Training Neural Networks for and by Interpolation

Leonard Berrada[1], Andrew Zisserman[1] and M. Pawan Kumar[1,2]
[1]Department of Engineering Science
University of Oxford
[2]Alan Turing Institute
{lberrada,az,pawan}@robots.ox.ac.uk

June 14, 2019

Abstract

The majority of modern deep learning models are able to interpolate the data: the empirical loss can be driven near zero on all samples simultaneously. In this work, we explicitly exploit this interpolation property for the design of a new optimization algorithm for deep learning. Specifically, we use it to compute an adaptive learning-rate given a stochastic gradient direction. This results in the Adaptive Learning-rates for Interpolation with Gradients (ALI-G) algorithm. ALI-G retains the advantages of SGD, which are low computational cost and provable convergence in the convex setting. But unlike SGD, the learning-rate of ALI-G can be computed inexpensively in closed-form and does not require a manual schedule. We provide a detailed analysis of ALI-G in the stochastic convex setting with explicit convergence rates. In order to obtain good empirical performance in deep learning, we extend the algorithm to use a maximal learning-rate, which gives a single hyper-parameter to tune. We show that employing such a maximal learning-rate has an intuitive proximal interpretation and preserves all convergence guarantees. We provide experiments on a variety of architectures and tasks: (i) learning a differentiable neural computer; (ii) training a wide residual network on the SVHN data set; (iii) training a Bi-LSTM on the SNLI data set; and (iv) training wide residual networks and densely connected networks on the CIFAR data sets. We empirically show that ALI-G outperforms adaptive gradient methods such as Adam, and provides comparable performance with SGD, although SGD benefits from manual learning rate schedules. We release PyTorch and Tensorflow implementations of ALI-G as standalone optimizers that can be used as a drop-in replacement in existing code (code available at `https://github.com/oval-group/ali-g`).

1 Introduction

Training a deep neural network is a challenging optimization problem: it involves minimizing the average of many high-dimensional non-convex functions. In practice, the main algorithms of choice are Stochastic Gradient Descent (SGD) [RM51] and adaptive gradient methods such as AdaGrad [DHS11] or Adam [KB15]. In recent work, the ability to interpolate – i.e. to achieve near zero loss on all training samples simultaneously – has been shown to help convergence of SGD [MBB18, VBS19, ZYZ+19]. This property is usually satisfied in supervised deep learning because of the empirical success of over-parameterized architectures. However, while the convergence analyses provide a better theoretical understanding of SGD, they do not help improve its practical behavior.

In this work, we open a different line of enquiry, namely: *can the interpolation property be used to design a robust and efficient optimization algorithm for deep learning?* In order to answer this question, we begin by giving the following two desiderata of an optimization algorithm for deep learning: (i) an inexpensive computational cost per iteration (typically a call to a stochastic first-order oracle); and (ii) adaptive learning-rates that do not require a manually designed schedule.

In this work, we present ALI-G (Adaptive Learning-rates for Interpolation with Gradients), an algorithm that takes advantage of interpolation by design and satisfies both properties mentioned above. Key to the ALI-G algorithm are the following two ideas. First, an adaptive learning-rate can be computed for the non-stochastic gradient direction when the minimum value of the objective function is known [Pol69,

Sho85, Brä95, NB01a, NB01b]. And second, one such minimum value is usually approximately known for interpolating models: for instance, it is close to zero for a model trained with the cross-entropy loss. By carefully combining these two ideas, we create a stochastic algorithm that provably converges fast in the convex setting and that can be used to train neural networks.

Procedurally, ALI-G can be seen as a variant of projected stochastic gradient descent with adaptive learning-rates. Crucially, since the adaptive learning-rates are computed in closed-form, ALI-G is completely hyper-parameter free in this form. In order to improve its empirical behavior with deep neural networks, we introduce a single hyper-parameter that controls the maximal learning-rate.

Contributions. We summarize the contributions of this work as follows:
- We formalize how the interpolation property can be exploited to compute a learning-rate in closed form at each iteration. The resulting algorithm has the same computational cost per iteration as SGD, but it requires no hyper-parameter for its learning-rate.
- We provide convergence rates of ALI-G in various convex settings.
- We extend the formulation to use a maximal learning-rate, which offers an intuitive interpretation through the lens of proximal optimization and retains all convergence guarantees.
- Empirically, we demonstrate the usefulness and efficiency of ALI-G on learning a differentiable neural computer; training variants of residual networks on the SVHN and CIFAR data sets; and training a Bi-LSTM on the Stanford Natural Language Inference data set.

This is a longer and more detailed version of the submitted paper. It provides additional theoretical results, and supplementary mathematical and visual interpretations.

2 Related Work

Interpolation in Deep Learning. As mentioned previously, recent works have successfully exploited the interpolation assumption to prove convergence of SGD in the context of deep learning [MBB18, VBS19, ZYZ+19]. Such works are complementary to ours in the sense that they provide a convergence analysis of an existing algorithm in the deep learning setting.

Adaptive Gradient Methods. Adaptive gradient methods are commonly used as alternatives to SGD to train deep neural networks. While the most popular ones are Adagrad [DHS11], RMSPROP [TH12], Adam [KB15] and AMSGrad [RKK18], there have been many other variants [Zei12, OP15, DB17, Lev17, MH17, ZK17, BWAA18, CG18, SS18, CLSH19, ZRS+18, LXLS19]. However, as pointed out in [WRS+17], adaptive gradient methods tend to give poor generalization in supervised learning. In our experiments, we compare ALI-G to the most popular adaptive gradient methods.

Incremental Subgradient Algorithms. Previous methods have exploited knowledge of the minimum value to compute an adaptive learning-rate, both in the deterministic case [Pol69, Sho85, Brä95], and in the randomized one [NB01a, NB01b]. However, even the randomized variants of these algorithms required periodical computations of the full (non-stochastic) objective function, which is expensive for large data sets. In contrast, by exploiting interpolation, ALI-G has the same computational cost per iteration as SGD. Furthermore, our extension with a maximal step-size is crucial for good practical performance with deep neural networks.

Adaptive Learning-Rate Algorithms. [RM18] proposed the L_4 algorithm that also employs an adaptive learning-rate based on the minimum value of the objective function. However, in contrast to our method, which utilizes interpolation, L_4 estimates the minimum value at each iteration based on past stochastic iterates. While being more broadly applicable, L_4 requires additional hyper-parameters to heuristically compensate for the noise. In addition, L_4 does not come with convergence guarantees. We compare the updates of L_4 and ALI-G in section 3.4, and we empirically evaluate the L_4 algorithms in our experiments. [BZK19] introduced the Deep Frank-Wolfe (DFW) algorithm to train deep neural networks by solving proximal linear support vector machine problems approximately. Though motivated differently, the DFW algorithm is procedurally close to ALI-G. However, DFW required the loss function to be piece-wise linear

convex, while ALI-G can use most loss functions, including the cross-entropy loss. In addition, ALI-G offers convergence guarantees for the convex setting, thereby making it more theoretically appealing. In a concurrent work to ours, [VML+19] show that one can use line search in a stochastic setting for interpolating models while guaranteeing convergence. However, in contrast to our work, the resulting algorithm requires more than one hyper-parameter (up to four), and the line-search is not computed in closed form. Note that we discuss their definition of interpolation and compare it to ours in section 3.3. Less closely related methods have proposed adaptive learning-rates without using the minimum for the computation of the learning rate [SZL13, TMDQ16, ZWW17, BCR+18, WWB18, LO19].

3 The ALI-G Algorithm

We begin by presenting the optimization problem that we aim to solve in this work. Then we present a simplified setting to give an intuition of the ALI-G algorithm. Finally, we introduce the ALI-G algorithm and briefly describe some convergence results for stochastic convex optimization.

3.1 Problem Setting

Loss Function. We consider a supervised learning task where the model is parameterized by $\mathbf{w} \in \mathbb{R}^p$. Usually, the objective function can be expressed as an expectation over $z \in \mathcal{Z}$, a random variable indexing the samples of the training set:

$$f(\mathbf{w}) \triangleq \mathbb{E}_{z \in \mathcal{Z}}[\ell_z(\mathbf{w})], \tag{1}$$

where each ℓ_z is the loss function associated with the sample z. We assume that each ℓ_z is lower-bounded, which is the case for the large majority of loss functions used in machine learning. For instance, suppose that the model is a deep neural network with weights $\mathbf{w}$ performing classification. Then for each sample z, $\ell_z(\mathbf{w})$ can represent the cross-entropy loss, which is always lower-bounded by zero. Other loss functions that are lower-bounded by zero include the structured or multi-class hinge loss, and the $\mathcal{L}_1$ or $\mathcal{L}_2$ loss functions for regression.

Regularization. It is sometimes desirable to employ a regularization function ϕ in order to promote generalization. In this work, we incorporate such regularization as a constraint on the feasible domain: $\Omega = \{\mathbf{w} \in \mathbb{R}^p : \ \phi(\mathbf{w}) \leq r\}$ for some value of r. In the deep learning setting, this will allow us to assume that the objective function can be driven close to zero without unrealistic assumptions about the regularization. Our framework can handle any constraint set Ω on which Euclidean projections are computationally efficient. This includes the feasible set induced by $\mathcal{L}_2$ regularization: $\Omega = \left\{\mathbf{w} \in \mathbb{R}^p : \ \|\mathbf{w}\|_2^2 \leq r\right\}$, for which the projection is given by a simple rescaling of $\mathbf{w}$. Finally, note that if we do not wish to use any regularization, we define $\Omega = \mathbb{R}^p$ and the corresponding projection is the identity.

Problem Formulation. The learning task can be expressed as the problem $(\mathcal{P})$ of finding a feasible vector of parameters $\mathbf{w}_\star \in \Omega$ that minimizes f:

$$\mathbf{w}_\star \in \operatorname*{argmin}_{\mathbf{w} \in \Omega} f(\mathbf{w}). \tag{$\mathcal{P}$}$$

Also note that $f_\star$ refers to the minimum value of f over Ω: $f_\star \triangleq \min_{\mathbf{w} \in \Omega} f(\mathbf{w})$.

3.2 A Simple Case: Non-Stochastic Update with Exact Minimum Known

Setting. In order to convey the intuition of our algorithm, we begin with a simple case. Specifically, we assume that $f_\star$ is known and we use non-stochastic updates: at each iteration, the full objective f and its derivative are evaluated. We now explain how this can be leveraged to inexpensively compute an adaptive learning-rate at each iteration. We denote by $\nabla f(\mathbf{w})$ the first-order derivative of f at $\mathbf{w}$ (e.g. $\nabla f(\mathbf{w})$ can be a sub-gradient or the gradient). In addition, we use the following notation: $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^p$, and $\Pi_\Omega(\mathbf{w})$ is the Euclidean projection of the vector $\mathbf{w} \in \mathbb{R}^p$ on the set Ω. Note that all proofs are deferred to Appendix B for space reasons.

Non-stochastic Update with $f_\star$ Known. At time-step t, we define the following update:

$$\mathbf{w}_{t+1} = \Pi_\Omega \left(\mathbf{w}_t - \gamma_t \nabla f(\mathbf{w}_t) \right), \text{ where } \gamma_t \triangleq \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2}, \tag{2}$$

where we loosely define $\frac{0}{0} = 0$ for simplicity purposes.

Interpretation. The update given by equation (2) is similar to projected gradient descent, with the crucial difference that the learning-rate γ_t is given in closed form by equation (2) instead of being a hyper-parameter of the method. It can be shown that $\mathbf{w}_{t+1}$ lies on the intersection between the linearization of f at $\mathbf{w}_t$ and the horizontal plane $z = f_\star$. Note that since $f_\star$ is the minimum of f, the learning-rate γ_t is necessarily non-negative.

The following proposition sheds further light on equation (2) in the unconstrained case:

Proposition 1. *Suppose that the problem is unconstrained: $\Omega = \mathbb{R}^p$. Let $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2} \nabla f(\mathbf{w}_t)$. Then $\mathbf{w}_{t+1}$ verifies:*

$$\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{w} - \mathbf{w}_t\| \text{ subject to: } f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t) = f_\star, \tag{3}$$

where we remind that $f_\star$ is the minimum of f, and $\mathbf{w} \mapsto f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t)$ is the linearization of f at $\mathbf{w}_t$.

