BIG)	2.946 ± 0.002	3.554 ± 0.001
CIF-REALNVP	$\mathbf{2.823 \pm 0.003}$	$\mathbf{3.477 \pm 0.019}$

of the prior. Moreover, in order to approximate such targets closely, Theorem 2.1 shows that the bi-Lipschitz constant of a flow must become arbitrarily large. To address these problems, we have proposed CIFs, which can "clean up" regions of mass that are placed outside the support of the target by a standard flow. CIFs perform well in practice and outperform baseline flows on several benchmark datasets.

While we have focussed on the use of CIFs for density estimation in this paper, it would also be interesting to apply CIFs in other contexts where normalising flows have been used successfully. As CIFs do not have an analytically available density, this would likely require the modification of existing numerical frameworks, but the expressiveness benefits provided by CIFs might make this additional effort worthwhile. We leave this direction for future work.

Acknowledgements

Rob Cornish is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems (EP/L015897/1) and NVIDIA. Anthony Caterini is a Commonwealth Scholar supported by the U.K. Government. Arnaud Doucet is partially supported by the U.S. Army Research Laboratory, the U.S. Army Research Office, and by the U.K. Ministry of Defence (MoD) grant EP/R013616/1 and the U.K. EPSRC under grant numbers EP/R034710/1 and EP/R018561/1.

References

Agakov, F. V. and Barber, D. An auxiliary variational method. In *International Conference on Neural Information Processing*, pp. 561–566. Springer, 2004.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Beatson, A. and Adams, R. P. Efficient optimization of loops and limits with randomized telescoping sums. In *International Conference on Machine Learning*, pp. 534–543, 2019.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. In *International Conference on Machine Learning*, pp. 573–582, 2019.

Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R. B., and Jacobsen, J.-H. On the invertibility of invertible neural networks, 2020. URL https://openreview.net/forum?id=BJlVeyHFwH.

Billingsley, P. *Probability and Measure*. John Wiley & Sons, 2008.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *ICLR*, 2016.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6571–6583, 2018.

Chen, T. Q., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pp. 9913–9923, 2019.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. In *ICLR Workshop*, 2015.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *ICLR*, 2017.

Dinh, L., Sohl-Dickstein, J., Pascanu, R., and Larochelle, H. A RAD approach to deep mixture models. In *ICLR Workshop*, 2019.

Duan, L. L. Transport Monte Carlo. *arXiv preprint arXiv:1907.10448*, 2019.

Dudley, R. M. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

Dupont, E., Doucet, A., and Teh, Y. W. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, pp. 3134–3144, 2019.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In *Advances in Neural Information Processing Systems*, pp. 7509–7520, 2019.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pp. 881–889, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

Grathwohl, W., Chen, R. T., Betterncourt, J., Sutskever, I., and Duvenaud, D. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *ICLR*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016b.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pp. 2722–2730, 2019.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. In *International Conference on Machine Learning*, pp. 2083–2092, 2018.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.

Jaini, P., Selby, K. A., and Yu, Y. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*, pp. 3009–3018, 2019.

Kahn, H. Use of different Monte Carlo sampling techniques. Technical report, Rand Corporation, 1955.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *ICLR*, 2014.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kumar, A., Poole, B., and Murphy, K. Learning generative samplers using relaxed injective flow. In *ICML Workshop on Invertible Neural Nets and Normalizing Flows*, 2019.

LeCun, Y. The MNIST database of handwritten digits, 1998. URL `http://yann.lecun.com/exdb/mnist/`.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., Simpson, D., et al. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423. IEEE, 2001.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pp. 2338–2347, 2017.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286, 2014.

Rhee, C.-h. and Glynn, P. W. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

Rudin, W. *Principles of Mathematical Analysis*, volume 3. McGraw-hill New York, 1964.

Rudin, W. *Real and Complex Analysis*. Tata McGraw-hill education, 2006.

Runde, V. *A Taste of Topology*. Springer, 2007.

Skilling, J. The eigenvalues of mega-dimensional matrices. In *Maximum Entropy and Bayesian Methods*, pp. 455–466. Springer, 1989.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 3738–3746, 2016.

Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. In *ICLR*, 2016.

van den Berg, R., Hasenclever, L., Tomczak, J. M., and Welling, M. Sylvester normalizing flows for variational inference. In *UAI 2018: The Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 393–402, 2018.

van den Oord, A. and Dambre, J. Locally-connected transformations for deep GMMs. In *International Conference on Machine Learning (ICML): Deep learning Workshop*, pp. 1–8, 2015.

Van den Oord, A. and Schrauwen, B. Factoring variations in natural images with deep Gaussian mixture models. In *Advances in Neural Information Processing Systems*, pp. 3518–3526, 2014.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows: Supplementary Material

A Guide to Notation

(a_n)	A sequence of elements $a_1, a_2, \ldots$
$a(n) = \Theta(b(n))$	$a(n)$ differs from $b(n)$ by at most a constant factor as $n \to \infty$
$u \odot v$	The elementwise product of tensors u and v
$\mathrm{LogSumExp}(a_1, \ldots, a_m)$	$\log\left(\sum_{i=1}^m \exp(a_i)\right)$
e^v, where $v \in \mathbb{R}^d$	$(e^{v_1}, \ldots, e^{v_d})$
$\|v\|$	The norm of a vector $v \in \mathbb{R}^d$ (our results are agnostic to the specific choice of $\|\cdot\|$)
$\|A\|_{\mathrm{op}}$	The operator norm of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ induced by $\|\cdot\|$
I_d	The $d \times d$ identity matrix
$\det A$	The determinant of a square matrix A
$\mathrm{D}f(z)$	The Jacobian matrix of a function f evaluated at z
$\mathrm{D}F(z; u)$	The Jacobian matrix of a function $\mathrm{D}F(\cdot; u)$ (i.e. with u fixed) evaluated at z
$\mathrm{Lip}\, f$	The Lipschitz constant of a function f
$\mathrm{BiLip}\, f$	The bi-Lipschitz constant of a function f
$\mathcal{A} \cong \mathcal{B}$	The topological spaces $\mathcal{A}$ and $\mathcal{B}$ are homeomorphic
$\overline{B}$	The topological closure of a set B
$\mathrm{int}(B)$	The interior of a set B
∂B	The boundary of a set B
$\mathrm{supp}\, \mu$	The support of a measure μ
$f\#\mu$	The pushforward of a measure μ by a function f
$\mu_n \xrightarrow{\mathcal{D}} \mu$	Weak convergence of the measures μ_n to μ

B Proofs

B.1 Preliminaries

We require some basic results that we include here for completeness. We will make use of standard definitions and results from topology and real analysis. A complete background to these topics can be found in Dudley (2002).

B.1.1 Supports of Measures

Recall that for a Borel measure μ on a topological space $\mathcal{Z}$, the *support* of μ, denoted $\mathrm{supp}\, \mu$, is the set of all $z \in \mathcal{Z}$ such that $\mu(N_z) > 0$ for every open set N_z containing z.

The following is an immediate consequence:

Proposition B.1. *Suppose μ and ν are Borel measures with μ absolutely continuous with respect to ν. Then*

$$\mathrm{supp}\, \mu \subseteq \mathrm{supp}\, \nu.$$

Proof. Suppose $z \notin \mathrm{supp}\, \nu$. Then there exists an open set N_z containing z such that $\nu(N_z) = 0$. By absolute continuity, we have also that $\mu(N_z) = 0$ and hence $z \notin \mathrm{supp}\, \mu$. □

In general the converse need not hold. For example, the Dirac measure on 0 has support contained within the Lebesgue measure on $\mathbb{R}$ (which has full support), but is not absolutely continuous with respect to it.

The following characterisation is useful:

Proposition B.2. *For any Borel measure μ,*

$$(\operatorname{supp}\mu)^c = \bigcup_{\substack{A \text{ open:} \\ \mu(A)=0}} A, \tag{B.1}$$

and hence $\operatorname{supp}\mu$ *is closed.*

Proof. This follows directly from the definitions, since $z \notin \operatorname{supp}\mu$ if and only if there exists open N_z with $z \in N_z$ and $\mu(N_z) = 0$, which is just another way of saying that z is contained in the right-hand side of (B.1). It follows that $(\operatorname{supp}\mu)^c$ is open, and hence $\operatorname{supp}\mu$ is closed. □

We mainly care about how the support of a measure is transformed by a pushforward function. The following proposition characterises what occurs in this case.

Proposition B.3. *Suppose $\mathcal{Z}$ and $\mathcal{X}$ are topological spaces. If μ is a Borel measure on $\mathcal{Z}$ such that $\mu((\operatorname{supp}\mu)^c) = 0$, and if $f : \mathcal{Z} \to \mathcal{X}$ is continuous, then*

$$\operatorname{supp} f\#\mu = \overline{f(\operatorname{supp}\mu)}.$$

Proof. Suppose $x \notin \overline{f(\operatorname{supp}\mu)}$. Then x must have an open neighbourhood N_x such that

$$N_x \cap f(\operatorname{supp}\mu) = \emptyset.$$

This implies

$$\begin{aligned} f^{-1}(N_x) \cap \operatorname{supp}\mu &\subseteq f^{-1}(N_x) \cap f^{-1}(f(\operatorname{supp}\mu)) \\ &= f^{-1}(N_x \cap f(\operatorname{supp}\mu)) \\ &= f^{-1}(\emptyset) \\ &= \emptyset. \end{aligned}$$

We then have

$$f\#\mu(N_x) = \mu(f^{-1}(N_x)) = \mu(f^{-1}(N_x) \cap \operatorname{supp}\mu) = 0,$$

where the second equality follows since we assumed $\mu((\operatorname{supp}\mu)^c) = 0$, and hence $x \notin \operatorname{supp} f\#\mu$. Consequently

$$\operatorname{supp} f\#\mu \subseteq \overline{f(\operatorname{supp}\mu)}.$$

In the other direction, suppose $x \in f(\operatorname{supp}\mu)$, so that $x = f(z)$ for some $z \in \operatorname{supp}\mu$. Given an open neighbourhood N_x it then follows from continuity that $f^{-1}(N_x)$ is an open neighbourhood of z, and so

$$f\#\mu(N_x) = \mu(f^{-1}(N_x)) > 0$$

since $z \in \operatorname{supp}\mu$. This entails $\operatorname{supp} f\#\mu \supseteq f(\operatorname{supp}\mu)$, which means

$$\operatorname{supp} f\#\mu = \overline{\operatorname{supp} f\#\mu} \supseteq \overline{f(\operatorname{supp}\mu)}$$

by Proposition B.2. □

Note that in general we need not have $\operatorname{supp} f\#\mu = f(\operatorname{supp}\mu)$. For example, if μ is Gaussian and $f = \arctan$, then

$$f(\operatorname{supp}\mu) = (-1, 1) \neq [-1, 1] = \operatorname{supp} f\#\mu.$$

Likewise, in general we do require the assumption $\mu((\operatorname{supp}\mu)^c) = 0$. This is because there exist examples of nontrivial Borel measures μ such that $\operatorname{supp}\mu = \emptyset$. Taking $f \equiv x_0$ to be any constant $x_0 \in \mathcal{X}$ (in which case f is certainly continuous) then gives

$$\overline{f(\operatorname{supp}\mu)} = \emptyset \neq \{x_0\} = \operatorname{supp} f\#\mu.$$

However, for our purposes, the following proposition shows that this is not a restriction.

Proposition B.4. *Suppose μ is a Borel measure on a separable metric space $\mathcal{Z}$. Then*

$$\mu((\operatorname{supp}\mu)^c) = 0.$$

Proof. Throughout the proof, for each z and $r > 0$, we will denote by $B(z, r)$ an open ball of radius r centered at z. Likewise, for each $z \notin \operatorname{supp}\mu$, let

$$r^\star(z) := \sup\{r > 0 \mid \mu(B(z, r)) = 0\}.$$

Observe that $r^\star$ is well-defined (but possibly infinite) since $z \notin \operatorname{supp}\mu$ means there must exist some $r > 0$ such that $\mu(B(z, r)) = 0$.

We first show that $\mu(B(z, r^\star(z))) = 0$ for all $z \notin \operatorname{supp}\mu$. To this end, fix z and choose a sequence $r_m \uparrow r^\star(z)$ with $r_m < r^\star(z)$. We then have

$$B(z, r^\star(z)) = \bigcup_{m=1}^{\infty} B(z, r_m),$$

and so

$$\mu(B(z, r^\star(z))) = \lim_{m\to\infty} \mu(B(z, r_m)) = 0$$

by continuity of measure.

Now, by separability, we can choose a countable sequence $(z_k) \subseteq (\operatorname{supp}\mu)^c$ such that $\overline{\{z_k\}} = \overline{(\operatorname{supp}\mu)^c}$. We show that

$$(\operatorname{supp}\mu)^c = \bigcup_{k=1}^{\infty} B(z_k, r^\star(z_k)),$$

from which the result follows by countable subadditivity. It is clear from (B.1) that the left-hand side is a superset of the right. In the other direction, let $z \in (\operatorname{supp}\mu)^c$. By construction of (z_k), there exists a subsequence $(z_{k'})$ such that $z_{k'} \to z$. For all k' large enough we then have $z_{k'} \in B(z, r^\star(z)/2)$ and hence

$$B(z_{k'}, r^\star(z)/2) \subseteq B(z, r^\star(z))$$

by triangle inequality. It follows that for such k' we have

$$\mu(B(z_{k'}, r^\star(z)/2)) \leq \mu(B(z, r^\star(z))) = 0,$$

and so $r^\star(z_{k'}) \geq r^\star(z)/2$ since $r^\star(z_{k'})$ is the supremum. But then we have

$$z \in B(z_{k'}, r^\star(z)/2) \subseteq B(z_{k'}, r^\star(z_{k'})),$$

so that

$$z \in \bigcup_{k=1}^{\infty} B(z_k, r^\star(z_k))$$

and we are done. □

B.2 Lipschitz and Bi-Lipschitz Functions

We assume that $\mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}}$, $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$, and $f : \mathcal{Z} \to \mathcal{X}$. Recall that the *Lipschitz* constant of f, denoted $\operatorname{Lip} f$, is defined as the infimum over $M \in [0, \infty]$ such that

$$\|f(z) - f(z')\| \leq M\|z - z'\|$$

for all $z, z' \in \mathcal{Z}$. Likewise the *bi-Lipschitz* constant $\operatorname{BiLip} f$ is defined as the infimum over $M \in [1, \infty]$ such that

$$M^{-1}\|z - z'\| \leq \|f(z) - f(z')\| \leq M\|z - z'\|$$

for all $z, z' \in \mathcal{Z}$. We prove some basic properties that follow from this definition.