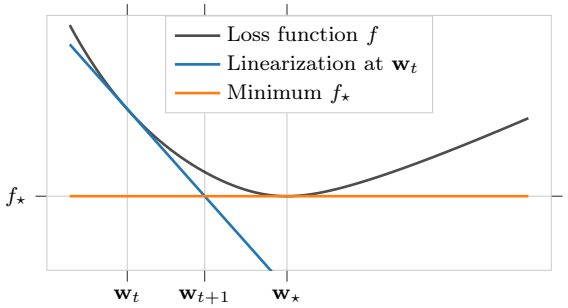

Figure 1: *Illustration of the update (2) in an unconstrained uni-dimensional case. In this setting, $\mathbf{w}_{t+1}$ is the intersection between the linearization of f at $\mathbf{w}_t$ and the minimum $f_\star$.*

In other words, $\mathbf{w}_{t+1}$ lies on the intersection between the linearization of f at $\mathbf{w}_t$ and the horizontal plane $z = f_\star$. Suppose that $\mathbf{w}_t$ is not already an optimal solution of problem $(\mathcal{P})$. Then in one dimension, this intersection is a single point, which thus defines $\mathbf{w}_{t+1}$ – see figure 1. In higher dimension, the intersection is a hyperplane of $\mathbb{R}^p$, and $\mathbf{w}_{t+1}$ is its member that is closest to $\mathbf{w}_t$.

Relationship to Existing Work. Consider the case $\Omega = \mathbb{R}$ and $f_\star = 0$, that is the unconstrained uni-dimensional setting where the minimum of f is 0. Then equation (2) coincides with an update of the Newton-Raphson method for finding roots.

In the more general case $\Omega \subseteq \mathbb{R}^p$ for some $p \geq 1$ and $f_\star \in \mathbb{R}$, the update defined in equation (2) has also been proposed in previous works [Pol69, Sho85, Brä95, NB01a, NB01b]. However, the simplified scenario presented in this section uses unrealistic assumptions and is not practically useful, as we explain below.

Limitations. Equation (2) has two major short-comings that prevent its applicability in a machine learning setting. First, the computation of γ_t requires exact knowledge of $f_\star$, the minimum of f over Ω. In practice, this exact knowledge is not usually available. Thus we wish to use instead an approximation that is known in advance. Second, the update requires a full evaluation of f and its derivative. Stochastic extensions have been proposed in [NB01a, NB01b], but they still require frequent evaluations of f. This is expensive in the large data setting, and even computationally infeasible when using massive data augmentation.

Therefore we would like to design an algorithm extending the update (2) so that (i) it requires an easily obtainable approximation of $f_\star$, and (ii) it relies only on stochastic estimates of f to compute the learning-rate γ_t. The next section introduces the notion of uniform lower bound, which will allow us to achieve these two goals.

3.3 Uniform Lower Bound & Interpolation

Formal Definitions. Intuitively, a uniform lower bound on the problem $(\mathcal{P})$ is a value that is lower than any value achievable by any single loss function ℓ_z on its unconstrained domain $\mathbb{R}^p$. We formalize this below:

Definition 1 (Uniform Lower Bound)**.** *We say that* $\underline{\mathrm{B}}$ *is a uniform lower bound on* $(\mathcal{P})$ *if:*

$$\underline{\mathrm{B}} \leq \inf_{z \in \mathcal{Z}} \inf_{\mathbf{w} \in \mathbb{R}^p} \ell_z(\mathbf{w}). \tag{4}$$

The definition above makes $\underline{\mathrm{B}}$ a useful statistic to analyze the behavior of each loss function ℓ_z around $\mathbf{w}_\star$, in a uniform way (that is, independently of z). The quality of a uniform lower bound $\underline{\mathrm{B}}$ can be quantified by the notion of ε-interpolation:

Definition 2 (ε-Interpolation)**.** *Let* $\underline{\mathrm{B}}$ *be a uniform lower bound on* $(\mathcal{P})$*,* $\mathbf{w}_\star$ *be a solution of* $(\mathcal{P})$ *and* $\varepsilon \geq 0$ *be a non-negative number. Then we say that* $\mathbf{w}_\star$ *is an* ε*-interpolation for* $((\mathcal{P}), \underline{\mathrm{B}})$ *if:*

$$\forall z \in \mathcal{Z},\ \ell_z(\mathbf{w}_\star) - \underline{\mathrm{B}} \leq \varepsilon. \tag{5}$$

Furthermore, when $\mathbf{w}_\star$ *is an* ε*-interpolation for* $((\mathcal{P}), \underline{\mathrm{B}})$ *with* $\varepsilon = 0$*, we say that* $\mathbf{w}_\star$ *is a perfect interpolation for* $((\mathcal{P}), \underline{\mathrm{B}})$*.*

By taking the expectation over equation (5), we can see that if $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, then we immediately have: $f_\star - \varepsilon \leq \underline{\mathrm{B}} \leq f_\star$. In other words, $\underline{\mathrm{B}}$ is also an approximation of $f_\star$ by below, and its quality is quantified by ε. We further note that $f_\star$ does not satisfy the definition of a uniform lower bound in the general case. However when $f_\star$ actually is a uniform lower bound, for any solution $\mathbf{w}_\star$ of $(\mathcal{P})$, $\mathbf{w}_\star$ is a perfect interpolation for $((\mathcal{P}), f_\star)$.

Applicability in Practice. We highlight that in many practical applications of deep learning, we already know a simple uniform lower bound for which the ε-interpolation assumption is verified. We detail this below.

Observation 1. *Suppose that we employ a non-negative loss function (such as cross-entropy, hinge loss, mean-squared-error etc.). Then* $\underline{\mathrm{B}} = 0$ *is a uniform lower bound on problem* $(\mathcal{P})$*. Furthermore, suppose that* $\mathbf{w}_\star$ *is a solution of* $(\mathcal{P})$ *such that* $\forall z \in \mathcal{Z}$*,* $\ell_z(\mathbf{w}_\star) \leq \varepsilon$*. Then* $\mathbf{w}_\star$ *is an* ε*-interpolation for* $((\mathcal{P}), \underline{\mathrm{B}} = 0)$*.*

To illustrate this further, consider the use case of a neural network performing classification, and suppose that the model is trained with the cross-entropy loss and $\mathcal{L}_2$ regularization. Let us further assume that the number of parameters of the neural network is significantly larger than the number of training samples, and that the data is consistently labeled. Due to the over-parameterization, the model is able to interpolate the data. In other words, it is possible to find parameter values $\mathbf{w}_\star$ such that the cross-entropy loss of each training sample is at most $\varepsilon \ll 1$. In such a case, $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}} = 0)$. Crucially, we have not made any assumption on the value of the $\mathcal{L}_2$ regularization, which can be significantly greater than 0 at $\mathbf{w}_\star$. This shows the advantage of including the regularization as a constraint rather than as an objective term: it avoids the unrealistic assumption of a regularization close to 0 at $\mathbf{w}_\star$.

On the Definition of Interpolation. Note that similarly to [MBB18], our definition of interpolation is based on function values, while the one of [VML+19] relies on gradient norms. These definitions are equivalent if the loss functions are convex, but they differ otherwise. In particular, they are different in the non-convex setting of optimization for deep learning. Our definition of interpolation, which uses function values, is more natural for our algorithm since our step-size exploits function values and function distances to the minimum.

3.4 The ALI-G Algorithm

Now that we have defined the notion of uniform lower bound, we assume that we have access to one such value (e.g. 0 for a neural network), and we denote it by $\underline{\mathrm{B}}$. In this section, we explain how the uniform lower bound can be leveraged to inexpensively compute an adaptive learning-rate at each iteration. In comparison with the update of equation (2), we will use $\underline{\mathrm{B}}$ instead of $f_\star$ to compute the learning-rate, and we will rely on stochastic estimates of f and its gradients only. In other words, each update will only use $(\underline{\mathrm{B}}, \ell_{z_t}(\mathbf{w}_t), \nabla \ell_{z_t}(\mathbf{w}_t))$ instead of $(f_\star, f(\mathbf{w}_t), \nabla f(\mathbf{w}_t))$.

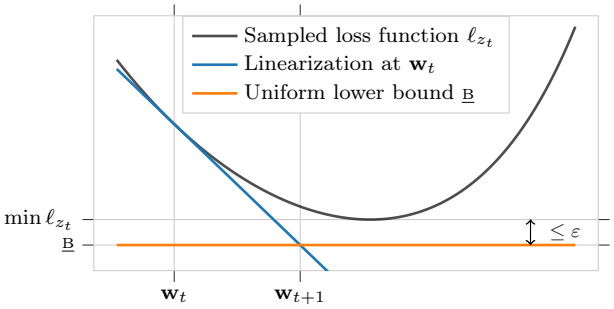

Figure 2: *Illustration of the ALI-G update in an unconstrained uni-dimensional case. Compared to figure 1, the learning-rate is computed for the sampled loss function ℓ_{z_t} instead of the full objective f, and uses a uniform lower bound $\underline{\mathrm{B}}$ rather than the exact minimum $f_\star$. If we assume ε-interpolation, then we have* $\min \ell_z - \varepsilon \leq \underline{\mathrm{B}} \leq \min \ell_z$.

Update. The update of ALI-G is illustrated in figure 2. At each time-step t, it can be written as:

$$\mathbf{w}_{t+1} = \Pi_\Omega \left(\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t) \right), \text{ where } \gamma_t \triangleq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}. \tag{6}$$

Note that the definition of the uniform lower bound $\underline{\mathrm{B}}$ guarantees that $\gamma_t \geq 0$. The term $\delta \geq 0$ is a positive constant for numerical stability, which we usually choose to have a very small positive value. This constant is required to be non-zero whenever $\underline{\mathrm{B}} < f_\star$ in order to avoid division by zero (while the numerator remains non-zero).

Comparison. In comparison with equation (2), we emphasize that we use $\underline{\mathrm{B}}$ instead of $f_\star$. We further note that given knowledge of $\underline{\mathrm{B}}$, the learning-rate given in equation (6) is inexpensive to compute. Indeed, it only requires $\ell_{z_t}(\mathbf{w}_t)$ and $\nabla \ell_{z_t}(\mathbf{w}_t)$, both of which are usually available in practice (in contrast to $f(\mathbf{w}_t)$ and $\nabla f(\mathbf{w}_t)$ as before). Furthermore, we point out that γ_t can again be interpreted as the solution of a proximal problem. Indeed, a result very similar to Proposition 1 can be shown in the case $\Omega = \mathbb{R}^p$ and $\delta = 0$, by simply replacing f by ℓ_z and $f_\star$ by $\underline{\mathrm{B}}$. The L_4 algorithms [RM18] use a similar update to (6). The crucial difference is that instead of the uniform lower bound $\underline{\mathrm{B}}$, they use an estimate of the minimum that is based on past iterations. This estimation requires three hyper-parameters in practice. In contrast, provided that $\underline{\mathrm{B}}$ is known in advance, ALI-G requires none.

4 The ALI-G Algorithm with a Maximal Learning-Rate

We wish to train deep neural networks with the ALI-G algorithm. However, the resulting steps are sometimes too large in practice, thus causing oscillations and preventing convergence. Indeed, the convergence guarantees of this work do not apply to the non-convex problem of training a deep neural network. Therefore we would like the algorithm to take more conservative steps in a controllable way. In this section, we show how truncating the learning-rate with a maximal value offers three advantages: (i) it is easy to interpret via a proximal formulation; (ii) it preserves the computational cost of the method; and (iii) it retains all guarantees of convergence. We denote by η the value of the maximal learning-rate. The resulting algorithm will still be

called ALI-G, and from now on, the version presented in the previous section will be referred to as ALI-G$^\infty$ (because it has an infinite maximal learning-rate).

Algorithm. Given a maximal learning-rate $\eta > 0$, ALI-G uses the following update at each iteration:

$$\mathbf{w}_{t+1} = \Pi_\Omega \left(\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t) \right), \text{ where } \gamma_t = \min \left\{ \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \eta \right\}. \tag{7}$$

Note that γ_t simply selects the smallest of the two elements $\frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ and η. In other words, the only difference between ALI-G$^\infty$ and ALI-G is that in the latter, each learning-rate γ_t is clipped to η.

The following two observations are direct consequences of equation (7):

Observation 2. *When we use the trivial uniform lower bound $\underline{\mathrm{B}} \to -\infty$, we always have $\gamma_t = \eta$ and we recover the projected SGD algorithm with constant learning-rate η.*

Theorem 1. *[Proximal Interpretation] Suppose that $\Omega = \mathbb{R}^p$ and let $\delta = 0$. We consider the update performed by SGD: $\mathbf{w}_{t+1}^{SGD} = \mathbf{w}_t - \eta \nabla \ell_{z_t}(\mathbf{w}_t)$; and the update performed by ALI-G: $\mathbf{w}_{t+1}^{ALI\text{-}G} = \mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)$, where $\gamma_t = \min \left\{ \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \eta \right\}$, Then we have:*

$$\mathbf{w}_{t+1}^{SGD} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^p} \left\{ \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_t\|^2 + \ell_{z_t}(\mathbf{w}_t) + \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t) \right\}, \tag{8}$$

$$\mathbf{w}_{t+1}^{ALI\text{-}G} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^p} \left\{ \frac{1}{2\eta} \|\mathbf{w} - \mathbf{w}_t\|^2 + \max \left\{ \ell_{z_t}(\mathbf{w}_t) + \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t), \underline{\mathrm{B}} \right\} \right\}. \tag{9}$$

In other words, at each iteration, ALI-G solves a proximal problem in closed form in a similar way to SGD. In both cases, the loss function ℓ_{z_t} is locally approximated by a first-order Taylor expansion at $\mathbf{w}_t$. The difference is that ALI-G also exploits the fact that ℓ_{z_t} is lower-bounded by $\underline{\mathrm{B}}$.

Relationship with Existing Algorithms. The formula of the learning-rate γ_t in equation (7) may remind the reader of existing algorithms, such as Frank-Wolfe [FW56] in some of its variants [LKTJ17], or other dual methods [LJJSP13, SSZ16]. This is because such methods are applicable to the dual of problem (9) which they happen to solve exactly in a single step. Therefore, if such methods were applied to problem (9), the resulting primal solution $\mathbf{w}_{t+1}$ would remain the same. We also point out that when no regularization is used, ALI-G and Deep Frank-Wolfe (DFW) [BZK19] are procedurally identical algorithms. This is because in such a setting, one iteration of DFW also amounts to solving (9) in closed-form. However, we point out the two fundamental advantages of ALI-G over DFW: (i) ALI-G can handle arbitrary (lower-bounded) loss functions, while DFW can only use convex piece-wise linear loss functions; and (ii) as will be seen shortly, ALI-G provides convergence guarantees in the convex setting.

Algorithm Overview. The ALI-G algorithm can be summarized as follows:

Algorithm 1 *The ALI-G algorithm*

Require: maximal learning-rate η, initial feasible $\mathbf{w}_0 \in \Omega$, uniform lower bound $\underline{\mathrm{B}}$, small constant $\delta > 0$

1: $t = 0$
2: **while** not converged **do**
3: Get $\ell_{z_t}(\mathbf{w}_t)$, $\nabla \ell_{z_t}(\mathbf{w}_t)$ with z_t drawn i.i.d.
4: $\gamma_t = \min \left\{ \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \eta \right\}$
5: $\mathbf{w}_{t+1} = \Pi_\Omega \left(\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t) \right)$
6: $t = t + 1$
7: **end while**

5 Convergence

We now establish various convergence rates of the ALI-G algorithm. A summary of the results is provided in Table 1:

Lipschitz	β-Smooth	α-Strongly Convex & β-Smooth
$\mathcal{O}(1/\sqrt{T})$	$\mathcal{O}(1/T)$	$\mathcal{O}(\exp(-\alpha T/8\beta))$

Table 1: Convergence results for ALI-G: distance to the minimum as a function of T, the number of iterations. The detailed results for ALI-G are described further, and their proofs are available in Appendix B.

Beyond the theoretical convergence speed, it is worth briefly discussing the computation of the learning-rate in practice. For the (S)GD algorithms, the choice of learning-rate depends on the properties of the objective function, such as the Lipschitz, strong convexity or smoothness constant. All of these are usually unknown to the user and are difficult to estimate. In contrast, provided that one has access to a uniform lower bound, the learning-rate of ALI-G only relies on information that is readily available to the user.

We now detail the convergence results for the ALI-G algorithm in three cases: Lipschitz convex functions; smooth convex functions; and smooth and strongly convex functions.

Theorem 2. *[Convex and Lipschitz] We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\underline{\mathrm{B}}$ be a uniform lower bound on $(\mathcal{P})$ and $\mathbf{w}_\star$ be a solution of $(\mathcal{P})$. Further suppose that $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$. Then ALI-G$^\infty$ applied to f satisfies:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \varepsilon\sqrt{\left(\frac{C^2}{\delta}+1\right)} + \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|\sqrt{C^2+\delta}}{\sqrt{T+1}}. \tag{10}$$

In other words, f approximately converges to $f_\star$ at a rate of $\mathcal{O}(1/\sqrt{T})$.

In addition, we note that the convergence rate of $\mathcal{O}(1/\sqrt{T})$ given by Theorem 2 is optimal among gradient-based algorithms ([Bub15], Theorem 3.13).

Theorem 3. *[Convex and Smooth] We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\underline{\mathrm{B}}$ be a uniform lower bound on $(\mathcal{P})$ and $\mathbf{w}_\star$ be a solution of $(\mathcal{P})$. Further suppose that $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, and that $\delta > 2\beta\varepsilon$. Then ALI-G$^\infty$ applied to f satisfies:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\delta}{\beta(1-\frac{2\beta\varepsilon}{\delta})} + \frac{2\beta}{1-\frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}. \tag{11}$$

In other words, f approximately converges to $f_\star$ at a rate of $\mathcal{O}(1/T)$.

Theorem 4. *[Strongly Convex and Smooth] We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\underline{\mathrm{B}}$ be a uniform lower bound on $(\mathcal{P})$ and $\mathbf{w}_\star$ be a solution of $(\mathcal{P})$. Further suppose that $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, and that $\delta > 2\beta\varepsilon$. Then ALI-G$^\infty$ applied to f satisfies:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \beta\exp\left(-\frac{\alpha T}{8\beta}\right)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{2\delta}{\alpha} + \left(10\frac{\beta}{\alpha} + 4\frac{\beta^2}{\alpha^2}\right)\varepsilon. \tag{12}$$

In other words, f approximately converges to $f_\star$ at a rate of $\mathcal{O}(\exp(-\alpha T/8\beta))$.

Note that in all previous results, we have given results with dependencies on ε and δ. If we happen to know $\varepsilon = 0$ (which happens when $\underline{\mathrm{B}} = f_\star$), we can safely use $\delta = 0$ and simplify all convergence results with the convention $\varepsilon/\delta = 0$; convergence then becomes exact rather than approximate.

It can be shown that ALI-G provably converges as fast as ALI-G$^\infty$ for η sufficiently large, and converges more slowly when η is small. For space reasons, the detailed results and their proofs are deferred to Appendices A and B.

6 Experiments

We compare ALI-G to the optimization algorithms most commonly used in deep learning. Our experiments span a variety of architectures and tasks: (i) learning a differentiable neural computer; (ii) training wide residual networks on SVHN; (iii) training a Bi-LSTM on the Stanford Natural Language Inference data set; and (iv) training wide residual networks and densely connected networks on the CIFAR data sets. In all these experiments, ALI-G will use the uniform lower bound of 0. Note that the tasks of training wide residual networks on SVHN and CIFAR-100 are part of the DeepOBS benchmark [SBH19], which aims at standardizing baselines for deep learning optimizers. In particular, these tasks are among the most difficult ones of the benchmark because the SGD baseline benefits from a manual schedule for the learning rate. Despite this, ALI-G obtains competitive performance with SGD. In addition, ALI-G is the best performing method with a single hyper-parameter on the difficult tasks of Bi-LSTM on SNLI and ResNet variants on CIFAR.

The code to reproduce our results is available at `https://github.com/oval-group/ali-g`. In the Tensorflow [AAB+15] experiment, we use the official and publicly available implementation of L4[1]. In the PyTorch [PGC+17] experiments, we use our implementation of L4, which we unit-test against the official Tensorflow implementation. In addition, we employ the official implementation of DFW[2] and we re-use their code for the experiments on SNLI and CIFAR. All experiments are performed either on a 12-core CPU (differentiable neural computer) or on a single GPU (SVHN, SNLI, CIFAR).

6.1 Differentiable Neural Computers

Setting. The Differentiable Neural Computer (DNC) [GWR+16] is a recurrent neural network that aims at performing computing tasks by learning from examples rather than by executing an explicit program. In this case, the DNC learns to repeatedly copy a fixed size string given as input. Although this learning task is relatively simple, the complex architecture of the DNC makes it an interesting benchmark problem for optimization algorithms.

Methods. We use the official and publicly available implementation of DNC[3]. We vary the initial learning rate as powers of ten between 10^{-4} and 10^4 for each method except for L4Adam and L4Mom. For L4Adam and L4Mom, since the main hyper-parameter α is designed to lie in $(0, 1)$, we vary it between 0.05 and 0.095 with a step of 0.1. The gradient norm is clipped for all methods except for ALI-G, L4Adam and L4Mom (as recommended by [RM18]).

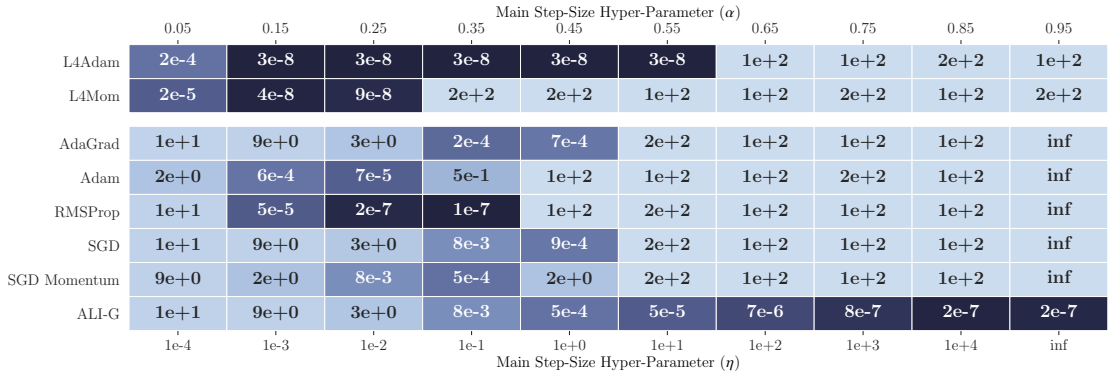

Figure 3: *Final objective function when training a Differentiable Neural Computer for* $10k$ *steps (lower is better). The intensity of each cell is log-proportional to the value of the objective function (darker is better). ALI-G obtains good performance for a very large range of* η *(any* $\eta \geq 0.1$*).*

[1] `https://github.com/martius-lab/l4-optimizer`
[2] `https://github.com/oval-group/dfw`
[3] `https://github.com/deepmind/dnc`

Results. We present the results in Figure 3. ALI-G provides accurate optimization for any $\eta \geq 0.1$, and is among the best performing methods by reaching an objective function of 2.10^{-7}. On this task, L4Adam and L4Mom also provide accurate and robust optimization. In contrast to ALI-G and the L4 methods, the most commonly used algorithms such as SGD, SGD with momentum and Adam are very sensitive to their main learning-rate hyper-parameter.

6.2 Wide Residual Networks on SVHN

Setting. The Street View House Numbers (SVHN) data set contains 73k training samples, 26k testing samples and 531k additional easier samples. From the 73k difficult training examples, we select 6k samples for validation; we use all remaining (both difficult and easy) examples for training, for a total of 598k samples. We train a wide residual network 16-4 following the protocol of [ZK16].