Proposition B.5. $\operatorname{BiLip} f < \infty$ *if and only if f is injective and* $\max(\operatorname{Lip} f, \operatorname{Lip} f^{-1}) < \infty$*, where $f^{-1} : f(\mathcal{Z}) \to \mathcal{Z}$. For all injective f, we then have* $\operatorname{BiLip} f = \max(\operatorname{Lip} f, \operatorname{Lip} f^{-1})$.

Proof. For the first statement, suppose $\operatorname{BiLip} f < \infty$. It is immediate that $\operatorname{BiLip} f \geq \operatorname{Lip} f$. To see that f is injective, note that for $z \neq z'$ we have

$$\|f(z) - f(z')\| \geq (\operatorname{BiLip} f)^{-1}\|z - z'\| > 0$$

and so $f(z) \neq f(z')$. On the other hand, for $x, x' \in f(\mathcal{Z})$, we have

$$(\operatorname{BiLip} f)^{-1}\|f^{-1}(x) - f^{-1}(x')\| \leq \|f(f^{-1}(x)) - f(f^{-1}(x'))\| = \|x - x'\|,$$

which gives that $\operatorname{BiLip} f \geq \operatorname{Lip} f^{-1}$. Altogether we have

$$\max(\operatorname{Lip} f, \operatorname{Lip} f^{-1}) \leq \operatorname{BiLip} f < \infty, \tag{B.2}$$

which gives the forward direction.

Next suppose f is injective and that

$$M := \max(\operatorname{Lip} f, \operatorname{Lip} f^{-1}) < \infty.$$

For $z, z' \in \mathcal{Z}$, we certainly have

$$\|f(z) - f(z')\| \leq M\|z - z'\|.$$

Likewise, since $f(z), f(z') \in f(\mathcal{Z})$,

$$\|z - z'\| = \|f^{-1}(f(z)) - f^{-1}(f(z'))\| \leq M\|f(z) - f(z')\|,$$

so that

$$M^{-1}\|z - z'\| \leq \|f(z) - f(z')\|$$

because injectivity of f means that $M > 0$. From this it follows that

$$\operatorname{BiLip} f \leq M < \infty, \tag{B.3}$$

which gives the reverse direction, proving the first statement.

For the second statement, suppose f is injective. Then if $\operatorname{BiLip} f < \infty$, (B.2) and (B.3) together give

$$\operatorname{BiLip} f = \max(\operatorname{Lip} f, \operatorname{Lip} f^{-1}).$$

On the other hand, if $\operatorname{BiLip} f = \infty$ then $\max(\operatorname{Lip} f, \operatorname{Lip} f^{-1}) = \infty$ since we would otherwise obtain a contradiction by the first statement of the proposition. This completes the proof. □

It follows directly that if $\operatorname{BiLip} f < \infty$, then f is a homeomorphism from $\mathcal{Z}$ to $f(\mathcal{Z})$.[14] Moreover, in this case f maps closed sets to closed sets, as the following result shows:

Proposition B.6. *If* $\operatorname{BiLip} f < \infty$ *and* $\mathcal{Z}$ *is closed in* $\mathbb{R}^{d_\mathcal{Z}}$*, then* $f(\mathcal{Z})$ *is closed in* $\mathbb{R}^{d_\mathcal{X}}$*.*

Proof. It is a straightforward consequence of Proposition B.5 that if $(x_n) \subseteq f(\mathcal{Z})$ is Cauchy, then $(f^{-1}(x_n))$ is Cauchy. Consequently $(f^{-1}(x_n))$ converges to some $z_\infty \in \mathcal{Z}$, since $\mathcal{Z}$ is a closed subset of a complete space and therefore complete. But then

$$\begin{aligned}\|x_n - f(z_\infty)\| &= \|f(f^{-1}(x_n)) - f(z_\infty)\| \\ &\leq M\|f^{-1}(x_n) - z_\infty\| \\ &\to 0\end{aligned}$$

as $n \to \infty$. Consequently $f(\mathcal{Z})$ is complete, and so $f(\mathcal{Z})$ is closed as desired since the ambient space $\mathbb{R}^{d_\mathcal{X}}$ is complete. □

The Lipschitz constant can be computed from the *operator norm* $\|\cdot\|_{\text{op}}$ of the Jacobian of f. Recall that $\|\cdot\|_{\text{op}}$ is defined as for a matrix $A \in \mathbb{R}^{d_\mathcal{X} \times d_\mathcal{Z}}$ as

$$\|A\|_{\text{op}} := \sup_{\substack{v \in \mathbb{R}^{d_\mathcal{Z}}: \\ \|v\|=1}} \|Av\|$$

where we think of elements of $\mathbb{R}^{d_\mathcal{Z}}$ as column vectors.

[14] Note however that the converse is not true in general: for example, exp is a homeomorphism from $\mathbb{R}$ to $(0, \infty)$, but $\operatorname{BiLip} \exp = \infty$.

Proposition B.7. *If $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$, and f is everywhere differentiable, then*

$$\operatorname{Lip} f = \sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}}.$$

Proof. If $v \in \mathcal{Z}$ with $\|v\| = 1$, then

$$\begin{aligned}
\|[\mathrm{D}f(z)]v\| &= \lim_{t \to 0} \frac{\|f(z + tv) - f(z)\|}{|t|} \\
&\leq \lim_{t \to 0} \frac{(\operatorname{Lip} f)\|(z + tv) - z\|}{|t|} \\
&= \operatorname{Lip} f.
\end{aligned}$$

It follows directly that

$$\|\mathrm{D}f(z)\|_{\mathrm{op}} \leq \operatorname{Lip} f.$$

On the other hand, suppose $\operatorname{Lip} f > M$. Then there exists $z, z' \in \mathcal{Z}$ such that

$$\|f(z) - f(z')\| > M\|z - z'\|.$$

Since f is differentiable, so too is the map $\varphi : [0, 1] \to \mathcal{X}$ defined by

$$\varphi(t) := f(tz + (1 - t)z').$$

By Theorem 5.19 of Rudin (1964), there exists $t_0 \in (0, 1)$ such that the derivative φ' satisfies

$$\|\varphi'(t_0)\| \geq \|f(z') - f(z)\| > M\|z - z'\|.$$

But, letting $z_0 := t_0 z + (1 - t_0)z'$, observe that

$$\begin{aligned}
\varphi'(t_0) &= \lim_{t \to 0} \frac{f(z_0 + t(z - z')) - f(z_0)}{t} \\
&= [\mathrm{D}f(z_0)](z - z'),
\end{aligned}$$

where we think of z, z' as column vectors. As such,

$$\begin{aligned}
\|\mathrm{D}f(z_0)\|_{\mathrm{op}}\|z - z'\| &\geq \|[\mathrm{D}f(z_0)](z - z')\| \\
&= \|\varphi'(t_0)\| \\
&> M\|z - z'\|
\end{aligned}$$

and so

$$\sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}} > M.$$

Since M was arbitrary this means that

$$\operatorname{Lip} f \leq \sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}}$$

which gives the result. □

Proposition B.5 and Proposition B.7 then immediately entail the following:

Corollary B.8. *Suppose $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$ and $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$. If f is injective, and if f and $f^{-1} : f(\mathcal{Z}) \to \mathcal{Z}$ are everywhere differentiable, then*

$$\operatorname{BiLip} f = \max\left(\sup_{z \in \mathcal{Z}} \|\mathrm{D}f(z)\|_{\mathrm{op}}, \sup_{x \in f(\mathcal{Z})} \|\mathrm{D}f^{-1}(x)\|_{\mathrm{op}}\right).$$

B.2.1 ARZELÀ-ASCOLI

Our proof of Theorem 2.1 makes use of the Arzelà-Ascoli theorem. This is a standard and foundational result in analysis, but we include a statement here for completeness. To this end, suppose we have a sequence of functions $f_n : \mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{X}} \to \mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$. We say that (f_n) is *pointwise bounded* if, for all $z \in \mathcal{Z}$,

$$\sup_n \|f_n(z)\| < \infty.$$

Likewise, (f_n) is *uniformly equicontinuous* if for every $\epsilon > 0$ there exists $\delta > 0$ such that, for all n,

$$\|f_n(z) - f_n(z')\| < \epsilon$$

whenever $\|z - z'\| < \delta$.

Theorem B.9 (Arzelà-Ascoli)**.** *If a sequence of functions $f_n : \mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}} \to \mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ is pointwise bounded and uniformly equicontinuous, then there exists a subsequence of (f_n) that converges uniformly on every compact subset of $\mathcal{Z}$.*

Proof. The case $d = 1$ is proven for example by Rudin (2006, Theorem 11.28). This can be extended to the case $d > 1$ by a standard argument. In particular, write

$$f_n =: (f_{n,1}, \ldots, f_{n,d}),$$

where $f_{n,i} : \mathcal{Z} \to \mathbb{R}$. Then extract a subsequence (f_{n_1}) of (f_n) such that $f_{n,1}$ converges uniformly on every compact subset of $\mathcal{Z}$. Then extract a subsequence of (f_{n_1}) such that the same holds for $f_{n,2}$, and so on. The result is a subsequence $(f_{n'})$ such that each $f_{n',i}$ converges uniformly on compact subsets of $\mathcal{Z}$, from which the same holds for $f_{n'}$ also by the triangle inequality. □

B.3 Pushforward Maps Require Unbounded Bi-Lipschitz Constants

Theorem 2.1. *Suppose P_Z and $P_X^\star$ are probability measures on $\mathbb{R}^{d_\mathcal{Z}}$ and $\mathbb{R}^{d_\mathcal{X}}$ respectively, and that $\operatorname{supp} P_Z \ncong \operatorname{supp} P_X^\star$. Then for any sequence of measurable $f_n : \mathbb{R}^{d_\mathcal{Z}} \to \mathbb{R}^{d_\mathcal{X}}$, we can have $f_n \# P_Z \xrightarrow{\mathcal{D}} P_X^\star$ only if*

$$\lim_{n\to\infty} \operatorname{BiLip} f_n = \infty.$$

Proof. We suppose that $f_n \# P_Z \xrightarrow{\mathcal{D}} P_X^\star$ and prove the contrapositive. That is, without loss of generality (pass to a subsequence if necessary) we assume

$$M := \sup_n \operatorname{BiLip} f_n < \infty, \tag{B.4}$$

and prove that $\operatorname{supp} P_Z \cong \operatorname{supp} P_X^\star$.

We first show that (f_n) is pointwise bounded. To this end, observe that Prokhorov's theorem (Dudley, 2002, Proposition 9.3.4) means that P_Z is tight and that the sequence $(f_n \# P_Z)$ is uniformly tight. As such, there exists compact $K \subseteq \mathbb{R}^{d_\mathcal{Z}}$ such that $P_Z(K) > 0$, and compact $K' \subseteq \mathbb{R}^{d_\mathcal{X}}$ such that

$$\inf_n f_n \# P_Z(K') > 1 - P_Z(K).$$

For each n, we must then have some $z_n \in K$ such that $f_n(z_n) \in K'$; otherwise $K' \subseteq f_n(K)^c$ and so

$$\begin{aligned} f_n \# P_Z(K') &\leq f_n \# P_Z(f_n(K)^c) \\ &= 1 - f_n \# P_Z(f_n(K)) \\ &= 1 - P_Z(f_n^{-1}(f_n(K))) \\ &= 1 - P_Z(K) \end{aligned}$$

since f_n is injective by Proposition B.5. But for any fixed $z \in \mathbb{R}^{d_Z}$, this entails

$$\begin{aligned}\sup_n \|f_n(z)\| &\leq \sup_n \|f_n(z_n)\| + \|f_n(z) - f_n(z_n)\| \\ &\leq \sup_{x \in K'} \|x\| + \sup_{z \in K} M\|z - z_n\| \\ &\leq \sup_{x \in K'} \|x\| + 2M \sup_{z \in K} \|z\| \\ &< \infty\end{aligned}$$

since K and K' are compact.

Next, observe that (B.4) easily means (f_n) is uniformly equicontinuous. In particular, for $\epsilon > 0$, choosing $\delta := \epsilon/M$ gives

$$\|f_n(z) - f_n(z')\| \leq M\|z - z'\| < \epsilon$$

for all n whenever $\|z - z'\| < \delta$

Theorem B.9 now entails the existence of a subsequence $(f_{n'})$ that converges uniformly on every compact subset of $\mathbb{R}^{d_Z}$. In particular, $(f_{n'})$ converges pointwise to a limit that we denote by f_∞. Moreover, f_∞ is bi-Lipschitz. To see this, recall that for all n' and $z, z' \in \mathbb{R}^{d_Z}$ we have

$$\frac{1}{M}\|z - z'\| \leq \|f_{n'}(z) - f_{n'}(z')\| \leq M\|z - z'\|,$$

by our assumption (B.4). Taking $n' \to \infty$ shows that $\operatorname{BiLip} f_\infty \leq M < \infty$.

We also have that

$$f_{n'} \# P_Z \xrightarrow{\mathcal{D}} f_\infty \# P_Z. \tag{B.5}$$

This follows from the Portmanteau theorem (Dudley, 2002, Theorem 11.3.3). In particular, suppose h is a bounded Lipschitz function, and let $B_r \subseteq \mathbb{R}^{d_Z}$ denote a ball of radius $r > 0$ at the origin. Then

$$\begin{aligned}\left|\int h(x)\, f_{n'} \# P_Z(\mathrm{d}x) - \int h(x)\, f_\infty \# P_Z(\mathrm{d}x)\right| &= \left|\int h(f_{n'}(z)) - h(f_\infty(z))\, P_Z(\mathrm{d}z)\right| \\ &\leq \int_{B_r} |h(f_{n'}(z)) - h(f_\infty(z))|\, P_Z(\mathrm{d}z) \\ &\quad + \int_{B_r^c} |h(f_{n'}(z))| + |h(f_\infty(z))|\, P_Z(\mathrm{d}z) \\ &\leq P_Z(B_r)(\operatorname{Lip} h) \sup_{z \in B_r} \|f_n(z) - f_\infty(z)\| + 2P_Z(B_r^c) \sup_{z \in \mathbb{R}^{d_Z}} |h(z)|\,.\end{aligned}$$

Hence

$$\limsup_{n' \to \infty} \left|\int h(x)\, f_{n'} \# P_Z(\mathrm{d}x) - \int h(x) f_\infty \# P_Z(\mathrm{d}x)\right| \leq 2P_Z(B_r^c) \sup_{z \in \mathbb{R}^{d_Z}} |h(z)|$$

by the uniform convergence of $f_{n'}$ to f_∞ on compact subsets, and since $\operatorname{Lip} h < \infty$. Taking $r \to \infty$, the right-hand side vanishes since h is bounded, and we obtain (B.5).