Method. For SGD, we use the manual schedule for the learning rate of [ZK16]. For L4Adam and L4Mom, we cross-validate the main learning-rate hyper-parameter α to be in $\{0.0015, 0.015, 0.15\}$ (0.15 is the value recommended by [RM18]). For other methods, the learning rate hyper-parameter is tuned as a power of 10. The $\mathcal{L}_2$ regularization is cross-validated in $\{0.0001, 0.0005\}$ for all methods but ALI-G. For ALI-G, the regularization is expressed as a constraint on the $\mathcal{L}_2$-norm of the (flattened) parameters, and its maximal value is set to 50. SGD, ALI-G and BPGrad use a Nesterov momentum of 0.9. All methods use a dropout rate of 0.4.

Results. The results are presented in Table 2.

Adagrad	Adam	AMSGrad	BPGrad	DFW	L4Adam	L4Mom	ALI-G	SGD	SGD†
98.0	97.9	97.9	98.1	98.1	**98.2**	19.6	98.1	98.3	98.4

Table 2: *Test Accuracy (%) on SVHN. In red, SGD benefits from a hand-designed schedule for its learning-rate. In black, adaptive methods, including ALI-G, have a single hyper-parameter for their learning-rate. SGD† refers to the performance reported by [ZK16].*

On this relatively easy task, most methods achieve about 98% test accuracy. Despite the cross-validation, L4Mom does not converge on this task. Even though SGD benefits from a hand-designed schedule, ALI-G and other adaptive methods obtain close performance to it.

6.3 Bi-LSTM on SNLI

Setting. We now train a Bi-LSTM of 47M parameters on the Stanford Natural Language Inference (SNLI) data set [BAPM15]. The SNLI data set consists in 570k pairs of sentences, with each pair labeled as entailment, neutral or contradiction. This large scale data set is commonly used as a pre-training corpus for transfer learning to many other natural language tasks where labeled data is scarcer [CKS$^+$17] – much like ImageNet is used for pre-training in computer vision. We follow the protocol of [BZK19]; we also use their code and re-use their existing results for the baselines.

Method. For L4Adam and L4Mom, the main hyper-parameter α is cross-validated in $\{0.015, 0.15\}$ – compared to the recommended value of 0.15, this helped convergence and considerably improved performance. The SGD algorithm benefits from a hand-designed schedule, where the learning-rate is decreased by 5 when the validation accuracy does not improve. Other methods use adaptive learning-rates and do not require such schedule. The value of the main hyper-parameter η is cross-validated as a power of ten for the ALI-G algorithm and for previously reported adaptive methods. Following the implementation by [CKS$^+$17], no $\mathcal{L}_2$ regularization is used. The algorithms are evaluated with the Cross-Entropy (CE) loss and the multi-class hinge loss (SVM), except for DFW which is designed for SVMs only.

Loss	Adagrad*	Adam*	AMSGrad*	BPGrad*	DFW*	L4Adam	L4Mom	ALI-G$^\infty$	ALI-G	SGD*	SGD†
CE	83.8	84.5	84.2	83.6	-	83.3	83.7	84.6	**84.8**	84.7	84.5
SVM	84.6	85.0	85.1	84.2	**85.2**	82.5	83.2	84.7	**85.2**	85.2	-

Table 3: *Test Accuracy (%) on SNLI. In red, SGD benefits from a hand-designed schedule for its learning-rate. In black, adaptive methods have a single hyper-parameter for their learning-rate. In blue, ALI-G$^\infty$ does not have any hyper-parameter for its learning-rate. With an SVM loss, DFW and ALI-G are procedurally identical algorithms – but in contrast to DFW, ALI-G can also employ the CE loss. Methods in the format X^* re-use results from [BZK19]. SGD† is the result from [CKS+17].*

Results. We present the results in Table 3. ALI-G$^\infty$ is the only method that requires no hyper-parameter for its learning-rate. Despite this, and the fact that SGD employs a learning-rate schedule that has been hand designed for good validation performance, ALI-G$^\infty$ is still able to obtain results that are competitive with SGD. Moreover, ALI-G, which requires a single hyper-parameter for the learning-rate, outperforms all other methods for both the SVM and the CE loss functions.

6.4 Wide Residual Networks and Densely Connected Networks on CIFAR

Setting. We follow the methodology of [BZK19]; we also re-use their code and we reproduce their results. We test two architectures: a Wide Residual Network (WRN) 40-4 [ZK16] and a bottleneck DenseNet (DN) 40-40 [HLWvdM17]. We use 45k samples for training and 5k for validation. The images are centered and normalized per channel. We apply standard data augmentation with random horizontal flipping and random crops.

Method. All optimization methods employ the cross-entropy loss, except for the DFW algorithm, which is designed to use an SVM loss. For DN and WRN respectively, SGD uses the manual learning rate schedules from [HLWvdM17] and [ZK16]. Following [BZK19], the batch-size is cross-validated in $\{64, 128, 256\}$ for the DN architecture, and $\{128, 256, 512\}$ for the WRN architecture. For L4Adam and L4Mom, the learning-rate hyper-parameter α is cross-validated in $\{0.015, 0.15\}$. For AMSGrad, DFW and ALI-G, the learning-rate hyper-parameter η is cross-validated as a power of 10 (in practice $\eta \in \{0.1, 1\}$ for ALI-G). SGD, DFW and ALI-G use a Nesterov momentum of 0.9. Following [BZK19], for all methods but ALI-G, the $\mathcal{L}_2$ regularization is cross-validated in $\{0.0001, 0.0005\}$ on the WRN architecture, and is set to 0.0001 for the DN architecture. For ALI-G, $\mathcal{L}_2$ regularization is expressed as a constraint on the norm on the vector of parameters; its maximal value is set to 100 for the WRN models, and to 75 for the DN ones. We present fewer baselines on this task due to the heavy computational cost of the cross-validation. AMSGrad was selected in [BZK19] because it was the best adaptive method on similar tasks, outperforming in particular Adam and Adagrad.

Results. We present the results in Table 4. In this setting again, ALI-G obtains competitive performance with SGD. Furthermore, ALI-G largely outperforms AMSGrad, and significantly bridges the gap between DFW and SGD on CIFAR-10 with the WRN model, and on CIFAR-100 with the DN one. Note that ALI-G is able to accurately minimize the objective function: the objective function reaches final values of about 0.001 on CIFAR-100, and 0.0001 on CIFAR-10.

Data Set	Architecture	AMSGrad	DFW	L4Adam	L4Mom	ALI-G	SGD	SGD†
CIFAR-10	WRN	90.8	94.2	90.5	91.6	**95.2**	95.3	95.4
	DN	91.7	94.6	90.8	91.9	**94.8**	95.1	-
CIFAR-100	WRN	68.7	**76.0**	61.7	61.4	75.8	77.8	78.8
	DN	69.4	73.2	60.5	62.6	**76.3**	76.3	-

Table 4: *Test Accuracy (%) on the CIFAR data sets. In red, SGD benefits from a hand-designed schedule for its learning-rate. In black, adaptive methods, including ALI-G, have a single hyper-parameter for their learning-rate. SGD† refers to the result from [ZK16]. Each reported result is an average over three independent runs.*

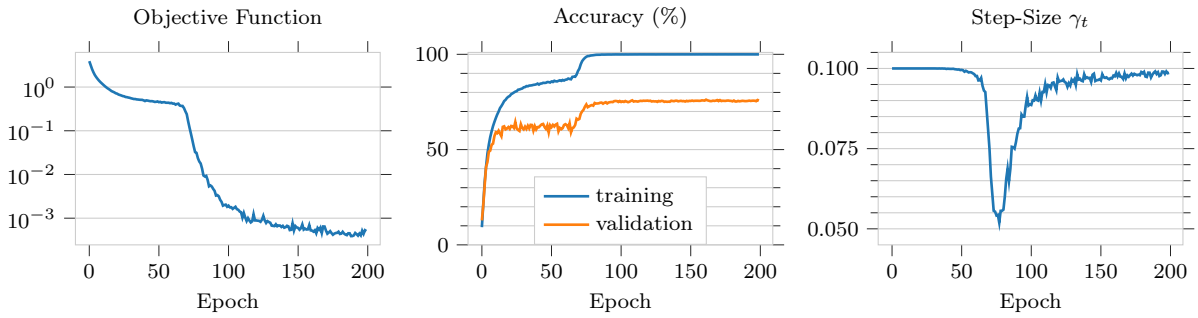

Figure 4: *Example of training with the ALI-G algorithm: a WRN model is trained on the CIFAR-100 data set. Each epoch takes about 50s, which is approximately the same as for SGD. One can observe that the objective function is accurately minimized, and reaches a final value of about 5e-4. In addition, one can notice the maximal value of $\eta = 0.1$ for the learning-rate γ_t. The sharp decrease in γ_t provokes the major improvement on both the objective function and on the accuracy around epoch 35. Perhaps surprisingly, the learning-rate increases again after this. We postulate that once the network has found parameter values that classify correctly each sample, the objective function is minimized by pushing the samples as far as possible from the decision boundary and ALI-G takes large steps for that.*

7 Discussion

We hope that the ALI-G algorithm is a helpful step towards efficient and reliable training of deep neural networks. ALI-G is readily applicable to a broad range of applications in deep learning where the model can interpolate the data. When that is not the case however, it would be interesting to design new algorithms that adapt the lower bound $\underline{B}$ online while requiring few hyper-parameters. This could be achieved for instance by building upon the works of [NB01b] and [RM18].

There are currently many parameters (too large initial learning rate, use of momentum, small batch-sizes etc.) that indirectly improve generalization while hurting optimization performance. We believe that a crucial next step is the design of more effective regularization methods removing such flukes, so that better optimization results in better learning.

Acknowledgments

This work was supported by the EPSRC grants AIMS CDT EP/L015987/1, Seebibyte EP/M013774/1, EP/P020658/1 and TU/B/000048, and by Yougov. We also thank the Nvidia Corporation for the GPU donation.

References

[AAB+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, GregS. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelioné Man, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernandaégas Vi, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015. Software available from tensorflow.org.

[BAPM15] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *Conference on Empirical Methods in Natural Language Processing*, 2015.

[BCR+18] Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. *International Conference on Learning Representations*, 2018.

[Brä95] Ulf Brännlund. A generalized subgradient method with relaxation step. *Mathematical Programming*, 1995.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

[BWAA18] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: Compressed optimisation for non-convex problems. *International Conference on Machine Learning*, 2018.

[BZK19] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Deep Frank-Wolfe for neural network optimization. *International Conference on Learning Representations*, 2019.

[CG18] Jinghui Chen and Quanquan Gu. Padam: Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint*, 2018.

[CKS+17] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *Conference on Empirical Methods in Natural Language Processing*, 2017.

[CLSH19] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *International Conference on Learning Representations*, 2019.

[DB17] Alexandre Défossez and Francis Bach. Adabatch: Efficient gradient aggregation rules for sequential and parallel stochastic gradient methods. *arXiv preprint*, 2017.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 1956.

[GWR+16] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016.

[HLWvdM17] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *Conference on Computer Vision and Pattern Recognition*, 2017.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[Lev17] Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. *Neural Information Processing Systems*, 2017.

[LJJSP13] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. *International Conference on Machine Learning*, 2013.

[LKTJ17] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. *International Conference on Artificial Intelligence and Statistics*, 2017.

[LO19] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. *International Conference on Artificial Intelligence and Statistics*, 2019.

[LXLS19] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *International Conference on Learning Representations*, 2019.

[MBB18] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *International Conference on Machine Learning*, 2018.

[MH17] Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. *International Conference on Machine Learning*, 2017.

[NB01a] Angelia Nedić and Dimitri Bertsekas. Convergence rate of incremental subgradient algorithms. *Stochastic optimization: algorithms and applications*, 2001.

[NB01b] Angelia Nedić and Dimitri Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 2001.

[OP15] Francesco Orabona and Dávid Pál. Scale-free algorithms for online linear optimization. *International Conference on Algorithmic Learning Theory*, 2015.