We are now ready to complete the proof. Since f_∞ is bi-Lipschitz, Proposition B.5 means that f_∞ is a homeomorphism from $\mathbb{R}^{d_Z}$ to $f_\infty(\mathbb{R}^{d_Z})$. This certainly gives

$$\operatorname{supp} P_Z \cong f_\infty(\operatorname{supp} P_Z).$$

But now Proposition B.6 means

$$f_\infty(\operatorname{supp} P_Z) = \overline{f_\infty(\operatorname{supp} P_Z)}$$

where the closure is taken in $\mathbb{R}^{d_X}$. However, from (B.5) we have

$$P_X^\star = f_\infty \# P_Z,$$

which by Proposition B.3 means that

$$\operatorname{supp} P_X^\star = \operatorname{supp} f_\infty \# P_Z = \overline{f_\infty(\operatorname{supp} P_Z)}.$$

Consequently

$$\operatorname{supp} P_X^\star = f_\infty(\operatorname{supp} P_Z) \cong \operatorname{supp} P_Z$$

as desired. □

The following corollary extends the above result to the case where $\operatorname{supp} P_X^\star$ may be homeomorphic to $\operatorname{supp} P_Z$, but $P_X^\star$ is very *close* to a probability measure with non-homeomorphic support to P_Z. Here ρ denotes any metric for the weak topology. In other words, ρ must be a metric on the space of distributions that satisfies $\rho(P_n, P) \to 0$ as $n \to \infty$ if and only if $P_n \xrightarrow{\mathcal{D}} P$. The Lévy-Prokhorov and bounded Lipschitz metrics provide standard examples of such ρ (Villani, 2008, Definition 3.3.10).

Corollary 2.2. *Suppose P_Z is a probability measure on $\mathbb{R}^{d_\mathcal{Z}}$, $P_X^\star$ and P_0 are probability measures on $\mathbb{R}^{d_\mathcal{X}}$, and that $\operatorname{supp} P_0 \ncong \operatorname{supp} P_Z$. Then there exists nonincreasing $M : [0, \infty) \to [1, \infty]$ with $M(\epsilon) \to \infty$ as $\epsilon \to 0$ such that $\operatorname{BiLip} f \geq M(\epsilon)$ if $\min\left(\rho(f\#P_Z, P_X^\star), \rho(P_X^\star, P_0)\right) \leq \epsilon$.*

Proof. Let

$$M(\epsilon) := \inf\left\{\operatorname{BiLip} f \mid f : \mathbb{R}^{d_\mathcal{Z}} \to \mathbb{R}^{d_\mathcal{X}}, \rho(f\#P_Z, P_0) \leq 2\epsilon\right\}.$$

Certainly $M : [0, \infty) \to [1, \infty]$ is nonincreasing. If $\min\left(\rho(f\#P_Z, P_X^\star), \rho(P_X^\star, P_0)\right) \leq \epsilon$, then the triangle inequality gives

$$\rho(f\#P_Z, P_0) \leq \rho(f\#P_Z, P_X^\star) + \rho(P_X^\star, P_0) \leq 2\epsilon$$

and so $\operatorname{BiLip} f \geq M(\epsilon)$. It remains only to show that $M(\epsilon) \to \infty$ as $\epsilon \to 0$. For contradiction, suppose there exists $\epsilon_n \to 0$ such that $M(\epsilon_n)$ remains bounded. From the definition of M, this means that for each n there exists $f_n : \mathbb{R}^{d_\mathcal{Z}} \to \mathbb{R}^{d_\mathcal{X}}$ such that $\operatorname{BiLip} f_n \leq M(\epsilon_n) + 1/2$, and hence $\operatorname{BiLip} f_n$ remains bounded also. But this contradicts Theorem 2.1, since we assumed $\operatorname{supp} P_0 \ncong \operatorname{supp} P_Z$. □

B.4 Variance of the Russian Roulette Estimator

In this section we briefly review the Russian roulette estimator used in Chen et al. (2019), and then discuss some scenarios in which we expect the variance of this estimator to increase unboundedly.

B.4.1 Russian Roulette Estimator

Residual Flows (ResFlows, (Chen et al., 2019)), building off of Invertible Residual Networks (iResNets, (Behrmann et al., 2019)), model the data by repeatedly stacking bijections of the form $f_\ell^{-1}(x) = x + g_\ell(x)$, where $\operatorname{Lip} g_\ell =: \kappa < 1$, as mentioned in (5). The change-of-variable formula for one layer of flow reads as, for $x \in \mathbb{R}^d$,

$$\log p_X(x) = \log p_Z(f_\ell^{-1}(x)) + \operatorname{tr}\left(\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \mathrm{D} g_\ell(x)^j\right). \tag{B.6}$$

To deal with this infinite series, iResNets truncate after a fixed number of terms – this provides a biased estimate of the log-likelihood of a point x under the model. ResFlows rely on an alternative method of estimating (B.6), first using a Russian roulette procedure to rewrite the series as follows:

$$\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \operatorname{tr}\left(\mathrm{D} g_\ell(x)^j\right) = \mathbb{E}_N\left[\sum_{j=1}^{N} \frac{(-1)^{j+1}}{j} \frac{\operatorname{tr}\left(\mathrm{D} g_\ell(x)^j\right)}{p_j}\right] =: S(x),$$

where $N \sim \operatorname{Geom}(p)$ is a geometric random variable, and $p_k := \mathbb{P}(N \geq k)$. Then, taking a single sample $N \sim \operatorname{Geom}(p)$, an unbiased estimator of S is given as S_N, where S_n is defined for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$ as

$$S_n(x) := \sum_{j=1}^{N} \frac{(-1)^{j+1}}{j} \frac{\operatorname{tr}\left(\mathrm{D} g_\ell(x)^j\right)}{p_j} \tag{B.7}$$

for any $x \in \mathbb{R}^d$. We will study the variance of S_N in this section.[15]

First, however, define the quantity $\alpha_j(x)$ for $j \in \mathbb{N}, x \in \mathbb{R}^d$ as

$$\alpha_j(x) \coloneqq \frac{(-1)^{j+1}}{j} \text{tr}\left(\mathrm{D}g(x)^j\right), \tag{B.8}$$

where we now drop the dependence of g on ℓ. Then, $S(x) = \sum_{j=1}^{\infty} \alpha_j(x)$, and $S_N(x) = \sum_{j=1}^{N} \alpha_j(x)/p_j$.

B.4.2 What might happen when $\kappa \to 1$?

We begin with an informal discussion on the variance of S_N as $\kappa \to 1$. First of all we know that, as $\kappa \to 1$, the mapping f^{-1} gets arbitrarily close to a non-invertible mapping: consider e.g. $g(x) = -\kappa x$, then $f^{-1} = (1-\kappa)\text{Id} \to 0$ as $\kappa \to 1$. This near non-invertibility has implications for the speed of convergence of both $S(x)$ and its gradient,[16] as noted in these two results from Behrmann et al. (2019):

1. **Theorem 3:** $\left|\sum_{j=1}^{n} \alpha_j(x) - \log\det\left(I + \mathrm{D}g(x)\right)\right| \leq -d\left(\log(1-\kappa) + \sum_{j=1}^{n} \frac{\kappa^j}{j}\right)$,
2. **Theorem 4:** $\left\|\nabla_\theta\left(\alpha_j(x) - \log\det\left(I + \mathrm{D}g(x)\right)\right)\right\|_\infty = \mathcal{O}\left(\kappa^n\right)$.

We can see that both bounds become very loose as $\kappa \to 1$, implying we cannot guarantee the fast convergence of either series. It then follows that we cannot invoke the results from Rhee & Glynn (2015) and Beatson & Adams (2019) to argue that the variance of the Russian roulette estimator S_N will be small. Indeed, in the next section, we will look at a specific example where this variance becomes *infinite*.

B.4.3 A Specific Example of Infinite Variance

Now consider the case where $d = 1$. We will show that when $\kappa^2 > 1 - p$, there is a set of x having positive Lebesgue measure such that $S_N(x)$ from (B.7) has infinite variance.

We note that here we have $\text{tr}\left(\mathrm{D}g(x)^j\right) = (g'(x))^j$ for any $j \in \mathbb{N}$. We can thus rewrite α_j from (B.8) as

$$\alpha_j(x) \coloneqq \frac{(-1)^{j+1}}{j}(g'(x))^j. \tag{B.9}$$

Also recall that $N \sim \text{Geom}(p)$ and $p_j \coloneqq \mathbb{P}(N \geq j)$ for all $j \in \mathbb{N}$.

Proposition B.10. *For any $x \in \mathbb{R}$ and random variable N satisfying $\operatorname{supp} N = \mathbb{N}$, $S_N(x)$ has finite expectation if $\kappa < 1$.*

Proof. Refer to Lyne et al. (2015, Proposition A.1). □

Proposition B.11. *Under the same conditions as Proposition B.10,*

$$\text{Var} S_N(x) \geq \lim_{n\to\infty} 2\sum_{j=1}^{n} \alpha_j(x) S_{j-1}(x) - \mathbb{E}[S_N(x)]^2.$$

Proof. This proof is taken from Lyne et al. (2015, Proposition A.2); we mostly rewrite the proof but adapt it to our specific setting and notation. Note that we will drop the dependence of S_j and α_j on x throughout the proof.

We know from Proposition B.10 that $\mathbb{E}[S_N(x)]$ is finite. Thus we will simply lower-bound $\mathbb{E}[S_N(x)^2]$.

We will first use induction to show the following holds for any $n \in \mathbb{N}$:

$$\sum_{j=1}^{n} S_j^2(p_j - p_{j+1}) = \alpha_1^2 + \sum_{j=2}^{n} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{n} \alpha_j S_{j-1} - S_n^2 p_{n+1}. \tag{B.10}$$

[15] Chen et al. (2019) additionally approximate $\text{tr}\left(\mathrm{D}g_\ell(x)^j\right)$ by the Hutchinson's trace estimator $v^T \mathrm{D}g_\ell(x)^j v$ for $v \sim \mathcal{N}(0, I)$. Since v is independent of N, their estimator has strictly higher variance than (B.7).

[16] With respect to the flow parameters θ

The base case is

$$S_1^2(p_1 - p_2) = \frac{\alpha_1^2}{p_1^2} p_1 - S_1^2 p_2 = \alpha_1^2 - S_1^2 p_2$$

since $p_1 = 1$. Now, assume (B.10) holds for some $m \in \mathbb{N}$. Then, for $n = m + 1$,

$$\begin{aligned}\sum_{j=1}^{m+1} S_j^2(p_j - p_{j+1}) &= \sum_{j=1}^{m} S_j^2(p_j - p_{j+1}) + S_{m+1}^2(p_{m+1} - p_{m+2}) \\ &= \alpha_1^2 + \sum_{j=2}^{m} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{m} \alpha_j S_{j-1} - S_m^2 p_{m+1} \\ &\quad + S_{m+1}^2(p_{m+1} - p_{m+2})\end{aligned} \tag{B.11}$$

by the inductive hypothesis. We also have

$$\begin{aligned}p_{m+1}(S_m^2 - S_{m+1}^2) &= p_{m+1}(S_m - S_{m+1})(S_m + S_{m+1}) \\ &= p_{m+1}\frac{\alpha_{m+1}}{p_{m+1}}\left(2S_m + \frac{\alpha_{m+1}}{p_{m+1}}\right) \\ &= \frac{\alpha_{m+1}^2}{p_{m+1}} + 2\alpha_{m+1}S_m.\end{aligned}$$

Substituting this result into (B.11) completes the induction and proves (B.10) for all $n \in \mathbb{N}$.

Now, by Jensen's inequality,

$$S_n^2 = \left(\sum_{j=1}^{n} p_j \frac{\frac{\alpha_j}{p_j}}{p_j}\right)^2 \leq \frac{\sum_{j=1}^{n} \frac{\alpha_j^2}{p_j}}{\sum_{j=1}^{n} p_j}.$$

This implies

$$p_{n+1}S_n^2 \leq p_n S_n^2 \leq \frac{p_n}{\sum_{j=1}^{n} p_j} \sum_{j=1}^{n} \frac{\alpha_j^2}{p_j} \leq \sum_{j=1}^{n} \frac{\alpha_j^2}{p_j}$$

since (p_n) is a positive sequence.

This finally implies the following lower bound for any $n \in \mathbb{N}$:

$$\begin{aligned}\sum_{j=1}^{n} S_j^2 \mathbb{P}(N = j) &= \sum_{j=1}^{n} S_j^2(p_j - p_{j+1}) \\ &= \alpha_1^2 + \sum_{j=2}^{n} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{n} \alpha_j S_{j-1} - S_n^2 p_{n+1} \\ &\geq \alpha_1^2 + \sum_{j=2}^{n} \frac{\alpha_j^2}{p_j} + 2\sum_{j=2}^{n} \alpha_j S_{j-1} - \sum_{j=1}^{n} \frac{\alpha_j^2}{p_j} \\ &= \alpha_1^2(1 - p_1^{-1}) + 2\sum_{j=2}^{n} \alpha_j S_{j-1} \\ &= 2\sum_{j=2}^{n} \alpha_j S_{j-1},\end{aligned}$$

where the final line follows because $p_1 = 1$.

Since $\mathbb{E}[S_N^2] = \lim_{n\to\infty} \sum_{j=1}^{n} S_j^2 \mathbb{P}(N = j)$, the proof is complete. □

We are about ready to prove the main result but require one more auxiliary result first.

Proposition B.12. *Suppose* $|b| > 1$. *Then,*

$$\lim_{n\to\infty} \frac{n}{b^n} \sum_{j=1}^{n-1} \frac{b^j}{j} = \frac{1}{b-1}.$$

Proof. We will first show that the limit exists, and then show that it equals $(b-1)^{-1}$. Let

$$c_n = \frac{n}{b^n} \sum_{j=1}^{n-1} \frac{b^j}{j}.$$

We can rewrite this as follows:

$$c_n = \sum_{j=1}^{n-1} \frac{n}{b^{n-j} j} = \sum_{j=1}^{n-1} \frac{n}{b^j (n-j)} = \sum_{j=1}^{n-1} \frac{1}{b^j} + \sum_{j=1}^{n-1} \frac{j}{b^j (n-j)}.$$

Since $b > 1$, the first sum is a convergent geometric series as $n \to \infty$. We can decompose the second sum into its positive and negative terms:

$$\sum_{j=1}^{n-1} \frac{j}{b^j (n-j)} = \sum_{j \geq 1 : b^j > 0}^{n-1} \frac{j}{b^j (n-j)} + \sum_{j \geq 1 : b^j < 0}^{n-1} \frac{j}{b^j (n-j)} \equiv ①_n + ②_n.$$

We can see, for all $n \in \mathbb{N}$,

$$①_n \geq -\sum_{j=1}^{n-1} \frac{j}{|b|^j (n-j)} \qquad \text{and} \qquad ②_n \leq \sum_{j=1}^{n-1} \frac{j}{|b|^j (n-j)}.$$

Furthermore, for all $j \in \{1, \ldots, n-1\}$, we have

$$\frac{j}{n-j} \leq j.$$

Now notice that the series $\sum_{j=1}^{\infty} \frac{j}{|b|^j}$ converges by the ratio test:

$$\lim_{j\to\infty} \left| \frac{\frac{j+1}{|b|^{j+1}}}{\frac{j}{|b|^j}} \right| = \lim_{j\to\infty} \frac{j+1}{j|b|} = \frac{1}{|b|} < 1.$$

This implies the existence of $\lim_{n\to\infty} \sum_{j=1}^{n-1} \frac{j}{|b|^j (n-j)}$. Since the sequence $(①_n)$ (resp. $(②_n)$) is negative, non-increasing, and bounded below (resp. positive, non-decreasing, and bounded above), this implies the existence of $\lim_{n\to\infty} ①_n$ (resp. $\lim_{n\to\infty} ②_n$). Altogether, this implies the existence of

$$\lim_{n\to\infty} \left(\sum_{j=1}^{n-1} \frac{1}{b^j} + \sum_{j=1}^{n-1} \frac{j}{b^j (n-j)} \right) = \lim_{n\to\infty} c_n =: c_\infty.$$

Now we will determine its precise value. Note the following recurrence for all $n \in \mathbb{N}$:

$$c_{n+1} = \frac{n+1}{bn} (1 + c_n).$$

Taking the limit of both sides as $n \to \infty$ gives

$$c_\infty = \frac{1}{b}(1 + c_\infty).$$

Solving this gives us $c_\infty = \frac{1}{b-1}$, which completes the proof. □

Proposition B.13. *Suppose $N \sim Geom(p)$, g is continuously differentiable, and $1 - p < \kappa^2 < 1$. Then*

$$\{x \in \mathbb{R} \mid \mathrm{Var} S_N(x) = \infty\}$$

has positive Lebesgue measure.

Proof. From Proposition B.11, for a given $x \in \mathbb{R}$, we can see that showing $\sum_{n=2}^{\infty} \alpha_n(x) S_{n-1}(x)$ diverges is sufficient to prove $\mathrm{Var} S_N(x)$ is infinite.

Consider using the ratio test to assess the convergence of the above series, with terms defined as $a_n(x) \coloneqq \alpha_n(x) S_{n-1}(x)$. We have the following for any $n \geq 2$:

$$\begin{aligned}
\left|\frac{a_{n+1}(x)}{a_n(x)}\right| &= \left|\frac{\alpha_{n+1}(x) S_n(x)}{\alpha_n(x) S_{n-1}(x)}\right| \\
&= \frac{\frac{|(g'(x))^{n+1}|}{n+1}}{\frac{|(g'(x))^n|}{n}} \cdot \left|\frac{\sum_{j=1}^{n} \frac{\alpha_j(x)}{p_j}}{\sum_{j=1}^{n-1} \frac{\alpha_j(x)}{p_j}}\right| \\
&= \frac{n|g'(x)|}{n+1} \cdot \left|\frac{(-1)^{n+1} \cdot (g'(x))^n}{np_n} \left(\sum_{j=1}^{n-1} \frac{(-1)^{j+1} \cdot (g'(x))^j}{jp_j}\right)^{-1} + 1\right|.
\end{aligned}$$

Recall $p_j = (1-p)^{j-1} \equiv q^{j-1}$. Then, writing $b = -\frac{g'(x)}{q}$, we have

$$\frac{(-1)^{n+1} \cdot (g'(x))^n}{np_n} \left(\sum_{j=1}^{n-1} \frac{(-1)^{j+1} \cdot (g'(x))^j}{jp_j}\right)^{-1} = \frac{1}{n} b^n \left(\sum_{j=1}^{n-1} \frac{1}{j} b^j\right)^{-1}.$$

Now let us assume that $|g'(x)|^2 > q$. We can see that $|g'(x)|^2 > q \implies |g'(x)| > q$ since $q \in (0,1)$, which then entails $|b| > 1$. Therefore, by Proposition B.12,

$$\lim_{n\to\infty} \frac{n}{b^n} \sum_{j=1}^{n-1} \frac{b^j}{j} = \frac{1}{b-1}.$$

This then implies

$$\begin{aligned}
\lim_{n\to\infty} \left|\frac{a_{n+1}(x)}{a_n(x)}\right| &= \lim_{n\to\infty} \frac{n|g'(x)|}{n+1} \left|\frac{1}{\frac{n}{b^n}\sum_{j=1}^{n-1} \frac{b^j}{j}} + 1\right| \\
&= |g'(x)| \left|\frac{1}{\frac{1}{b-1}} + 1\right| = \frac{|g'(x)|^2}{q} > 1
\end{aligned}$$

since we have assumed that $|g'(x)|^2 > q$. Thus, for all x in the set

$$V_{g,q} \coloneqq \{x \in \mathbb{R} \mid |g'(x)|^2 > q\},$$

the series $\sum_{n=2}^{\infty} \alpha_n(x) S_{n-1}(x)$ diverges by the ratio test. This means that $\mathrm{Var} S_N(x) = \infty$ for all $x \in V_{g,q}$.

Finally, we will prove the set $\{x \in \mathbb{R} \mid \mathrm{Var} S_N(x) = \infty\}$ has positive Lebesgue measure. Recall that $\mathrm{Lip}\, g = \kappa$, which directly implies $\sup_{x\in\mathbb{R}} |g'(x)| = \kappa$ from Proposition B.7 and thus $\sup_{x\in\mathbb{R}} |g'(x)|^2 = \kappa^2$. Then, since $\kappa^2 > q$, there exists $x_0 \in \mathbb{R}$ such that $|g'(x_0)|^2 \in (q, \kappa^2)$. By the continuity of $|g'|$, there is open ball of nonzero radius around x_0, denoted $\mathcal{B}(x_0)$, such that $|g'(x)| > q$ for all $x \in \mathcal{B}(x_0)$. Since $\mathcal{B}(x_0)$ is open and non-empty, it has positive Lebesgue measure. The inclusions

$$\mathcal{B}(x_0) \subseteq V_{g,q} \subseteq \{x \in \mathbb{R} \mid \mathrm{Var} S_N(x) = \infty\}$$

thus conclude the proof.

□

B.4.4 DISCUSSION

Changing p as κ increases An obvious strategy to avoid satisfying the conditions of Proposition B.13 is to set p such that $1 - \kappa^2 > p$. However, lowering p in this way incurs additional computational cost: the average number of iterations per training step is equal to p^{-1}, or is lower-bounded by $(1 - \kappa^2)^{-1}$ if $p < 1 - \kappa^2$. Thus, if we send $\kappa \to 1$ to mitigate the bi-Lipschitz constraint (6), we will either incur an infinite computational cost or run the risk of encountering infinite variance.

Higher dimensions Although Proposition B.13 only applies for $d = 1$, it is conceivable that similar results can be derived for $d > 1$, especially when considering the discussion in Section B.4.2. We leave a deeper investigation for future work.

B.5 Density of a CIF

We make precise our heuristic derivation of the density (11) via the following result.

Proposition B.14. *Suppose $\mathcal{Z}, \mathcal{X} \subseteq \mathbb{R}^d$ are open, and that $F(\cdot; u) : \mathcal{Z} \to \mathcal{X}$ is a continuously differentiable bijection with everywhere invertible Jacobian for each $u \in \mathcal{U}$. Under the generative model (8), (X, U) has joint density*

$$p_Z(F^{-1}(x;u))\, p_{U|Z}(u|F^{-1}(x;u))\, |\det \mathrm{D}F^{-1}(x;u)|.$$

Proof. Suppose $h : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is a bounded measurable test function. Then

$$\begin{aligned}
\mathbb{E}[h(X,U)] &= \mathbb{E}[h(F(Z;U),U)] \\
&= \int \left[\int h(F(z;u),u)\, p_Z(z)\, p_{U|Z}(u|z)\, \mathrm{d}z \right] \mathrm{d}u \\
&= \int h(x,u)\, p_Z(F^{-1}(x;z)) p_{U|Z}(u|F^{-1}(x;z))\, |\det \mathrm{D}F^{-1}(x;u)|\, \mathrm{d}z\, \mathrm{d}u,
\end{aligned}$$

where in the third line we substitute $x := F(z;u)$ on the inner integral, which is valid by Theorem 17.2 of Billingsley (2008). Now for $A \subseteq \mathcal{X} \times \mathcal{U}$, let $h := \mathbb{I}_A$. It follows that

$$\mathbb{P}((X,U) \in A) = \mathbb{E}[\mathbb{I}_A(X,U)] = \int_A p_Z(F^{-1}(x;z))\, p_{U|Z}(u|F^{-1}(x;z))\, |\det \mathrm{D}F^{-1}(x;u)|\, \mathrm{d}z\, \mathrm{d}u,$$

which gives the result since A was arbitrary. □

B.6 Our Approximate Posterior Does Not Sacrifice Generality

The following result shows that our parameterisation of the approximate posterior $q_{U_{1:L}|X}$ in (15) does not lose generality. In particular, provided each $q_{U_\ell|Z_\ell}$ is sufficiently expressive, we can always recover the exact posterior.

Proposition B.15. *Under the generative model (8), the posterior factors like*

$$p_{U_{1:L}|X}(u_{1:L}|x) = \prod_{\ell=1}^{L} p_{U_\ell|Z_\ell}(u_\ell|z_\ell),$$

where $z_L := x$ and $z_\ell := F_{\ell+1}^{-1}(z_{\ell+1}; u_{\ell+1})$ for $\ell \in \{1, \ldots, L-1\}$.

Proof. Writing $p_{U_{1:L}|X}$ autoregressively gives

$$p_{U_{1:L}|X}(u_{1:L}|x) = \prod_{\ell=1}^{L} p_{U_\ell|U_{\ell+1:L},X}(u_\ell|u_{\ell+1:L},x).$$

But now it is clear from the generative model (8) that U_ℓ is conditionally independent of $(U_{\ell+1:L}, X)$ given Z_ℓ, and as such

$$p_{U_\ell|U_{\ell+1:L},X}(u_\ell|u_{\ell+1:L},x) = p_{U_\ell|Z_\ell}(u_\ell|z_\ell).$$

Substituting this into the above expression then gives the result. □

B.7 Conditions for a CIF to Outperform an Underlying Normalising Flow

For this result, the components of our model are assumed to be parameterised by $\theta \in \Theta$, which we will indicate by by F_θ, $p^\theta_{U|Z}$, and $q^\theta_{U|X}$. We will also use θ to indicate quantities that result from the choice of parameters θ (e.g. P^θ_X for the distribution obtained), and will denote by $\mathcal{L}^\theta$ the corresponding ELBO (14).

Proposition 4.1. *Suppose there exists $\phi \in \Theta$ such that, for some bijection $f : \mathcal{Z} \to \mathcal{X}$, $F_\phi(\cdot; u) = f(\cdot)$ for all $u \in \mathcal{U}$. Likewise, suppose $p^\phi_{U|Z}$ and $q^\phi_{U|X}$ are such that, for some density r on $\mathcal{U}$, $p^\phi_{U|Z}(\cdot|z) = q^\phi_{U|X}(\cdot|x) = r(\cdot)$ for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. If $\mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\theta(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)]$, then*

$$D_{\mathrm{KL}}\big(P^\star_X \,\big\|\, P^\theta_X\big) \leq D_{\mathrm{KL}}(P^\star_X \,\|\, f\#P_Z).$$

Proof. Observe from (11) that

$$p^\phi_{X,U}(x,u) = p_Z(f^{-1}(x))\,|\det \mathrm{D}f^{-1}(x)|\,p_{U|Z}(u|f^{-1}(x)).$$

It then follows from (13) that, under ϕ, the model has density

$$p^\phi_X(x) = p_Z(f^{-1}(x))\,|\det \mathrm{D}f^{-1}(x)| \underbrace{\int p_{U|Z}(u|f^{-1}(x))\,\mathrm{d}u}_{=1}$$

which is exactly the density of the normalising flow $f\#P_Z$. We also obtain the posterior

$$\begin{aligned} p^\phi_{U|X}(u|x) &= \frac{p^\phi_{X,U}(x,u)}{p^\phi_X(x)} \\ &= p_{U|Z}(u|f^{-1}(x)) \\ &= r(u). \end{aligned}$$

Since each $q^\phi_{U|X}(\cdot|x) = r(\cdot)$ also, it follows that $\mathcal{L}^\phi$ is tight, so that $\mathcal{L}^\phi(x) = \log p^\phi_X(x)$ for all $x \in \mathcal{X}$.

Now suppose some $\theta \in \Theta$ has

$$\mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\theta(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)].$$

It follows that

$$\mathbb{E}_{x\sim P^\star_X}[\log p^\theta_X(x)] \geq \mathbb{E}_{x\sim P^\star_X}[\mathcal{L}^\phi(x)] = \mathbb{E}_{x\sim P^\star_X}[\log p^\phi_X(x)].$$

Subtracting $\mathbb{E}_{x\sim P^\star_X}[\log p^\star_X(x)]$ from both sides and negating gives

$$D_{\mathrm{KL}}\big(P^\star_X \,\big\|\, P^\theta_X\big) \leq D_{\mathrm{KL}}\Big(P^\star_X \,\Big\|\, P^\phi_X\Big) = D_{\mathrm{KL}}(P^\star_X \,\|\, f\#P_Z).$$

□

B.8 CIFs Can Learn Target Supports Exactly

In this section we give necessary and sufficient conditions for a CIF to learn the support of a target distribution exactly, without needing changes to F. However, our argument applies more generally and does not make specific use of the bijective structure of F. To make this clear, we formulate our result here in terms of a generalisation of the model (7). In particular, we will take P_X as the marginal in X of

$$Z \sim P_Z, \quad U \sim P_{U|Z}(\cdot|Z), \quad X := G(Z,U), \tag{B.12}$$

where $G : \mathcal{Z} \times \mathcal{U} \to \mathcal{X}$. We will assume that

- $\mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}}$, $\mathcal{U} \subseteq \mathbb{R}^{d_\mathcal{U}}$, and $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$ are equipped with the subspace topology;
- P_Z and each $P_{U|Z}(\cdot|z)$ are Borel probability measures on $\mathcal{Z}$ and $\mathcal{U}$ respectively;

- G is continuous with respect to the product topology $\mathcal{Z} \times \mathcal{U}$.