[PGC+17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS Autodiff Workshop*, 2017.

[Pol69] Boris Teodorovich Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 1969.

[RKK18] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *International Conference on Learning Representations*, 2018.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.

[RM18] Michal Rolinek and Georg Martius. L4: Practical loss-based stepsize adaptation for deep learning. *Neural Information Processing Systems*, 2018.

[SBH19] Frank Schneider, Lukas Balles, and Philipp Hennig. DeepOBS: A deep learning optimizer benchmark suite. *International Conference on Learning Representations*, 2019.

[Sho85] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions.* Springer Series in Computational Mathematics, 1985.

[SS18] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. *International Conference on Machine Learning*, 2018.

[SSZ16] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 2016.

[SZL13] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. *International Conference on Machine Learning*, 2013.

[TH12] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 2012.

[TMDQ16] Conghui Tan, Shiqian Ma, Yu-Hong Dai, and Yuqiu Qian. Barzilai-borwein step size for stochastic gradient descent. *Neural Information Processing Systems*, 2016.

[VBS19] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *International Conference on Artificial Intelligence and Statistics*, 2019.

[VML+19] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *arXiv preprint*, 2019.

[WRS+17] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Neural Information Processing Systems*, 2017.

[WWB18] Xiaoxia Wu, Rachel Ward, and Léon Bottou. WNGrad: Learn the learning rate in gradient descent. *arXiv preprint*, 2018.

[Zei12] Matthew Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint*, 2012.

[ZK16] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *British Machine Vision Conference*, 2016.

[ZK17] Shuai Zheng and James T Kwok. Follow the moving leader in deep learning. *International Conference on Machine Learning*, 2017.

[ZRS+18] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Neural Information Processing Systems*, 2018.

[ZWW17] Ziming Zhang, Yuanwei Wu, and Guanghui Wang. Bpgrad: Towards global optimality in deep learning via branch and pruning. *Conference on Computer Vision and Pattern Recognition*, 2017.

[ZYZ+19] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *International Conference on Learning Representations*, 2019.

A Convergence Results for ALI-G with a Maximal Learning-Rate

In this section, we show that using a maximal learning-rate does not affect the convergence guarantees of the ALI-G algorithm (other than by changing the constants in the convergence results). As previously, we consider the following three cases: (i) convex and Lipschitz functions, (ii) convex and β-smooth functions and finally (iii) α-strongly convex and β-smooth functions.

A.1 Lipschitz Convex Functions

We give the two following results in the convex and Lipschitz case. We note that the first result only holds for a large value of η, while the second result is informative in the complementary case where η has a small value.

Theorem 5. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\mathbf{w}_\star$ be an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$. We further assume that $\eta > \frac{\varepsilon}{\delta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{(\eta - \frac{\varepsilon}{\delta})(T+1)} + \frac{\varepsilon^2}{\delta(\eta - \frac{\varepsilon}{\delta})} + \sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} + \varepsilon\sqrt{\frac{C^2}{\delta} + 1}. \tag{13}$$

Proof: See Appendix B.6. ■

We note that for very large values of η ($\eta \to \infty$), Theorem 5 gives the exact same result as Theorem 2. However when η is small, the convergence error of Theorem 5 is large. This is corrected in the following result which is informative in the regime where η is small:

Theorem 6. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\mathbf{w}_\star$ be an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \varepsilon + \sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} + \eta\varepsilon\sqrt{C^2+\delta}. \tag{14}$$

Proof: See Appendix B.7. ■

A.2 Smooth Convex Functions

We now tackle the convex and β-smooth case. Our proof techniques naturally produce two distinct cases: $\eta \geq \frac{1}{2\beta}$ and $\eta \leq \frac{1}{2\beta}$. When $\eta \geq \frac{1}{2\beta}$, the convergence result is exactly the same as in Theorem 3 (which corresponds to $\eta \to \infty$). When $\eta \leq \frac{1}{2\beta}$, the speed of convergence is limited by the value of η.

Theorem 7. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \geq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\delta}{\beta(1 - \frac{2\beta\varepsilon}{\delta})} + \frac{2\beta}{1 - \frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}. \tag{15}$$

Proof: See Appendix B.8. ■

Theorem 8. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \leq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \frac{\delta}{2\beta} + \varepsilon. \tag{16}$$

Proof: See Appendix B.9. ■

A.3 Smooth and Strongly Convex Functions

Finally, we consider the α-strongly convex and β-smooth case. Again, our proof yields a natural distinction between $\eta \geq \frac{1}{2\beta}$ and $\eta \leq \frac{1}{2\beta}$. In a similar way to the β-smooth case, when $\eta \geq \frac{1}{2\beta}$, Theorem 9 gives the exact same result as Theorem 4 (which corresponds to $\eta \to \infty$). And when $\eta \leq \frac{1}{2\beta}$, the rate of convergence given by Theorem 10 is limited by the value of η.

Theorem 9. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for ($\mathcal{P}$), B), and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \geq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \beta \exp\left(-\frac{\alpha T}{8\beta}\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{2\delta}{\alpha} + \left(10\frac{\beta}{\alpha} + 4\frac{\beta^2}{\alpha^2}\right)\varepsilon. \tag{17}$$

Proof: See Appendix B.10. ■

Theorem 10. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for ($\mathcal{P}$), B), and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \leq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \beta \exp\left(-\frac{\alpha\eta T}{4}\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{2\delta}{\alpha} + \frac{14\varepsilon\beta}{\alpha}. \tag{18}$$

Proof: See Appendix B.11. ■

B Proofs

B.1 Proposition 1

Proposition 1. *Suppose that the problem is unconstrained: $\Omega = \mathbb{R}^p$. Let $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2}\nabla f(\mathbf{w}_t)$. Then $\mathbf{w}_{t+1}$ verifies:*

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p} \|\mathbf{w} - \mathbf{w}_t\| \text{ subject to: } f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top(\mathbf{w} - \mathbf{w}_t) = f_\star, \tag{3}$$

where we remind that $f_\star$ is the minimum of f, and $\mathbf{w} \mapsto f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top(\mathbf{w} - \mathbf{w}_t)$ is the linearization of f at $\mathbf{w}_t$.

Proof: First we show that $\mathbf{w}_{t+1}$ satisfies the linear equality constraint:

$$\begin{aligned} &f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) \\ &\quad= f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top\left(-\frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2}\nabla f(\mathbf{w}_t)\right), \\ &\quad= f(\mathbf{w}_t) - f(\mathbf{w}_t) + f_\star, \\ &\quad= f_\star. \end{aligned} \tag{19}$$

Now let us show that it has a minimal distance to $\mathbf{w}_t$.

We take $\hat{\mathbf{w}} \in \mathbb{R}^p$ a solution of the linear equality constraint, and we will show that $\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \leq \|\hat{\mathbf{w}} - \mathbf{w}_t\|$. By definition, we have that $\hat{\mathbf{w}}$ satisfies:

$$f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top(\hat{\mathbf{w}} - \mathbf{w}_t) = f_\star. \tag{20}$$

Now we can write:

$$\begin{aligned}
\|\mathbf{w}_{t+1}-\mathbf{w}_t\| &= \|\frac{f(\mathbf{w}_t)-f_\star}{\|\nabla f(\mathbf{w}_t)\|^2}\nabla f(\mathbf{w}_t)\|,\\
&= \frac{f(\mathbf{w}_t)-f_\star}{\|\nabla f(\mathbf{w}_t)\|},\\
&= \frac{|\nabla f(\mathbf{w}_t)^\top(\hat{\mathbf{w}}-\mathbf{w}_t)|}{\|\nabla f(\mathbf{w}_t)\|}, \qquad (21)\\
&\leq \frac{\|\nabla f(\mathbf{w}_t)\|\|\hat{\mathbf{w}}-\mathbf{w}_t\|}{\|\nabla f(\mathbf{w}_t)\|}, \quad \text{(Cauchy-Schwarz)}\\
&= \|\hat{\mathbf{w}}-\mathbf{w}_t\|.
\end{aligned}$$

■

B.2 Theorem 1

Theorem 1. *[Proximal Interpretation] Suppose that $\Omega=\mathbb{R}^p$ and let $\delta=0$. We consider the update performed by SGD: $\mathbf{w}_{t+1}^{SGD}=\mathbf{w}_t-\eta\nabla\ell_{z_t}(\mathbf{w}_t)$; and the update performed by ALI-G: $\mathbf{w}_{t+1}^{ALI\text{-}G}=\mathbf{w}_t-\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)$, where $\gamma_t=\min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta},\eta\right\}$, Then we have:*

$$\mathbf{w}_{t+1}^{SGD}=\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\ell_{z_t}(\mathbf{w}_t)+\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}-\mathbf{w}_t)\right\}, \qquad (8)$$

$$\mathbf{w}_{t+1}^{ALI\text{-}G}=\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\max\left\{\ell_{z_t}(\mathbf{w}_t)+\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}-\mathbf{w}_t),\underline{\mathrm{B}}\right\}\right\}. \qquad (9)$$

Proof: In order to make the notation simpler, we use $\boldsymbol{d}_t\triangleq\nabla\ell_{z_t}(\mathbf{w}_t)$ and $l_t\triangleq\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}$.
First, let us consider $\boldsymbol{d}_t=\mathbf{0}$.

Then we choose $\gamma_t=0$ and it is clear that $\mathbf{w}_{t+1}=\mathbf{w}_t-\eta\gamma_t\boldsymbol{d}_t=\mathbf{w}_t$ is the optimal solution of problem (9).

We now assume $\boldsymbol{d}_t\neq\mathbf{0}$.

We can successively re-write the proximal problem (9) as :

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\max\left\{\ell_{z_t}(\mathbf{w}_t)+\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}-\mathbf{w}_t),\underline{\mathrm{B}}\right\}\right\},\\
&\min_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\max\left\{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}+\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}-\mathbf{w}_t),0\right\}\right\},\\
&\min_{\mathbf{w}\in\mathbb{R}^p}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\max\left\{l_t+\boldsymbol{d}_t^\top(\mathbf{w}-\mathbf{w}_t),0\right\}\right\},\\
&\min_{\mathbf{w}\in\mathbb{R}^p,\upsilon}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\upsilon\right\}\text{ subject to: }\ \upsilon\geq0,\ \upsilon\geq l_t+\boldsymbol{d}_t^\top(\mathbf{w}-\mathbf{w}_t)\\
&\min_{\mathbf{w}\in\mathbb{R}^p,\upsilon}\sup_{\mu,\nu\geq0}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\upsilon-\mu\upsilon-\nu(\upsilon-l_t-\boldsymbol{d}_t^\top(\mathbf{w}-\mathbf{w}_t))\right\}\\
&\sup_{\mu,\nu\geq0}\min_{\mathbf{w}\in\mathbb{R}^p,\upsilon}\left\{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_t\|^2+\upsilon-\mu\upsilon-\nu(\upsilon-l_t-\boldsymbol{d}_t^\top(\mathbf{w}-\mathbf{w}_t))\right\}\quad\text{(strong duality)} \qquad (22)
\end{aligned}$$

The inner problem is now smooth in $\mathbf{w}$ and υ. We write its KKT conditions:

$$\frac{\partial\cdot}{\partial\upsilon}=0:\quad 1-\mu-\nu=0 \qquad (23)$$

$$\frac{\partial\cdot}{\partial\mathbf{w}}=0:\quad \frac{1}{\eta}(\mathbf{w}-\mathbf{w}_t)+\nu\boldsymbol{d}_t=\mathbf{0} \qquad (24)$$

We plug in these results and obtain:

$$\begin{aligned}
&\sup_{\mu,\nu\geq 0}\left\{\frac{1}{2\eta}\|\eta\nu\boldsymbol{d}_t\|^2+\nu(l_t+\boldsymbol{d}_t^\top(-\eta\nu\boldsymbol{d}_t))\right\}\\
&\text{st:}\quad \mu+\nu=1\\
&\sup_{\nu\in[0,1]}\left\{\frac{\eta}{2}\nu^2\|\boldsymbol{d}_t\|^2+\nu l_t-\eta\nu^2\|\boldsymbol{d}_t^\top\|^2\right\}\\
&\sup_{\nu\in[0,1]}\left\{-\frac{\eta}{2}\nu^2\|\boldsymbol{d}_t\|^2+\nu l_t\right\}
\end{aligned}\tag{25}$$