We then have the following formula for $\operatorname{supp} P_X$:

Lemma B.16. *Under the model (B.12),*

$$\operatorname{supp} P_X = \overline{\bigcup_{z \in \operatorname{supp} P_Z} G(\{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z))}.$$

Proof. Denote the joint distribution of (Z, U) by $P_{Z,U}$. Observe from Proposition B.3 that

$$\operatorname{supp} P_X = \overline{G(\operatorname{supp} P_{Z,U})}.$$

Let

$$B := \bigcup_{z \in \operatorname{supp} Z} \{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z).$$

The result follows if we can show that

$$\operatorname{supp} P_{Z,U} = \overline{B},$$

since $\overline{G(\overline{B})} = \overline{G(B)}$ because G is continuous.

We first show that $\operatorname{supp} P_{Z,U} \supseteq \overline{B}$. Suppose $(z, u) \in B$, and let $N_{(z,u)} \subseteq \mathcal{Z} \times \mathcal{U}$ be an open set containing (z, u). Then there exists open N_z and N_u containing z and u respectively such that $N_z \times N_u \subseteq N_{(z,u)}$, since the open rectangles form a base for the product topology. It follows that

$$\begin{aligned} P_{Z,U}(N_{(z,u)}) &\geq P_{Z,U}(N_z \times N_u) \\ &= \int_{N_z} P_{U|Z}(N_u|z')\, P_Z(\mathrm{d}z') \\ &> 0, \end{aligned}$$

since by the definition of B we have $P_Z(N_z) > 0$ and $P_{U|Z}(N_u|z) > 0$ for each $u \in N_z$. From this we have $\operatorname{supp} P_{Z,U} \supseteq B$, and taking the closure of each side gives $\operatorname{supp} P_{Z,U} \supseteq \overline{B}$.

In the other direction, suppose that $(z, u) \notin \overline{B}$. Then there exist open sets N_z and N_u containing z and u respectively such that

$$(N_z \times N_u) \cap \overline{B} = \emptyset.$$

By the definition of B, it follows that if $(z', u') \in N_z \times N_u$ and $z' \in \operatorname{supp} P_Z$, then $u' \notin \operatorname{supp} P_{U|Z}(\cdot|z')$. Otherwise stated, if $z' \in N_z \cap \operatorname{supp} P_Z$, then

$$N_u \cap \operatorname{supp} P_{U|Z}(\cdot|z') = \emptyset.$$

Thus

$$\begin{aligned} P_{Z,U}(N_z \times N_u) &= \int_{N_z} \left[\int_{N_u} P_{U|Z}(\mathrm{d}u'|z') \right] P_Z(\mathrm{d}z') \\ &= \int_{N_z \cap \operatorname{supp} P_Z} \left[\int_{N_u \cap \operatorname{supp} P_{U|Z}(\cdot|z')} P_{U|Z}(\mathrm{d}u'|z') \right] P_Z(\mathrm{d}z') \\ &= 0, \end{aligned}$$

where the second line follows from Proposition B.4. Consequently $(z, u) \notin \operatorname{supp} P_{Z,U}$, which gives $\operatorname{supp} P_{Z,U} \subseteq \overline{B}$. □

We now give necessary and sufficient conditions for the model (B.12) to learn a given target support exactly.

Proposition B.17. *Suppose $P_X^\star(\partial \operatorname{supp} P_X^\star) = 0$ and that*

$$\overline{G(\operatorname{supp} P_Z \times \mathcal{U})} \supseteq \operatorname{supp} P_X^\star. \tag{B.13}$$

Then there exists $P_{U|Z}$ such that $\operatorname{supp} P_X = \operatorname{supp} P_X^\star$ if and only if, for all $z \in \operatorname{supp} P_Z$, there exists $u \in \mathcal{U}$ with

$$G(z, u) \in \operatorname{supp} P_X^\star.$$

Proof. ($\Rightarrow$) Choose $P_{U|Z}$ such that $\operatorname{supp} P_X = \operatorname{supp} P_X^\star$. Lemma B.16 gives

$$\bigcup_{z \in \operatorname{supp} P_Z} G(\{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z)) \subseteq \operatorname{supp} P_X^\star.$$

Suppose $z \in \operatorname{supp} P_Z$. Then for indeed all $u \in \operatorname{supp} P_{U|Z}(\cdot|z)$ we must have $G(z,u) \in \operatorname{supp} P_X^\star$, which proves this direction since $\operatorname{supp} P_{U|Z}(\cdot|z) \neq 0$ by Proposition B.4.

($\Leftarrow$) For $z \in \operatorname{supp} P_Z$, let

$$A_z := \{u \in \mathcal{U} : G(z,u) \in \operatorname{int}(\operatorname{supp} P_X^\star)\}, \tag{B.14}$$

where int denotes the *interior* operator. If $A_z = \emptyset$, define $P_{U|Z}(\cdot|z)$ to be Dirac on some u such that $G(z,u) \in \operatorname{supp} P_X^\star$, which exists by assumption. Otherwise, we let $P_{U|Z}(\cdot|z)$ be a probability measure with support $\overline{A_z}$. To show that such a measure exists, observe that A_z is open since G is continuous. Since $\mathcal{U}$ is separable, we can therefore write

$$A_z = \bigcup_{n=1}^{\infty} B_n$$

for a countable collection of open sets $B_n \subseteq A_z$. We can then define a probability measure μ by

$$\mu(C) := \sum_{n=1}^{\infty} 2^{-n}\, \mathbb{I}(C \cap B_n \neq \emptyset)$$

for measurable $C \subseteq \mathcal{U}$. Since $A_z \neq \emptyset$, it is straightforward to see that this is a probability measure with $\mu(A_z) = 1$. Consequently $\operatorname{supp} \mu = \overline{A_z}$ by Proposition B.2, since A_z is open and $\operatorname{supp} \mu$ is the smallest closed set with μ-probability 1.

We show this construction gives $\operatorname{supp} P_X \subseteq \operatorname{supp} P_X^\star$. To this end, we first prove that if $z \in \operatorname{supp} P_Z$ and $u \in \operatorname{supp} P_{U|Z}(\cdot|z)$ then

$$G(z,u) \in \operatorname{supp} P_X^\star.$$

If $A_z = \emptyset$ this is immediate. Otherwise, since $\operatorname{supp} P_{U|Z}(\cdot|z) = \overline{A_z}$, there exists $(u_n) \subseteq A_z$ such that $u_n \to u$. By (B.14), each $G(z,u_n) \in \operatorname{supp} P_X^\star$. By continuity we then have

$$G(z,u_n) \to G(z,u) \in \operatorname{supp} P_X^\star$$

since $\operatorname{supp} P_X^\star$ is closed. It follows that

$$\bigcup_{z \in \operatorname{supp} P_Z} G(\{z\} \times \operatorname{supp} P_{U|Z}(\cdot|z)) \subseteq \operatorname{supp} P_X^\star,$$

which gives $\operatorname{supp} P_X \subseteq \operatorname{supp} P_X^\star$ from Lemma B.16 since $\operatorname{supp} P_X^\star$ is closed.

We now show $\operatorname{supp} P_X \supseteq \operatorname{supp} P_X^\star$. Since $P_X^\star(\partial \operatorname{supp} P_X^\star) = 0$ we have

$$\operatorname{supp} P_X^\star = \overline{\operatorname{int}(\operatorname{supp} P_X^\star)}$$

by Proposition B.2, so that $\operatorname{supp} P_X \supseteq \operatorname{supp} P_X^\star$ if $\operatorname{supp} P_X \supseteq \operatorname{int}(\operatorname{supp} P_X^\star)$. Now suppose $x \in \operatorname{int}(\operatorname{supp} P_X^\star)$. Then there exists $(z_n) \subseteq \operatorname{supp} P_Z$ and $(u_n) \subseteq \mathcal{U}$ such that $G(z_n,u_n) \to x$ by (B.13). But then we must have $G(z_n,u_n) \in \operatorname{int}(\operatorname{supp} P_X^\star)$ for n large enough because x lies in the interior. Consequently, for n large enough,

$$u_n \in A_{z_n} \subseteq \operatorname{supp} P_{U|Z}(\cdot|z_n)$$

and hence $G(z_n,u_n) \in \operatorname{supp} P_X$ by Lemma B.16. This means $x \in \operatorname{supp} P_X$ since $\operatorname{supp} P_X$ is closed. □

The following proposition then gives a straightforward condition under which it is additionally possible to recover the *target* exactly (i.e. not just its support). In our experiments we do not enforce this condition explicitly. However, since we learn the parameters of G here, we can expect our model will approximate this behaviour if doing so produces a better density estimator.

Proposition B.18. *If $G(z, \cdot)$ is surjective for each $z \in \mathcal{Z}$, then there exists $P_{U|Z}$ such that $P_X = P_X^\star$.*

Proof. Fix $z \in \mathcal{Z}$. Surjectivity of $G(z, \cdot)$ means that, for $x \in \mathcal{X}$, there exists $u \in \mathcal{U}$ such that $G(z, u) = x$. Thus we can define $H_z : \mathcal{X} \to \mathcal{U}$ such that

$$G(z, H_z(x)) = x$$

for all $x \in \mathcal{X}$. We then define each

$$P_{U|Z}(\cdot|z) := H_z \# P_X^\star.$$

From this it follows that $P_X = P_X^\star$. For, letting $B \subseteq \mathcal{X}$ be measurable,

$$\begin{aligned}
P_X(B) &= \int_{G^{-1}(B)} P_{U|Z}(\mathrm{d}u|z)\, P_Z(\mathrm{d}z) \\
&= \int \left[\int \mathbb{I}_B(G(z,u))\, H_z \# P_X^\star(\mathrm{d}u) \right] P_Z(\mathrm{d}z) \\
&= \int \left[\int \mathbb{I}_B(G(z, H_z(x)))\, P_X^\star(\mathrm{d}x) \right] P_Z(\mathrm{d}z) \\
&= \int \left[\int \mathbb{I}_B(x)\, P_X^\star(\mathrm{d}x) \right] P_Z(\mathrm{d}z) \\
&= P_X^\star(B),
\end{aligned}$$

which gives the result. □

C Experimental Details

Our choices (10) and (17) required parameterising s, t, μ^p, Σ^p, μ^q, and Σ^q. Since these terms are naturally paired, at each layer of our model we set

$$\begin{aligned}
[s(u), t(u)] &:= \mathtt{NN}_F(u), \\
[\mu^p(z), \varsigma^p(z)] &:= \mathtt{NN}_p(z), \\
\Sigma^p(z) &:= \mathrm{diag}(e^{\varsigma^p(z)}), \\
[\mu^q(x), \varsigma^q(x)] &:= \mathtt{NN}_q(x), \\
\Sigma^q(x) &:= \mathrm{diag}(e^{\varsigma^q(x)}),
\end{aligned}$$

where $\mathtt{NN}$ denotes a separate neural network and $\varsigma^p(z), \varsigma^q(x) \in \mathbb{R}^d$.

In all experiments we trained our models to maximise either the log-likelihood (for the baseline flows) or the ELBO (for the CIFs) using the ADAM optimiser (Kingma & Ba, 2015) with default hyperparameters and no weight decay. The ELBO was estimated using a single sample per datapoint (i.e. a single call to Algorithm 1). We used a held-out validation set and trained each model until its validation score stopped improving, except for the NSF tabular data experiments where we train for a fixed number of epochs as specified in Durkan et al. (2019). After training, we used validation performance to select the best parameters found during training for use at test time (again except for the NSF experiments, where we just test with the final model). Both validation and test scores were computed using the exact log-likelihood for the baseline and the importance sampling estimate (16) for the CIFs, with $m = 5$ samples for validation and $m = 100$ for testing.

C.1 Tabular Data Experiments

Following Papamakarios et al. (2017), we experimented with the POWER, GAS, HEPMASS, and MINIBOONE datasets from the UCI repository (Bache & Lichman, 2013), as well as a dataset of 8×8 image patches extracted from the BSDS300 dataset (Martin et al., 2001). We preprocessed these datasets identically to Papamakarios et al. (2017), and used the same train/validation/test splits. For all CIF-ResFlow models, we used a batch size of 1000 and a learning rate of 10^{-3}. For the MAF experiments, we used a batch size of 1000 and a learning rate of 10^{-3}, except for BSDS300 where we used a learning rate of 10^{-4} to control the instability of the baseline. For the NSF experiments, we used batch sizes and learning rates as dictated by Durkan et al. (2019, Table 5), along with their cosine learning rate annealing scheme.

Table C.4: MAF and CIF-MAF parameter configurations for POWER and GAS.

	LAYERS (L)	AUTOREGRESSIVE NETWORK SIZE	$\texttt{NN}_p$ SIZE	$\texttt{NN}_q$ SIZE	$\texttt{NN}_F$ SIZE
MAF	5, 10, 20	2×100, 2×200, 2×400	-	-	-
CIF-MAF	5, 10	2×128	$2 \times 100, 2 \times 200$	$2 \times 100, 2 \times 200$	2×128

Table C.5: MAF and CIF-MAF parameter configurations for HEPMASS and MINIBOONE

	LAYERS (L)	AUTOREGRESSIVE NETWORK SIZE	$\texttt{NN}_p$ SIZE	$\texttt{NN}_q$ SIZE	$\texttt{NN}_F$ SIZE
MAF	5, 10, 20	2×128, 2×512, 2×1024	-	-	-
CIF-MAF	5, 10	2×128	$2 \times 128, 2 \times 512$	$2 \times 128, 2 \times 512$	2×128

Also, for all CIF models, each U_ℓ had the same dimension $d_{\mathcal{U}}$, which we took to be roughly a quarter of the dimensionality of the data (except in Section C.1.4 for which $d_{\mathcal{U}} = d_{\mathcal{X}}$). In particular, we set $d_{\mathcal{U}} := 2$ for POWER and GAS, $d_{\mathcal{U}} := 5$ for HEPMASS, $d_{\mathcal{U}} := 10$ for MINIBOONE, and $d_{\mathcal{U}} := 15$ for BSDS300.