This is a one-dimensional quadratic problem in ν. It can be solved in closed-form by finding the global maximum of the quadratic objective, and projecting the solution on $[0,1]$. We have:

$$\frac{\partial\cdot}{\partial\nu}=0:-\eta\nu\|\boldsymbol{d}_t\|^2+l_t=0\tag{26}$$

Since $\boldsymbol{d}_t\neq\mathbf{0}$ and $\eta\neq 0$, this gives the optimal solution:

$$\nu=\min\left\{\max\left\{\frac{l_t}{\eta\|\boldsymbol{d}_t\|^2},0\right\},1\right\}=\min\left\{\frac{l_t}{\eta\|\boldsymbol{d}_t\|^2},1\right\},\tag{27}$$

since $l_t,\eta,\|\boldsymbol{d}_t\|^2\geq 0$.
Plugging this back in the KKT conditions, we obtain that the solution $\mathbf{w}_{t+1}$ of the primal problem can be written as:

$$\begin{aligned}
\mathbf{w}_{t+1}&=\mathbf{w}_t-\eta\nu\boldsymbol{d}_t,\\
&=\mathbf{w}_t-\eta\min\left\{\frac{l_t}{\eta\|\boldsymbol{d}_t\|^2},1\right\}\boldsymbol{d}_t,\\
&=\mathbf{w}_t-\eta\min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\eta\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2},1\right\}\nabla\ell_{z_t}(\mathbf{w}_t),\\
&=\mathbf{w}_t-\min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2},\eta\right\}\nabla\ell_{z_t}(\mathbf{w}_t).
\end{aligned}\tag{28}$$

■

B.3 Theorem 2

Theorem 2. *[Convex and Lipschitz] We assume that $\mathcal{X}$ is a convex set, and that for every $z\in\mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\underline{\mathrm{B}}$ be a uniform lower bound on $(\mathcal{P})$ and $\mathbf{w}_\star$ be a solution of $(\mathcal{P})$. Further suppose that $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}),\underline{\mathrm{B}})$. Then ALI-G$^\infty$ applied to f satisfies:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star\leq\varepsilon\sqrt{\left(\frac{C^2}{\delta}+1\right)}+\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|\sqrt{C^2+\delta}}{\sqrt{T+1}}.\tag{10}$$

In other words, f approximately converges to $f_\star$ at a rate of $\mathcal{O}(1/\sqrt{T})$.

Proof:

We consider the update at time t, which we condition on the draw of $z_t\in\mathcal{Z}$:

$$\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\\
&=\|\Pi_\Omega(\mathbf{w}_t-\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t))-\mathbf{w}_\star\|^2\\
&\leq\|\mathbf{w}_t-\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)-\mathbf{w}_\star\|^2\quad(\Pi_\Omega\text{ projection})\\
&=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t-\mathbf{w}_\star)+\gamma_t^2\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2\\
&=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t-\mathbf{w}_\star)+\gamma_t\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2\\
&\qquad(\text{definition of }\gamma_t)
\end{aligned}$$

$$
\begin{aligned}
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2} \|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \\
&\quad \text{(because } \ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}} \geq 0 \text{ and } \delta \geq 0) \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t (\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}) \quad \text{(convexity of } \ell_{z_t}) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t \Big((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - (\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \Big) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \Big((\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}})(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&\quad - (\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}})(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \Big) \quad \text{(definition of } \gamma_t) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \Big((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&\quad + (\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \\
&\quad - (\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \Big) \\
&\quad \text{(we use twice } \ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}} = \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \Big((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 - (\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})^2 \Big) \\
&\quad \text{(middle terms cancel out)} \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} + \frac{(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})^2}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})^2}{\delta} \quad (0 \leq \|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \leq C) \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\varepsilon^2}{\delta} \quad \text{(definition of } \varepsilon) \qquad (29)
\end{aligned}
$$

We re-write this inequality as:

$$
(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 \leq \varepsilon^2 \left(\frac{C^2}{\delta} + 1 \right) + \left(C^2 + \delta \right) \left(\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \right) \qquad (30)
$$

We can now use the Cauchy-Schwarz inequality to bound the sum over the iterations:

$$
(T+1) \sum_{t=0}^{T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 \geq \left(\sum_{t=0}^{T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \right)^2 \qquad (31)
$$

Therefore we can write:

$$
\begin{aligned}
&\left(\sum_{t=0}^{T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \right)^2 \\
&\leq (T+1) \sum_{t=0}^{T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 \\
&\leq (T+1)^2 \varepsilon^2 \left(\frac{C^2}{\delta} + 1 \right) + (T+1) \left(C^2 + \delta \right) \left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \right), \\
&\leq (T+1)^2 \varepsilon^2 \left(\frac{C^2}{\delta} + 1 \right) + (T+1) \left(C^2 + \delta \right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2,
\end{aligned} \qquad (32)
$$

which yields:

$$
\begin{aligned}
\sum_{t=0}^{T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) &\leq \sqrt{(T+1)^2 \varepsilon^2 \left(\frac{C^2}{\delta} + 1 \right) + (T+1) \left(C^2 + \delta \right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2}, \\
&\leq (T+1) \varepsilon \sqrt{\frac{C^2}{\delta} + 1} + \sqrt{(T+1) \left(C^2 + \delta \right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2},
\end{aligned} \qquad (33)
$$

We can now take the expectation over the z_t:

$$\sum_{t=0}^{T} f(\mathbf{w}_t) - f_\star \leq (T+1)\,\varepsilon\sqrt{\frac{C^2}{\delta}+1} + \sqrt{(T+1)\,(C^2+\delta)\,\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}. \tag{34}$$

Dividing by $T+1$ and exploiting convexity of f, we finally get:

$$\begin{aligned} f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star &\leq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}_t) - f_\star \quad \text{(convexity of } f\text{)} \\ &\leq \varepsilon\sqrt{\frac{C^2}{\delta}+1} + \sqrt{\frac{\left(C^2+\delta\right)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}}. \end{aligned} \tag{35}$$

■

B.4 Theorem 3

Lemma 1. *Let $z \in \mathcal{Z}$. Assume that ℓ_z is convex, β-smooth and is lower-bounded on $\mathbb{R}^p$ by $\underline{\mathrm{B}} \in \mathbb{R}$. Then we have:*

$$\forall\, \mathbf{w} \in \mathbb{R}^p,\ \ell_z(\mathbf{w}) - \underline{\mathrm{B}} \geq \frac{1}{2\beta}\|\nabla\ell_z(\mathbf{w})\|^2 \tag{36}$$

Proof: Let $\mathbf{w} \in \mathbb{R}^p$ and suppose that ℓ_z reaches its infimum at $\underline{\mathbf{w}} \in (\mathbb{R} \cup \{-\infty, +\infty\})^p$.

First, let us consider the case $\underline{\mathbf{w}} \in \mathbb{R}^p$. Then by Lemma 3.5 of [Bub15], we have:

$$\begin{aligned} \ell_z(\underline{\mathbf{w}}) - \ell_z(\mathbf{w}) &\leq \nabla\ell_z(\underline{\mathbf{w}})^\top(\underline{\mathbf{w}} - \mathbf{w}) - \frac{1}{2\beta}\|\nabla\ell_z(\underline{\mathbf{w}}) - \nabla\ell_z(\mathbf{w})\|^2, \\ &= -\frac{1}{2\beta}\|\nabla\ell_z(\mathbf{w})\|^2 \quad (\nabla\ell_z(\underline{\mathbf{w}}) = \mathbf{0}). \end{aligned} \tag{37}$$

Therefore we can write:

$$\ell_z(\mathbf{w}) - \underline{\mathrm{B}} \geq \ell_z(\mathbf{w}) - \ell_z(\underline{\mathbf{w}}) \geq \frac{1}{2\beta}\|\nabla\ell_z(\mathbf{w})\|^2. \tag{38}$$

Now let us assume that $\underline{\mathbf{w}} \notin \mathbb{R}^p$. Then we can construct a sequence $(\underline{\mathbf{w}}_k)_{k\in\mathbb{N}} \in (\mathbb{R}^p)^{\mathbb{N}}$ that converges to $\underline{\mathbf{w}}$. Since ℓ_z and $\nabla\ell_z$ are continuous functions (they are respectively convex and smooth), we have:

$$\begin{aligned} \lim_{k\to\infty} \ell_z(\underline{\mathbf{w}}_k) &= \inf \ell_z, \\ \lim_{k\to\infty} \nabla\ell_z(\underline{\mathbf{w}}_k) &= \mathbf{0}. \end{aligned} \tag{39}$$

Therefore the previous case gives the wanted result by using $\underline{\mathbf{w}}_k$ in place of $\underline{\mathbf{w}}$ and then taking the limit $k \to \infty$.

■

Lemma 2. *Let $z \in \mathcal{Z}$. Assume that ℓ_z is convex, β-smooth and is lower-bounded on $\mathbb{R}^p$ by $\underline{\mathrm{B}} \in \mathbb{R}$. Then we have:*

$$\forall\, \mathbf{w} \in \mathbb{R}^p,\ \frac{\ell_z(\mathbf{w}) - \underline{\mathrm{B}}}{\|\nabla\ell_z(\mathbf{w})\|^2 + \delta} \geq \frac{1}{2\beta} - \frac{\delta}{4\beta^2(\ell_z(\mathbf{w}) - \underline{\mathrm{B}}))} \tag{40}$$

Proof:

Let $\mathbf{w} \in \mathbb{R}^p$. We apply Lemma 1 and we write successively:

$$\begin{aligned}
\frac{\ell_z(\mathbf{w}) - \underline{\mathrm{B}}}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} &\geq \frac{\ell_z(\mathbf{w}) - \underline{\mathrm{B}}}{2\beta(\ell_z(\mathbf{w}) - \underline{\mathrm{B}}) + \delta}, \quad \text{(Lemma 1)} \\
&= \frac{\ell_z(\mathbf{w}) - \underline{\mathrm{B}} + \frac{\delta}{2\beta} - \frac{\delta}{2\beta}}{2\beta(\ell_z(\mathbf{w}) - \underline{\mathrm{B}} + \frac{\delta}{2\beta})}, \\
&= \frac{1}{2\beta} - \frac{\frac{\delta}{2\beta}}{2\beta(\ell_z(\mathbf{w}) - \underline{\mathrm{B}} + \frac{\delta}{2\beta})}, \\
&\geq \frac{1}{2\beta} - \frac{\delta}{4\beta^2(\ell_z(\mathbf{w}) - \underline{\mathrm{B}})}. \quad (\delta \geq 0)
\end{aligned} \tag{41}$$

■

Theorem 3. *[Convex and Smooth] We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\underline{\mathrm{B}}$ be a uniform lower bound on $(\mathcal{P})$ and $\mathbf{w}_\star$ be a solution of $(\mathcal{P})$. Further suppose that $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, and that $\delta > 2\beta\varepsilon$. Then ALI-G$^\infty$ applied to f satisfies:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\delta}{\beta(1 - \frac{2\beta\varepsilon}{\delta})} + \frac{2\beta}{1 - \frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}. \tag{11}$$

In other words, f approximately converges to $f_\star$ at a rate of $\mathcal{O}(1/T)$.

Proof:

We consider the update at time t, which we condition on the draw of $z_t \in \mathcal{Z}$:

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\
&\quad = \|\Pi_\Omega(\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)) - \mathbf{w}_\star\|^2 \\
&\quad \leq \|\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t) - \mathbf{w}_\star\|^2 \quad (\Pi_\Omega \text{ projection}) \\
&\quad = \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t^2 \|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \\
&\quad = \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \\
&\qquad \text{(definition of } \gamma_t) \\
&\quad \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2}\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \\
&\qquad \text{(because } \ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}} \geq 0 \text{ and } \delta \geq 0) \\
&\quad \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}) \quad (\text{convexity of } \ell_{z_t}) \\
&\quad = \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \\
&\quad = \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})
\end{aligned} \tag{42}$$

We now lower bound $\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))$ and upper bound $\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})$ individually.

We begin with $\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))$, for which we distinguish two cases according to its sign:

Suppose $(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0$. Then we can write:

$$\begin{aligned}
&\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&\quad= \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)), \quad \text{(definition of } \gamma_t) \\
&\quad\geq \left(\frac{1}{2\beta} - \frac{\delta}{4\beta^2(\ell_z(\mathbf{w}_t) - \underline{\mathrm{B}})}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \quad \text{(Lemma 2, } (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0) \\
&\quad= \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2}\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)}{(\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}})} \\
&\quad\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \quad (\ell_{z_t}(\mathbf{w}_\star) \geq \underline{\mathrm{B}},\, (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0)
\end{aligned} \tag{43}$$

Now suppose $(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \leq 0$, and let us show that the same result holds. We have:

$$\begin{aligned}
\gamma_t &= \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\leq \frac{\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad ((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \leq 0) \\
&\leq \frac{\varepsilon}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad \text{(definition of } \varepsilon) \\
&\leq \frac{\varepsilon}{\delta} \quad (\|\nabla \ell_{z_t}(\mathbf{w}_t)\| \geq 0) \\
&\leq \frac{1}{2\beta} \quad (\delta \geq 2\beta\varepsilon)
\end{aligned} \tag{44}$$

Therefore we have:

$$\begin{aligned}
&\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&\quad\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \quad ((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \leq 0) \\
&\quad\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \quad (\delta \geq 0)
\end{aligned} \tag{45}$$

In conclusion, regardless of the sign, it always holds true that:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \tag{46}$$

We now upper bound $\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})$:

$$\begin{aligned}
\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) &= \frac{(\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}})(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \quad \text{(definition of } \gamma_t) \\
&\leq \frac{(\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}})(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})}{\delta}, \quad (\|\nabla \ell_{z_t}(\mathbf{w}_t)\| \geq 0) \\
&\leq \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \varepsilon)\varepsilon}{\delta}, \quad \text{(definition of } \varepsilon \text{ twice)} \\
&= \frac{\varepsilon}{\delta}((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}.
\end{aligned} \tag{47}$$

Putting inequalities (42), (46) and (47) together, we obtain:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{\delta}((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}. \tag{48}$$

Therefore we have:

$$\left(\frac{1}{2\beta}-\frac{\varepsilon}{\delta}\right)\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)-\left(\frac{\delta}{4\beta^2}+\frac{\varepsilon^2}{\delta}\right)\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2. \tag{49}$$

By summing over t and taking the expectation over the z_t, we obtain:

$$\frac{\delta-2\beta\varepsilon}{2\beta\delta}\sum_{t=0}^{T}\left(f(\mathbf{w}_t)-f(\mathbf{w}_\star)-\frac{\delta^2+4\beta^2\varepsilon^2}{4\beta^2\delta}\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2-\mathbb{E}\left[\|\mathbf{w}_{T+1}-\mathbf{w}_\star\|^2\right]\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2. \tag{50}$$

By assumption, we have that $\delta-2\beta\varepsilon>0$. Dividing by $T+1$ and using the convexity of f, we finally obtain:

$$\begin{aligned}
f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star &\leq \frac{1}{T+1}\sum_{t=0}^{T}f(\mathbf{w}_t)-f_\star \quad \text{(convexity of } f\text{)},\\
&= \frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\delta^2+4\beta^2\varepsilon^2}{4\beta^2\delta}+\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1},\\
&= \frac{\delta^2+4\beta^2\varepsilon^2}{2\beta(\delta-2\beta\varepsilon)}+\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1},\\
&\leq \frac{\delta^2}{\beta(\delta-2\beta\varepsilon)}+\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}, \quad (\delta-2\beta\varepsilon\geq 0)\\
&= \frac{\delta}{\beta(1-\frac{2\beta\varepsilon}{\delta})}+\frac{2\beta}{1-\frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}.
\end{aligned} \tag{51}$$

■

B.5 Theorem 4

Lemma 3. *For any $a,b\in\mathbb{R}^p$, we have that:*

$$\|a\|^2+\|b\|^2\geq\frac{1}{2}\|a-b\|^2 \tag{52}$$

Proof: This is a simple application of the parallelogram law, but we give the proof here for completeness.

$$\begin{aligned}
\|a\|^2+\|b\|^2-\frac{1}{2}\|a-b\|^2 &= \|a\|^2+\|b\|^2-\frac{1}{2}\|a\|^2-\frac{1}{2}\|b\|^2+a^\top b\\
&= \frac{1}{2}\|a\|^2+\frac{1}{2}\|b\|^2+a^\top b\\
&= \frac{1}{2}\|a+b\|^2\\
&\geq 0
\end{aligned}$$

■

Lemma 4. *Let $z\in\mathcal{Z}$. Assume that ℓ_z is α-strongly convex and is lower-bounded on $\mathbb{R}^p$ by $\underline{\mathrm{B}}\in\mathbb{R}$ such that $\inf\ell_z-\underline{\mathrm{B}}\leq\varepsilon$. In addition, suppose that $\delta\geq 2\alpha\varepsilon$. Then we have:*

$$\forall\,\mathbf{w}\in\mathbb{R}^p,\ \frac{\ell_z(\mathbf{w})-\underline{\mathrm{B}}}{\|\nabla\ell_z(\mathbf{w})\|^2+\delta}\leq\frac{1}{2\alpha}. \tag{53}$$

Proof:

Let $\mathbf{w}\in\mathbb{R}^p$ and suppose that ℓ_z reaches its minimum at $\underline{\mathbf{w}}\in\mathbb{R}^p$ (this minimum exists because of strong convexity). By definition of strong convexity, we have that:

$$\forall\,\hat{\mathbf{w}}\in\mathbb{R}^p,\ \ell_z(\hat{\mathbf{w}})\geq\ell_z(\mathbf{w})+\nabla\ell_z(\mathbf{w})^\top(\hat{\mathbf{w}}-\mathbf{w})+\frac{\alpha}{2}\|\hat{\mathbf{w}}-\mathbf{w}\|^2 \tag{54}$$

We minimize the right hand-side over $\hat{\mathbf{w}}$, which gives:

$$\forall \hat{\mathbf{w}} \in \mathbb{R}^p,\ \ell_z(\hat{\mathbf{w}}) \geq \ell_z(\mathbf{w}) + \nabla \ell_z(\mathbf{w})^\top (\hat{\mathbf{w}} - \mathbf{w}) + \frac{\alpha}{2}\|\hat{\mathbf{w}} - \mathbf{w}\|^2 \geq \ell_z(\mathbf{w}) - \frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 \tag{55}$$

Thus by choosing $\hat{\mathbf{w}} = \underline{\mathbf{w}}$ and re-ordering, we obtain the following result (a.k.a. the Polyak-Lojasiewicz inequality):

$$\ell_z(\mathbf{w}) - \ell_z(\underline{\mathbf{w}}) \leq \frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 \tag{56}$$

Therefore we can write:

$$\frac{\ell_z(\mathbf{w}) - \underline{\mathrm{B}}}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} \leq \frac{\ell_z(\mathbf{w}) - \ell_z(\underline{\mathbf{w}}) + \varepsilon}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} \leq \frac{\frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 + \varepsilon}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta}. \tag{57}$$

We introduce the function $\psi : x \in \mathbb{R}^+ \mapsto \dfrac{\frac{1}{2\alpha}x + \varepsilon}{x + \delta}$, and we compute its derivative:

$$\begin{aligned} \psi'(x) &= \frac{\frac{1}{2\alpha}(x+\delta) - \frac{1}{2\alpha}x - \varepsilon}{(x+\delta)^2}, \\ &= \frac{\frac{\delta}{2\alpha} - \varepsilon}{(x+\delta)^2} \geq 0. \quad (\delta \geq 2\alpha\varepsilon) \end{aligned} \tag{58}$$

Therefore ψ is monotonically increasing. As a result, we have:

$$\forall\ x \in \mathbb{R}^+,\ \psi(x) \leq \lim_{x\to\infty} \psi(x) = \frac{1}{2\alpha}. \tag{59}$$

Therefore we have that:

$$\frac{\frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 + \varepsilon}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} = \psi\left(\|\nabla \ell_z(\mathbf{w})\|^2\right) \leq \frac{1}{2\alpha}, \tag{60}$$

which concludes the proof. ■

Theorem 4. *[Strongly Convex and Smooth] We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\underline{\mathrm{B}}$ be a uniform lower bound on $(\mathcal{P})$ and $\mathbf{w}_\star$ be a solution of $(\mathcal{P})$. Further suppose that $\mathbf{w}_\star$ is an ε-interpolation for $((\mathcal{P}), \underline{\mathrm{B}})$, and that $\delta > 2\beta\varepsilon$. Then ALI-G$^\infty$ applied to f satisfies:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \beta \exp\left(-\frac{\alpha T}{8\beta}\right)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{2\delta}{\alpha} + \left(10\frac{\beta}{\alpha} + 4\frac{\beta^2}{\alpha^2}\right)\varepsilon. \tag{12}$$

In other words, f approximately converges to $f_\star$ at a rate of $\mathcal{O}(\exp(-\alpha T/8\beta))$.

Proof:

We condition the update on z_t drawn at random. The beginning of the proof is identical to that of Theorem 3 (and in particular requires $\delta > 2\beta\varepsilon$). In addition, we remark that $\delta > 2\beta\varepsilon \geq 2\alpha\varepsilon$, because it always holds true that $\beta \geq \alpha$. Combining inequalities (42) and (46), we obtain:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}), \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \gamma_t\varepsilon, \quad \text{(definition of } \varepsilon) \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}. \quad \text{(Lemma 4)} \end{aligned} \tag{61}$$

Let $\underline{\mathbf{w}}^{(z_t)}$ be the minimizer of ℓ_{z_t} on its unconstrained domain $\mathbb{R}^p$ (its existence is guaranteed by the strong convexity property). Then we exploit strong convexity to lower bound the progress made:

$$\begin{aligned}
\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) &= \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) + \ell_{z_t}(\mathbf{w}_\star) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) - 2(\ell_{z_t}(\mathbf{w}_\star) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)})) \\
&\geq \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) + \ell_{z_t}(\mathbf{w}_\star) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) - 2(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \\
&\quad \text{(because } \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) \geq \underline{\mathrm{B}}) \\
&\geq \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) + \ell_{z_t}(\mathbf{w}_\star) - \ell_{z_t}(\underline{\mathbf{w}}^{(z_t)}) - 2\varepsilon \quad \text{(definition of } \varepsilon) \\
&\geq \frac{\alpha}{2}\|\mathbf{w}_t - \underline{\mathbf{w}}^{(z_t)}\|^2 + \frac{\alpha}{2}\|\mathbf{w}_\star - \underline{\mathbf{w}}^{(z_t)}\|^2 - 2\varepsilon \quad (\alpha\text{-strong convexity}) \\
&\geq \frac{\alpha}{4}\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\varepsilon \quad \text{(Lemma 3)}
\end{aligned} \tag{62}$$

We now combine inequalities (61) and (62):

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \left(1 - \frac{\alpha}{8\beta}\right)\|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \frac{\varepsilon}{\beta} + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha} \tag{63}$$

We use a trivial induction over t and write:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \left(1 - \frac{\alpha}{8\beta}\right)\|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \frac{\varepsilon}{\beta} + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}, \\
&\leq \left(1 - \frac{\alpha}{8\beta}\right)^t \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \sum_{k=0}^{t}\left(1 - \frac{\alpha}{8\beta}\right)^{t-k}\left(\frac{\varepsilon}{\beta} + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}\right), \\
&\leq \left(1 - \frac{\alpha}{8\beta}\right)^t \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \sum_{k=0}^{\infty}\left(1 - \frac{\alpha}{8\beta}\right)^{k}\left(\frac{\varepsilon}{\beta} + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}\right), \\
&= \left(1 - \frac{\alpha}{8\beta}\right)^t \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{1}{\frac{\alpha}{8\beta}}\left(\frac{\varepsilon}{\beta} + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}\right), \\
&= \left(1 - \frac{\alpha}{8\beta}\right)^t \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{8\beta}{\alpha}\left(\frac{\varepsilon}{\beta} + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}\right).
\end{aligned} \tag{64}$$

In particular, we remark that the right hand-side of the equation is independent on z_t.