C.1.1 RESIDUAL FLOWS

The residual blocks in all ResFlow models used multilayer perceptrons (MLPs) with 4 hidden layers of 128 hidden units (denoted 4×128), LipSwish nonlinearities (Chen et al., 2019, (10)) before each linear layer, and a residual connection from the input to the output. We did not use any kind of normalisation (e.g. ActNorm or BatchNorm) for these experiments. For all models we set $\kappa = 0.9$ in (5) to match the value for the 2-D experiments in the codebase of Chen et al. (2019). Other design choices followed Chen et al. (2019). In particular:

- We always exactly computed several terms at the beginning of the series expansion of the log Jacobian, and then used Russian Roulette sampling (Kahn, 1955) to estimate the sum of the remaining terms. In particular, at training time we computed 2 exact terms, while at test time we computed 20 exact terms;
- We used a geometric distribution with parameter 0.5 for the number of terms to compute in our Russian Roulette estimators;
- We used the Skilling-Hutchinson trace estimator (Skilling, 1989; Hutchinson, 1990) to estimate the trace in the log Jacobian term;
- At both training and test time, we used a single Monte Carlo sample of (n, v) to estimate (6) of Chen et al. (2019);

However, note that for these experiments, for the sake of simplicity, we did not use the memory-saving techniques in (8) and (9) of Chen et al. (2019), nor the adaptive power iteration scheme described in their Appendix E.

For $\texttt{NN}_F$, $\texttt{NN}_p$, and $\texttt{NN}_q$ we used 2×10 MLPs with $\tanh$ nonlinearities These networks were much smaller than 4×128, and hence the CIF-ResFlows had only roughly 1.5-4.5% more parameters (depending on the dimension of the dataset) than the otherwise identical 10-layer ResFlows, and roughly 10% of the parameters of the 100-layer ResFlows.

The 100-layer ResFlows were significantly slower to train than the 10-layer models, and for POWER, GAS, and BSDS300 we were forced to stop these before their validation loss had converged. However, to ensure a fair comparison, we allocated more total computing power to these models than to the 10-layer models, which were terminated properly. In particular, we trained each 100-layer ResFlow on POWER and GAS for a total of 10 days on a single NVIDIA GeForce GTX 1080 Ti, and on BSDS300 for a total of 7 days. In contrast, the 10-layer ResFlows converged after around 1 day on POWER, 4.5 days on GAS, and around 3 days on BSDS300. Likewise, the 10-layer CIF-ResFlows converged after around 1 day on POWER, 6 days on GAS, and 2 days on BSDS300.

C.1.2 MASKED AUTOREGRESSIVE FLOWS

The experiment comparing MAF baselines to CIF-MAFs was inspired by the experimental setup in Papamakarios et al. (2017). For each dataset, we specified a set of hyperparameters over which to search for both the baselines and the CIFs;

Table C.6: MAF and CIF-MAF parameter configurations for BSDS300

	LAYERS (L)	AUTOREGRESSIVE NETWORK SIZE	$\texttt{NN}_p$ SIZE	$\texttt{NN}_q$ SIZE	$\texttt{NN}_F$ SIZE
MAF	5, 10, 20	2×512, 2×1024, 2×2048	-	-	-
CIF-MAF	5, 10	2×512	$2 \times 128, 2 \times 512$	$2 \times 128, 2 \times 512$	2×128

Table C.7: CIF-NSF configurations for all tabular datasets. The number of hidden features in the autoregressive network is referred to as n_h.

	$\texttt{NN}_p$ SIZE	$\texttt{NN}_q$ SIZE	$\texttt{NN}_F$ SIZE	n_h VS. BASELINE
CIF-NSF-1 (MINIBOONE)	3×50	2×10	3×25	FEWER
CIF-NSF-1 (NON-MINIBOONE)	3×200	2×10	3×100	FEWER
CIF-NSF-2	3×200	2×10	3×100	SAME

these hyperparameters are provided in Table C.4, Table C.5, and Table C.6. Then, we trained each model until no validation improvement had been observed for 50 epochs. We then evaluated the model with the best validation score among all candidate models on the test dataset to obtain a log-likelihood score. We performed this procedure with three separate random seeds, and report the average and standard error across the runs in Table 1.

We searched over all combinations of parameters listed in Table C.4, Table C.5, and Table C.6. For example, on HEPMASS or MINIBOONE, our set of candidate MAF models included: for $L = 5$, an autoregressive network of size of either 2×128, 2×512, or 2×1024; for $L = 10$, an autoregressive network size of either 2×128, 2×512, or 2×1024; and for $L = 20$, again an autoregressive network size of either 2×128, 2×512, or 2×1024; this gave us a total of 9 candidate MAF models for each seed. The set of candidate CIF-MAF models can similarly be determined via the table and gave us a total of 8 candidate models for each seed. We maintained this split of 9 candidates for MAF and 8 candidates for CIF-MAF across datasets to fairly compare against the baseline by allowing them more configurations. We also considered deeper and wider MAF models to compensate for the additional parameters introduced by $\texttt{NN}_F$, $\texttt{NN}_p$, and $\texttt{NN}_q$ in the CIF-MAFs. Finally, we allowed the baseline MAF models to use batch normalization between MADE layers as recommended by Papamakarios et al. (2017), but we do not use them within CIF-MAFs as the structure of our F generalises this transformation.

We should note that our evaluation of models is slightly different from Papamakarios et al. (2017). For the model which scores best on the validation set, Papamakarios et al. (2017) report the average and standard deviation of log-likelihood across the points in the test dataset. However, our error bars emerge as the error in average test-set log-likelihood across *multiple* runs of the same experiment; this style of evaluation is often employed in other works as well (e.g. FFJORD (Grathwohl et al., 2019), NAF (Huang et al., 2018), and SOS (Jaini et al., 2019) as noted in Durkan et al. (2019, Table 1)).

C.1.3 NEURAL SPLINE FLOWS

The experiment comparing NSF baselines to CIF-NSFs mirrors the experimental setup in Durkan et al. (2019). Specifically, we constructed baseline NSFs that exactly copied the settings in Durkan et al. (2019, Table 5). We also built CIF-NSFs using these baseline settings, although for the CIF-NSF-1 model we lowered the number of hidden channels in the autoregressive networks so that the total number of trainable parameters matched that of the baseline. Our parameter settings are provided in Table C.7; note that parameter settings are homogeneous across datasets, besides MINIBOONE for which we reduced the size $\texttt{NN}_p$ and $\texttt{NN}_F$ by a factor of 4 as per Durkan et al. (2019)[17]. We trained both NSFs and CIF-NSFs for a number of training epochs corresponding to the number of training steps divided by the number of batches in the training set, i.e.

$$n_e = \lceil n_s / \left(n_t / n_b\right) \rceil ,$$

where n_e is the number of epochs, n_s is the number of training steps, n_b is the batch size, and n_t is the number of training data points. Note that n_s and n_b are from Durkan et al. (2019, Table 5), and n_t is fixed by the pre-processing steps from Papamakarios et al. (2017). We then evaluated the test-set performance of each model after the pre-specified number of

[17] Indeed, there was no choice of n_h which would allow us to achieve the same number of parameters as the baseline for the models noted in row 2 of Table C.7.

Table C.8: Mean $\pm$ standard error of average test set log-likelihood (higher is better). Best performing runs are shown in bold. CIF-Id-1 had $s \equiv 0$ and $t = \mathrm{Id}$. CIF-Id-2 had $s \equiv 0$ and $t = \mathtt{NN}_F$. CIF-Id-3 had $(s, t) = \mathtt{NN}_F$.

	POWER	GAS	HEPMASS	MINIBOONE
CIF-ID-1 ($\mathtt{NN}_q = 10 \times 2$)	0.43 ± 0.01	10.92 ± 0.10	-17.06 ± 0.05	-11.26 ± 0.03
CIF-ID-1 ($\mathtt{NN}_q = 100 \times 4$)	0.42 ± 0.01	10.86 ± 0.16	-17.44 ± 0.09	-10.91 ± 0.04
CIF-ID-2 ($\mathtt{NN}_q = 10 \times 2$)	0.45 ± 0.01	10.43 ± 0.08	-17.63 ± 0.10	-11.13 ± 0.08
CIF-ID-2 ($\mathtt{NN}_q = 100 \times 4$)	0.47 ± 0.01	10.89 ± 0.18	-17.51 ± 0.09	-10.75 ± 0.07
CIF-ID-3 ($\mathtt{NN}_q = 10 \times 2$)	$\mathbf{0.50 \pm 0.01}$	$\mathbf{11.32 \pm 0.14}$	-17.08 ± 0.02	-10.45 ± 0.04
CIF-ID-3 ($\mathtt{NN}_q = 100 \times 4$)	$\mathbf{0.50 \pm 0.01}$	$\mathbf{11.58 \pm 0.12}$	$\mathbf{-16.68 \pm 0.07}$	$\mathbf{-10.01 \pm 0.04}$

epochs, averaging across three seeds, and put the results in Table 1. We again average randomness across seeds, rather than across points in the test set, as discussed in the previous section.

We quickly note here that we selected our parameters after trying a few settings on various UCI datasets. There were other settings which performed better for individual datasets that are not included here, as we would like the proposed configurations to be as homogeneous as possible. It appeared as though the NSF models were already fairly good at modelling the data, which allowed us to make $\mathtt{NN}_q$ much smaller while still achieving good inference.

We also should note that we wrapped our code around the NSF bijection code from `https://github.com/bayesiains/nsf`. We also disable weight decay in all of these experiments without observing any problems with convergence.

C.1.4 ABLATING f

We ran ablation experiments to gain some insight into the relative importance of f in (9). In particular, we considered a 10 layer model ($L = 10$) where at each layer U_ℓ had the same dimension as the data and $f = \mathrm{Id}$ was the identity. We refer to this model as CIF-Id.

We considered three parameterisations of CIF-Id. The first had $s \equiv 0$ and $t = \mathrm{Id}$, which from our choice (10) of $p_{U|Z}$ corresponds to stacking the following generative process:

$$\begin{aligned} Z &\sim P_Z \\ \epsilon &\sim \mathrm{Normal}(0, I_d) \\ X &\coloneqq Z - \mu^p(Z) - e^{\varsigma^p(Z)} \odot \epsilon. \end{aligned} \tag{C.1}$$

Observe this generalise ResFlows, since (5) can be realised by sending $\varsigma^p \to -\infty$ and having $\mu^p < 1$. Accordingly, we took $\mathtt{NN}_p$ to be a 4×128 MLP to match the size of the residual blocks used in our tabular ResFlow experiments.

The second CIF-Id parameterisation had $s \equiv 0$ and $t = \mathtt{NN}_F$, which amounts to replacing (C.1) with

$$X \coloneqq Z - t\left(\mu^p(Z) + e^{\varsigma^p(Z)} \odot \epsilon\right).$$

To align with the first CIF-Id, we took $\mathtt{NN}_F$ and $\mathtt{NN}_p$ to be 2×128 MLPs, and zeroed out the s output of $\mathtt{NN}_F$ to obtain $s \equiv 0$. The third parameterisation had $(s, t) = \mathtt{NN}_F$, which replaces (C.1) with

$$X \coloneqq \exp\left(-s\left(\mu^p(Z) + e^{\varsigma^p(Z)} \odot \epsilon\right)\right) \odot Z - t\left(\mu^p(Z) + e^{\varsigma^p(Z)} \odot \epsilon\right).$$

Again, we took $\mathtt{NN}_F$ and $\mathtt{NN}_p$ to be 2×128 MLPs in this case.

We ran all configurations with two different choices of $\mathtt{NN}_q$: a 2×10 MLP as in our tabular ResFlow experiments, as well as a 4×100 MLP. The results are given in Table C.8.[18] Observe that these models performed comparably or better than the 100-layer ResFlows, but worse than the CIF-ResFlows and CIF-MAFs in Table 1. As discussed in Section 4.1.1,

[18] Due to computational constraints we did not run these experiments on BSDS300.

we conjecture this occurs because a CIF-Id requires greater complexity from $p_{U|Z}$ to make up for its simple choice of f, which in turn makes inference harder and hence the ELBO (14) looser, resulting in a poorer model that is learned overall. Likewise, note that the best performance in all cases was obtained when $(s, t) = \texttt{NN}_F$. This provides some justification for the generality of our choice of (9), as opposed to simpler alternatives that omit s or t.

C.2 Image Experiments

In all our image experiments we applied the same uniform dequantisation scheme as Theis et al. (2016), after which we applied the logit transform of Dinh et al. (2017) with $\alpha = 10^{-5}$ for Fashion-MNIST and $\alpha = 0.05$ for CIFAR10.

C.2.1 ResFlow

For our baseline ResFlow experiments we used the same architecture as Chen et al. (2019). In particular, our convolutional residual blocks (denoted Conv-ResBlock) had the form

$$\text{LipSwish} \to 3 \times 3 \text{ Conv} \to \text{LipSwish} \to 1 \times 1 \text{ Conv} \to \text{LipSwish} \to 3 \times 3 \text{ Conv},$$

while our fully connected residual blocks (denoted FC-ResBlock) had the form

$$\text{LipSwish} \to \text{Linear} \to \text{LipSwish} \to \text{Linear},$$

with a residual connection from the input to the output in both cases. The overall architecture of the flow in all cases was:

$$\text{Image} \to \text{LogitTransform}(\alpha) \to k\times \text{ Conv-ResBlock} \to [\text{Squeeze} \to k\times \text{ Conv-ResBlock}] \times 2 \to 4\times \text{ FC-ResBlock},$$

where the Squeeze operation was as defined by Dinh et al. (2017). Like Chen et al. (2019), we used ActNorm layers (Kingma & Dhariwal, 2018) before and after each residual block.

Due to computational constraints, the models we considered were smaller than those used by Chen et al. (2019). In particular, our smaller ResFlow models used 128 hidden channels in their Conv-ResBlocks, 64 hidden channels in the linear layers of their FC-ResBlocks, and had $k = 4$. Our larger ResFlow models used 256 hidden channels in their Conv-ResBlocks, 128 hidden channels in the linear layers of their FC-ResBlocks, and had $k = 6$. In contrast, Chen et al. (2019) used 512 hidden channels in their Conv-ResBlocks, 128 hidden channels in their FC-ResBlocks, and had $k = 16$.

As described for our tabular experiments, we used the same estimation scheme as Chen et al. (2019). Additionally:

- We took $\kappa = 0.98$;
- We used the Neumann gradient series expression for the log Jacobian (Chen et al., 2019, (8)) and computed gradients in the forward pass (Chen et al., 2019, (9)) to reduce memory overhead;
- We used an adaptive rather than a fixed number of power iterations for spectral normalisation (Gouk et al., 2018), with a tolerance of 0.001;

For the CIF-ResFlows, we augmented the smaller baseline ResFlow by treating each composition of ActNorm $\to$ ResBlock, as well as the final ActNorm, as an instance of f in (9). Each $\texttt{NN}_F$, $\texttt{NN}_p$, and $\texttt{NN}_q$ was a ResNet (He et al., 2016a;b) consisting of 2 residual blocks with 32 hidden channels (denoted 2×32). We gave each U_ℓ the same shape as a single channel of Z_ℓ, and upsampled to the dimension of Z_ℓ by adding channels at the output of each $\texttt{NN}_F$. Note that we did not experiment with using the larger baseline ResFlow model as the basis for a CIF.

For all models we used a learning rate of 10^{-3} and a batch size of 64.

Figure C.3 through to Figure C.8 show samples synthesised from the ResFlow and CIF-ResFlow density models trained on MNIST and CIFAR-10.

C.2.2 RealNVP

For our RealNVP-based image experiments, we took the baseline to be a RealNVP with the same architecture used by Dinh et al. (2017) for their CIFAR-10 experiments. In particular, we used 10 affine coupling layers with the corresponding

alternating channelwise and checkerboard masks. Each coupling layer used a ResNet (He et al., 2016a;b) consisting of 8 residual blocks of 64 channels (denoted 8×64). We replicated the multi-scale architecture of Dinh et al. (2017), squeezing the channel dimension after the first 3 coupling layers, and splitting off half the dimensions after the first 6. This model had 5.94M parameters for Fashion-MNIST and 6.01M parameters for CIFAR-10.

For the CIF-RealNVP, we considered each affine coupling layer to be an instance of f in (9). When choosing the size of our networks, we sought to maintain roughly the same depth over which gradients were propagated as in the baseline. To this end, our coupling networks were 4×64 ResNets, each $\texttt{NN}_p$ and $\texttt{NN}_q$ were 2×64 ResNets, and each $\texttt{NN}_F$ was a 2×8 ResNet. We gave each U_ℓ the same shape as a single channel of Z_ℓ, and upsampled to the dimension of Z_ℓ by adding channels at the output of $\texttt{NN}_F$. Our model had 5.99M parameters for Fashion-MNIST and 6.07M parameters for CIFAR-10.

For completeness, we also trained a RealNVP model with coupler networks of size 4×64 to match our CIF-RealNVP configuration. This model had 2.99M parameters for Fashion-MNIST and 3.05M for CIFAR-10.

In all cases for these experiments we used a learning rate of 10^{-4} and a batch size of 100.

Figure C.9 through to Figure C.14 show samples synthesised from the RealNVP and CIF-RealNVP density models trained on Fashion-MNIST and CIFAR-10.

C.3 2-D Experiments

To gain intuition about our model, we ran experiments on some simple 2-D datasets. For the datasets in Figure 1, we used a 10-layer ResFlow, a 100-layer ResFlow, and 10-layer CIF-ResFlow. For the CIF-ResFlows we took $d_{\mathcal{U}} = 1$. Other architectural and training details were the same as for the tabular experiments described in Section 5.1 and Section C.1.1. The resulting average test set log-likelihoods for the top dataset were:

- -1.501 for the 10-layer ResFlow
- -1.419 for the 100-layer ResFlow
- -1.409 for the 10-layer CIF-ResFlow

The final average test set log likelihoods for the bottom dataset were:

- -2.357 for the 10-layer ResFlow
- -2.287 for the 100-layer ResFlow
- -2.275 for the 10-layer CIF-ResFlow

Note that in both cases the CIF-ResFlow slightly outperformed the 100-layer ResFlow.

We additionally ran several experiments comparing a baseline MAF against a CIF-MAF on the 2-D datasets shown in Figure C.15. The baseline MAFs had 20 autoregressive layers, while the CIF-MAFs had 5. The network used at each layer had 4 hidden layers of 50 hidden units (denoted 4×50). For the CIF-MAF, we took $d_{\mathcal{U}} = 1$, and used 2×10 MLPs for $\texttt{NN}_F$ and 4×50 MLPs for $\texttt{NN}_p$ and $\texttt{NN}_q$. In total the baseline MAF had 160160 parameters, while our model had 119910 parameters.

The results of these experiments are shown in Figure C.15. Observe that CIF-MAF consistently produces a more faithful representation of the target distribution than the baseline, and in all cases achieved higher average test set log probability. A failure mode of our approach is exhibited in the spiral dataset, where our model still lacks the power to fully capture the topology of the target. However, we did not find it difficult to improve on this: by increasing the size of $\texttt{NN}_p$ to 8×50 (and keeping all other parameters fixed), we were able to obtain the result shown in Figure C.16. This model had a total of 221910 parameters. We also tried a larger MAF model with autoregressive networks of size 8×50, (obtaining 364160 parameters total). This model diverged after approximately 160 epochs. The result after 150 epochs is shown in Figure C.16.

Figure C.3: Synthetic MNIST samples generated by the small baseline ResFlow model

Figure C.4: Synthetic MNIST samples generated by the large baseline ResFlow model

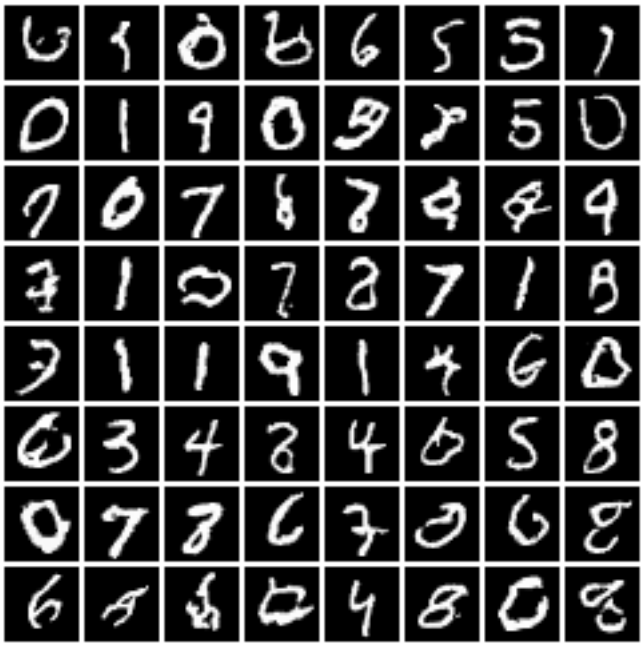


Figure C.5: Synthetic MNIST samples generated by the CIF-ResFlow model

Figure C.6: Synthetic CIFAR-10 samples generated by the small baseline ResFlow model

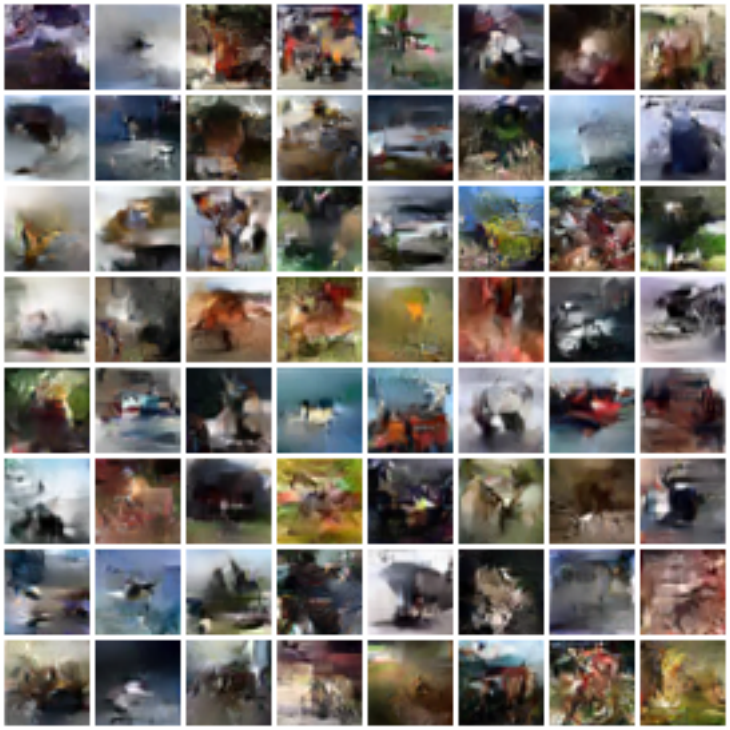


Figure C.7: Synthetic CIFAR-10 samples generated by the large baseline ResFlow model

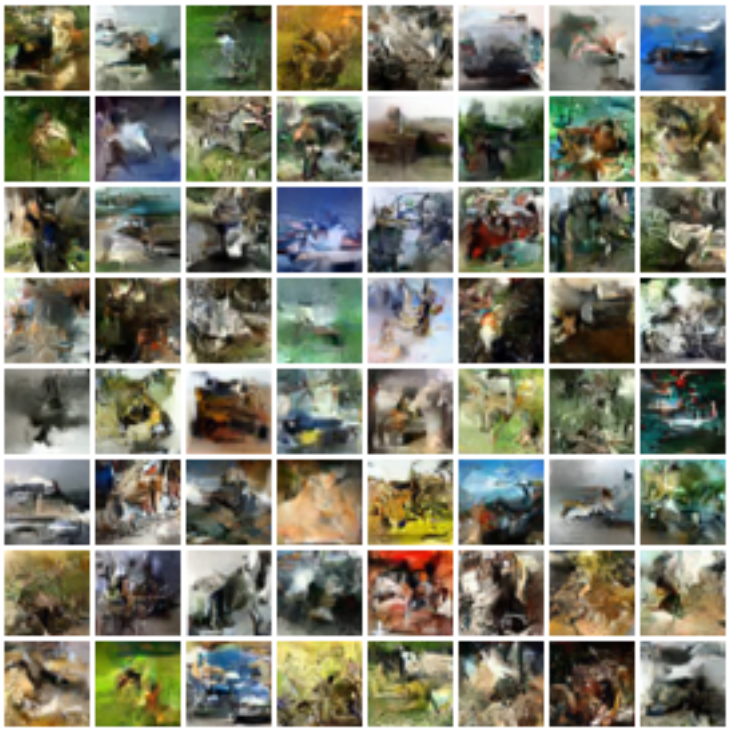


Figure C.8: Synthetic CIFAR-10 samples generated by the CIF-ResFlow model

Figure C.9: Synthetic Fashion-MNIST samples generated by RealNVP with coupling networks of size 4×64

Figure C.10: Synthetic Fashion-MNIST samples generated by RealNVP with coupling networks of size 8×64

Figure C.11: Synthetic Fashion-MNIST samples generated by CIF-RealNVP

Figure C.12: Synthetic CIFAR-10 samples generated by RealNVP with coupling networks of size 4×64

Figure C.13: Synthetic CIFAR-10 samples generated by RealNVP with coupling networks of size 8×64

Figure C.14: Synthetic CIFAR-10 samples generated by CIF-RealNVP

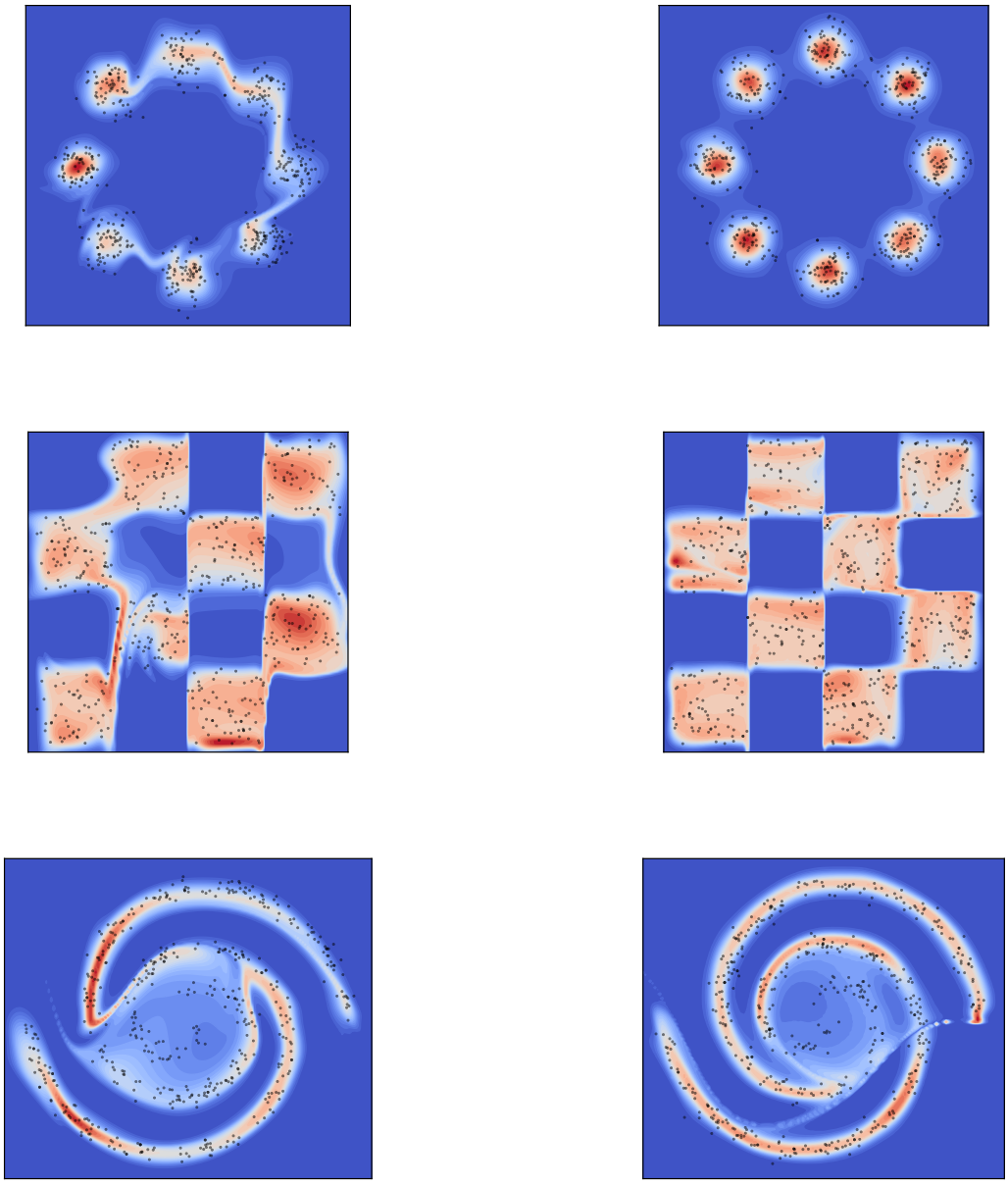


Figure C.15: Density models learned by a standard 20 layer MAF (left) and by a 5 layer CIF-MAF (right) for a variety of 2-D target distributions. Samples from the target are shown in black.

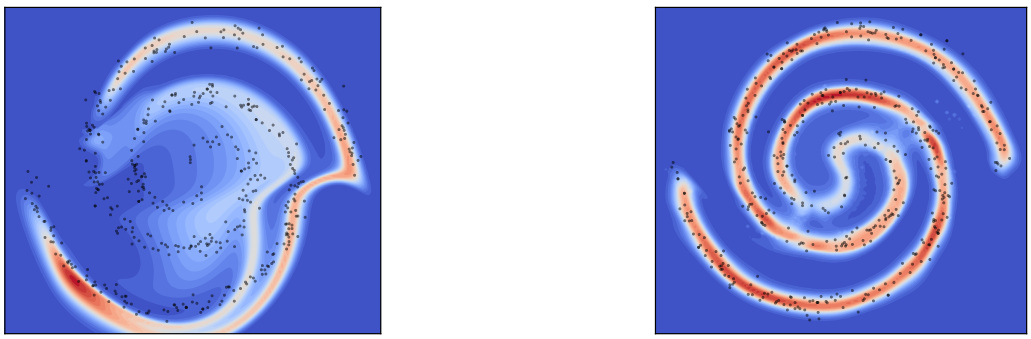


Figure C.16: Density models learned by a larger 20 layer MAF (left) and a larger 5 layer CIF-MAF (right) for the spirals dataset.

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	Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows
Publication Status	□Published ☑ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Accepted for publication at ICML 2020

Student Confirmation

Student Name:	Rob Cornish		
Contribution to the Paper	Identified topological restrictions with existing flow-based methods. Conceived of CIF architecture as a generalisation of normalising flows to address their identified limitations. Formulated and proved all theorems apart from those in Section B.4, which were proved by the second author (Anthony Caterini). Implemented CIFs in collaboration with Anthony Caterini and obtained experimental results. Wrote the majority of the initial paper draft.		
Signature 		Date	08/07/2020

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: Prof. Arnaud Doucet		
Supervisor comments: I do certify that the candidate made a substantial contribution to the publication, and the description above is accurate.		
Signature 	Date	08/07/2020

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

5 Conclusions

This thesis has presented three pieces of work aimed at addressing problems of scale that occur within the inference and modelling steps of the Bayesian workflow. In this final section we summarise these results and suggest directions for future work along these lines.

5.1 Contributions

In Chapter 2, we addressed the problem of performing inference via MCMC on Bayesian posteriors with very large datasets. We presented an asymptotically exact, MH-style algorithm that combines a factorised acceptance probability with a method for fast sampling of non-homogeneous Bernoulli processes to allow iterating without necessarily computing each likelihood term at each iteration. We used control variates to ensure a sublinear cost per iteration and nonvanishing acceptance probability in the limit of large data, providing rigorous theoretical guarantees to this effect.

In Chapter 3, we considered the implications of nesting Monte Carlo schemes, which occurs in contexts such as probabilistic programming for the purpose of modelling certain complex phenomena [Stuhlmüller and Goodman, 2014, Le et al., 2016b]. We give conditions under nested schemes are consistent, but showed

that in general nesting incurs an exponential cost in the number of layers in order to achieve a given performance. We discussed the consequences of our results for several applications.

In Chapter 4, we considered the problem of learning models automatically from data, which seems essential in order to scale Bayesian reasoning to large, complicated problems in artificial intelligence and elsewhere. We identified a limitation of existing normalising flow methods, which must preserve the topology of the support of the prior, hence limiting their expressiveness in practice. We proposed a modification to this framework involving stacked continuous mixtures of normalising flows that are trained variationally, and showed the empirical benefits of this approach on several problems.

5.2 Future work

5.2.1 Scalable Metropolis–Hastings

In many ways our results in Chapter 2 are negative, and suggest limitations with several existing lines of research into exact subsampling MCMC methods. Most notably, recent schemes involving nonreversible continuous-time piecewise deterministic Markov processes have attracted significant interest in part because they provide a mechanism for subsampling [Bouchard-Côté et al., 2018, Bierkens et al., 2019, Vanetti et al., 2017]. Our approach indicates that these same ideas can be applied in essentially a standard Metropolis–Hastings framework, and the potential benefits are not necessarily a consequence of being nonreversible or of operating in continuous-time.

In a more positive direction, it would be very useful in contexts such as probabilistic programming to allow the automatic application of SMH to a potentially arbitrary model. To do so using our current framework would require a means for programmatically deriving the bounds

$$\overline{U}_{k+1,i} \geq \sup_{\substack{\theta \in \Theta \\ |\beta|=k+1}} |\partial^{\beta} U_i(\theta)|,$$

which are required by our suggested factorisation. More ambitiously, for certain classes of models, it might also be possible to retain the theoretical properties of SMH using a different factorisation that achieves a tighter bound than $\overline{U}_{k+1,i}$, and hence better performance. To some extent the general applicability of our framework – we require only $\overline{U}_{k+1,i}$ and concentration of the target – is potentially also a shortcoming that reduces the practical benefits of SMH for real problems.

5.2.2 Nested Monte Carlo

Like Chapter 2, our results in Chapter 3 have negative implications. While we demonstrate that consistent estimation in the nested context is possible, the requirement that the number of samples to achieve a given accuracy must grow exponentially in the depth of nesting seems a significant limitation, and suggests a need to pursue alternative inference strategies to Monte Carlo for applications that require nesting. In particular, inference amortisation strategies [Kingma and Welling, 2013, Rezende et al., 2014, Gershman and Goodman, 2014, Paige and Wood, 2016, Le et al., 2016a] may be necessary to obtain good performance for large-scale nested models [Le et al., 2016b].

Subsequent work to ours has considered the implications of our results for existing probabilistic programming systems [Rainforth, 2018]. As an alternative direction, it also seems worth investigating conditions under which the bounds we provide are loose, so that good performance can be achieved without exponentially many samples in the depth of nesting. Examples certainly seem to exist. For instance, recall that the ELBO for CIFs consists of a nested expectation of the form

$$\mathcal{L}_\ell(z_\ell) = \mathbb{E}_{q_{U_\ell|Z_\ell}(u_\ell|z_\ell)}[\mathcal{L}_{\ell-1}(F_\ell^{-1}(z_\ell;u_\ell)) + \log p_{U_\ell|Z_{\ell-1}}(u_\ell|z_{\ell-1}) + \log|\det DF_\ell^{-1}(z_\ell;u_\ell)| - \log q_{U_\ell|Z_\ell}(u_\ell|z_\ell)], \quad (5.1)$$

which we estimate during training using Monte Carlo. While our results in Chapter 3 would seem to suggest that performance should degrade as ℓ grows larger, we found the opposite: deeper models performed better, even when using only a single Monte Carlo sample to approximate each expectation. It seems likely that this is due in

part to the fact that $q_{U_\ell|Z_\ell}$ is explicitly trained to approximate the posterior $p_{U_\ell|Z_\ell}$ and thereby to produce low variance estimates of the ELBO. We also speculate that, by increasing the depth of our model, the complexity required at each layer is reduced, which makes inference more tractable still and reduces the Monte Carlo error even futher. Establishing precise conditions that alleviate the difficulties of nesting could potentially make this behaviour more transparent and predictable.

5.2.3 Continuously Indexed Flows

Several directions seem possible for future work related to Chapter 4. It seems interesting to consider a continuous limit of CIFs along the lines of Chen et al. [2018], Grathwohl et al. [2018]. Tzen and Raginsky [2019] recently studied the continuous (stochastic) limit of deep latent Gaussian models [Rezende et al., 2014], which can be recovered as an instance of CIFs. It may be possible to extend their approach to apply more generally to our context also.

Alternatively, Nalisnick et al. [2018] made the intriguing observation that many classes of deep generative models assign higher probability density to out-of-sample data. For instance, they train a normalising flow model on FashionMNIST, and discover that MNIST has on average much higher log probability under the model. The artefacts we observed for our 2-D experiments – for example, the mass occuring between the two modes in Figure 1 – suggest that this may occur in part due to the topological restrictions that we identify. It would be interesting to consider whether CIFs can therefore yield better out-of-sample behaviour than standard normalising flows and related generative models.

Finally, our work on CIFs also highlighted to us the need for a systematic comparison study of different normalising flow architectures in a common training environment. Normalising flows have attracted significant recent interest and a flurry of architectures have emerged [Dinh et al., 2014, Rezende and Mohamed, 2015, Kingma et al., 2016, Papamakarios et al., 2017, Kingma and Dhariwal, 2018, Berg et al., 2018, Oliva et al., 2018, Huang et al., 2018, Chen et al., 2018, Grathwohl et al., 2018, Behrmann et al., 2018, Jaini et al., 2019, Ho et al., 2019,

Dinh et al., 2019, De Cao et al., 2019, Durkan et al., 2019, Chen et al., 2019]. Each architecture is reported to improve performance over a fairly standard set of benchmarks (UCI datasets, MNIST/FashionMNIST, CIFAR10). However, rarely are the baselines considered by these methods independently reproduced, with the results from earlier studies simply re-stated. As such, the improvements described may be in part the consequence of confounding factors such as data preprocessing methods or the training regime used.

We considered a variety of flows during the development of CIFs, and in several cases encountered difficulties in reproducing the results described in the original papers. For example, we found the *sum-of-squares polynomial flow* [Jaini et al., 2019] to be highly unstable to the extent that, for the UCI datasets, we did not achieve a non-NaN value for the average test-set log likelihood on any run (i.e. at least one test value was NaN). We believe this is due to the fact that the configuration suggested by Jaini et al. [2019] means their flow involves 9th order polynomials, which can produce extremely large values when given inputs with values outside the range $[-1, 1]^d$.

We likewise encountered problems with the *block neural autoregressive flow* [De Cao et al., 2019]. In particular, we note that the default choice of tanh nonlinearity suggested in the paper (and which we believe their results are based on) means that their model is not a bijection, since tanh is not surjective. The suggested alternative choice of LeakyReLU does fix this, but then introduces a second problem of removing the gradient signal on the Jacobian terms in the loss, which become constant almost everywhere. We sought to address this by considering a soft version of LeakyReLU defined by $x \mapsto \epsilon x + (1 - \epsilon) \log(1 + e^x)$, where $\epsilon \in (0, 1)$ corresponds to the slope on the negative part of the real line. However, while improving over LeakyReLU, this did not train successfully even for simple 2-D experiments.

Overall, since most flow-based architectures are justified principally on their ability to produce higher log probability scores than previous methods, it therefore seems of significant value to assess properly whether these improvements are in fact due to architectural innovations, or are instead the result of peripheral confounding factors.

Bibliography

Christophe Andrieu, Gareth O Roberts, et al. The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

Marco Banterle, Clara Grazian, Anthony Lee, and Christian P Robert. Accelerating metropolis-hastings algorithms by delayed acceptance. *arXiv preprint arXiv:1503.00996*, 2015.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1): 1515–1557, 2017.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. *arXiv preprint arXiv:1811.00995*, 2018.

Rianne van den Berg, Leonard Hasenclever, Jakub M Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*, 2018.

Joris Bierkens, Paul Fearnhead, Gareth Roberts, et al. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *The Annals of Statistics*, 47(3):1288–1320, 2019.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877, 2017.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Alexandre Bouchard-Côté, Sebastian J Vollmer, and Arnaud Doucet. The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

Ricky TQ Chen, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *arXiv preprint arXiv:1906.02735*, 2019.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

J. Andrés Christen and Colin Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005. ISSN 10618600. URL http://www.jstor.org/stable/27594150.

Nicola De Cao, Ivan Titov, and Wilker Aziz. Block neural autoregressive flow. *arXiv preprint arXiv:1904.04676*, 2019.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, Razvan Pascanu, and Hugo Larochelle. A rad approach to deep mixture models. *arXiv preprint arXiv:1903.07714*, 2019.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *arXiv preprint arXiv:1906.04032*, 2019.

SM Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410): 398–409, 1990.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Andrew Gelman et al. Induction and deduction in bayesian data analysis. 2011.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

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*, pages 2672–2680, 2014.

Noah Goodman, Vikash Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. *arXiv preprint arXiv:1206.3255*, 2012.

Will Grathwohl, Ricky TQ Chen, Jesse Betterncourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. *arXiv preprint arXiv:1902.00275*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*, 2018.

Pierre E Jacob, Alexandre H Thiery, et al. On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784, 2015.

Priyank Jaini, Kira A Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. *arXiv preprint arXiv:1905.02325*, 2019.

ET Jaynes. Probability theory: The logic of science. 2002.

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.

Michael I Jordan et al. On statistics, computation and scalability. *Bernoulli*, 19(4): 1378–1390, 2013.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

Durk 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*, pages 4743–4751, 2016.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Tuan Anh Le, Atilim Gunes Baydin, and Frank Wood. Inference compilation and universal probabilistic programming. *arXiv preprint arXiv:1610.09900*, 2016a.

Tuan Anh Le, Atılım Günes Baydin, and Frank Wood. Nested compiled inference for hierarchical reinforcement learning. In *NIPS Workshop on Bayesian Deep Learning*, 2016b.

Vikash Mansinghka, Daniel Selsam, and Yura Perov. Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*, 2014.

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.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017.

Junier B Oliva, Avinava Dubey, Manzil Zaheer, Barnabás Póczos, Ruslan Salakhutdinov, Eric P Xing, and Jeff Schneider. Transformation autoregressive networks. *arXiv preprint arXiv:1801.09819*, 2018.

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 preprint arXiv:1609.03499*, 2016a.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016b.

Brooks Paige and Frank Wood. Inference networks for sequential monte carlo in graphical models. In *International Conference on Machine Learning*, pages 3040–3049, 2016.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

Tom Rainforth. Nesting probabilistic programs. *arXiv preprint arXiv:1803.06328*, 2018.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Andreas Stuhlmüller and Noah D Goodman. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99, 2014.

Alexander Terenin and David Draper. Cox's theorem and the jaynesian interpretation of probability. *arXiv preprint arXiv:1507.06597*, 2015.

Dustin Tran, Alp Kucukelbir, Adji B Dieng, Maja Rudolph, Dawen Liang, and David M Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

Roberto Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104, 2008.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Paul Vanetti, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Piecewise-deterministic markov chain monte carlo. *arXiv preprint arXiv:1707.05296*, 2017.

Max Welling, Yee Whye Teh, Christophe Andrieu, Jakub Kominiarczuk, Ted Meeds, Babak Shahbaba, and Sebastian Vollmer. Bayesian inference with big data: a snapshot from a workshop. 2014.

Frank Wood, Jan Willem Meent, and Vikash Mansinghka. A new approach to probabilistic programming inference. In *Artificial Intelligence and Statistics*, pages 1024–1032, 2014.