Given an arbitrary $\mathbf{w} \in \mathbb{R}^p$, we now wish to relate the distance $\|\mathbf{w} - \mathbf{w}_\star\|^2$ to the function values $f(\mathbf{w}) - f(\mathbf{w}_\star)$.

Since each ℓ_z is α-strongly convex and β-smooth, so is $f = \mathbb{E}_z[\ell_z]$. We introduce $\underline{\mathbf{w}}$ the minimizer of f on its unconstrained domain $\mathbb{R}^p$. Then we can write that for any $\mathbf{w} \in \mathbb{R}^p$:

$$\begin{aligned}
f(\mathbf{w}) - f(\mathbf{w}_\star) &\leq f(\mathbf{w}) - f(\underline{\mathbf{w}}), \quad (f(\underline{\mathbf{w}}) \leq f(\mathbf{w}_\star)) \\
&\leq \nabla f(\underline{\mathbf{w}})^\top(\mathbf{w} - \underline{\mathbf{w}}) + \frac{\beta}{2}\|\mathbf{w} - \underline{\mathbf{w}}\|^2, \quad (f \text{ is } \beta\text{-smooth}) \\
&= \frac{\beta}{2}\|\mathbf{w} - \underline{\mathbf{w}}\|^2, \quad (\nabla f(\underline{\mathbf{w}}) = \mathbf{0}) \\
&\leq \beta(\|\mathbf{w} - \mathbf{w}_\star\|^2 + \|\mathbf{w}_\star - \underline{\mathbf{w}}\|^2), \quad \text{(Lemma 3)} \\
&\leq \beta\|\mathbf{w} - \mathbf{w}_\star\|^2 + \frac{2\beta}{\alpha}\left(f(\mathbf{w}_\star) - f(\underline{\mathbf{w}})\right), \quad (f \text{ is } \alpha\text{-strongly convex}) \\
&\leq \beta\|\mathbf{w} - \mathbf{w}_\star\|^2 + \frac{2\beta}{\alpha}\left(f(\mathbf{w}_\star) - \underline{\mathrm{B}}\right), \quad (\underline{\mathrm{B}} \leq f(\underline{\mathbf{w}}) \text{ by definition}) \\
&\leq \beta\|\mathbf{w} - \mathbf{w}_\star\|^2 + 2\frac{\beta\varepsilon}{\alpha}, \quad \text{(definition of } \varepsilon)
\end{aligned} \tag{65}$$

We combine the results to obtain the final result:

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}_\star) \leq \beta\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 + 2\frac{\beta\varepsilon}{\alpha},$$

$$
\begin{aligned}
&\leq \beta\left(\left(1-\frac{\alpha}{8\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{8\beta}{\alpha}\left(\frac{\varepsilon}{\beta}+\frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}\right)\right)+2\frac{\beta\varepsilon}{\alpha},\\
&=\beta\left(1-\frac{\alpha}{8\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{8\beta}{\alpha}\left(\varepsilon+\frac{\delta}{4\beta}+\frac{\varepsilon\beta}{2\alpha}\right)+2\frac{\beta\varepsilon}{\alpha},\\
&=\beta\left(1-\frac{\alpha}{8\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{2\delta}{\alpha}+\left(10\frac{\beta}{\alpha}+4\frac{\beta^2}{\alpha^2}\right)\varepsilon,\\
&\leq\beta\exp\left(-\frac{\alpha t}{8\beta}\right)\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{2\delta}{\alpha}+\left(10\frac{\beta}{\alpha}+4\frac{\beta^2}{\alpha^2}\right)\varepsilon. \qquad (66)
\end{aligned}
$$

■

B.6 Theorem 5

Theorem 5. *We assume that $\mathcal{X}$ is a convex set, and that for every $z\in\mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\mathbf{w}_\star$ be an ε-interpolation for ($(\mathcal{P})$, $\underline{\mathrm{B}}$). We further assume that $\eta>\frac{\varepsilon}{\delta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$
f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star\leq\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{(\eta-\frac{\varepsilon}{\delta})(T+1)}+\frac{\varepsilon^2}{\delta(\eta-\frac{\varepsilon}{\delta})}+\sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}}+\varepsilon\sqrt{\frac{C^2}{\delta}+1}. \qquad (13)
$$

Proof:

We consider the update at time t, which we condition on the draw of $z_t\in\mathcal{Z}$:

$$
\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\\
&\quad=\|\Pi_\Omega(\mathbf{w}_t-\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t))-\mathbf{w}_\star\|^2\\
&\quad\leq\|\mathbf{w}_t-\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)-\mathbf{w}_\star\|^2\quad(\Pi_\Omega\text{ projection})\\
&\quad=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t-\mathbf{w}_\star)+\gamma_t^2\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t-\mathbf{w}_\star)+\gamma_t\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2\\
&\qquad(\text{because }\gamma_t\leq\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta})\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t\nabla\ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t-\mathbf{w}_\star)+\gamma_t\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2}\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2\\
&\qquad(\text{because }\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}\geq0\text{ and }\delta\geq0)\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}})\quad(\text{convexity of }\ell_{z_t})\\
&\quad=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-2\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}})\\
&\quad=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}) \qquad (67)
\end{aligned}
$$

We now consider different cases, according to the value that γ_t takes: $\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$ or $\gamma_t=\eta$.

First, suppose that $\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$. Then we can follow the proof of Theorem 2 to obtain:

$$
\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\frac{(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))^2}{C^2+\delta}+\frac{\varepsilon^2}{\delta}. \qquad (68)
$$

Now suppose $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. We can use $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ to write:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}), \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\ &\quad + \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}),\end{aligned} \tag{69}$$

where the last inequality has used $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$, $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$ and $\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}} \geq 0$. Therefore we are exactly in the same situation as the first case (where we used $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$), and thus we have again:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\varepsilon^2}{\delta}. \tag{70}$$

Now suppose that $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. The inequality (67) gives:

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}), \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}), \quad (\gamma_t = \eta) \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t\varepsilon. \quad \text{(definition of } \varepsilon,\ \gamma_t \geq 0) \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \\ &\qquad \text{(because } \gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta},\ \varepsilon \geq 0) \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\delta}, \quad (\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \geq 0) \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}}{\delta}, \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \varepsilon}{\delta}, \\ &\qquad \text{(because } \ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}} \leq \varepsilon) \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \left(\eta - \frac{\varepsilon}{\delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}.\end{aligned} \tag{71}$$

We now introduce $\mathcal{I}_T$ and $\mathcal{J}_T$ as follows:

$$\begin{aligned}\mathcal{I}_T &\triangleq \{t \in \{0, ..., T\} : \gamma_t = \eta \text{ and } \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0\} \\ \mathcal{J}_T &\triangleq \{0, ..., T\} \setminus \mathcal{I}_T\end{aligned} \tag{72}$$

Then, by combining inequalities (68), (70) and (71), and using a telescopic sum, we obtain:

$$\begin{aligned}\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 &+ \sum_{t \in \mathcal{J}_T}\left(-\frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\varepsilon^2}{\delta}\right) \\ &+ \sum_{t \in \mathcal{I}_T}\left(-\left(\eta - \frac{\varepsilon}{\delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}\right)\end{aligned} \tag{73}$$

Using $\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \geq 0$, we obtain:

$$\frac{1}{C^2 + \delta}\sum_{t \in \mathcal{J}_T}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 + \left(\eta - \frac{\varepsilon}{\delta}\right)\sum_{t \in \mathcal{I}_T}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta} \tag{74}$$

In particular, the inequality (74) gives that:

$$\left(\eta - \frac{\varepsilon}{\delta}\right) \sum_{t \in \mathcal{I}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}. \tag{75}$$

Furthermore, for every $t \in \mathcal{I}_T$, we have $(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0$, which yields $\left(\eta - \frac{\varepsilon}{\delta}\right) \sum_{t \in \mathcal{I}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0$ since $\eta > \frac{\epsilon}{\delta}$. Thus the inequality (74) also gives:

$$\frac{1}{C^2 + \delta} \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}. \tag{76}$$

Using the Cauchy-Schwarz inequality, we can further write:

$$\left(\sum_{t \in \mathcal{J}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)\right)^2 \leq |\mathcal{J}_T| \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2. \tag{77}$$

Therefore we have:

$$\begin{aligned} \sum_{t \in \mathcal{J}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) &\leq \sqrt{|\mathcal{J}_T| \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}, \\ &\leq \sqrt{|\mathcal{J}_T|(C^2 + \delta)\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\right)}. \end{aligned} \tag{78}$$

We can now put together inequalities (75) and (77) by writing:

$$\begin{aligned} &\sum_{t=0}^{T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \\ &\quad = \sum_{t \in \mathcal{I}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \sum_{t \in \mathcal{J}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \\ &\quad \leq \frac{1}{\eta - \frac{\varepsilon}{\delta}}\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\right) + \sqrt{|\mathcal{J}_T|(C^2 + \delta)\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\right)} \\ &\quad \leq \frac{1}{\eta - \frac{\varepsilon}{\delta}}\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\right) + \sqrt{(T+1)(C^2 + \delta)\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\right)} \end{aligned} \tag{79}$$

Dividing by $T+1$ and taking the expectation, we obtain:

$$\begin{aligned} f\left(\frac{1}{T+1}\sum_{t=0}^{T} \mathbf{w}_t\right) - f_\star &\leq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}_t) - f_\star, \quad (f \text{ is convex}) \\ &\leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{(\eta - \frac{\varepsilon}{\delta})(T+1)} + \frac{\varepsilon^2}{\delta(\eta - \frac{\varepsilon}{\delta})} + \sqrt{(C^2 + \delta)\left(\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1} + \frac{\varepsilon^2}{\delta}\right)}, \\ &\leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{(\eta - \frac{\varepsilon}{\delta})(T+1)} + \frac{\varepsilon^2}{\delta(\eta - \frac{\varepsilon}{\delta})} + \sqrt{\frac{(C^2 + \delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} \\ &\quad + \varepsilon\sqrt{\frac{C^2}{\delta} + 1}. \end{aligned} \tag{80}$$

■

B.7 Theorem 6

Theorem 6. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\mathbf{w}_\star$ be an ε-interpolation for ($(\mathcal{P})$, B). Then if we apply ALI-G with a maximal learning-rate of η to f, we*

have:

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star\leq\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{\eta(T+1)}+\varepsilon+\sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}}+\eta\varepsilon\sqrt{C^2+\delta}. \tag{14}$$

Proof:

We consider the update at time t, which we condition on the draw of $z_t\in\mathcal{Z}$. We re-use the inequality (67) from the proof of Theorem 5:

$$\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}) \tag{81}$$

We consider again different cases, according to the value of γ_t and the sign of $\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)$.

Suppose that $\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\leq 0$. Then the inequality (81) gives:

$$\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}),\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}),\\
&\qquad\text{(because }\gamma_t\leq\eta,\ \ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\leq 0)\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t\varepsilon,\quad\text{(definition of }\varepsilon,\ \gamma_t\geq 0)\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta\varepsilon,\quad(\gamma_t\leq\eta,\ \varepsilon\geq 0)
\end{aligned} \tag{82}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0$ and $\gamma_t=\eta$. Then the inequality (81) gives:

$$\begin{aligned}
\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2&\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}),\\
&=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}),\quad(\gamma_t=\eta)\\
&\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta\varepsilon,\quad\text{(definition of }\varepsilon,\ \eta\geq 0)
\end{aligned} \tag{83}$$

Finally, suppose that $\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0$ and $\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$. Then the inequality (81) gives:

$$\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}),\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta(\ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}),\quad(\gamma_t\leq\eta,\ \ell_{z_t}(\mathbf{w}_\star)-\underline{\mathrm{B}}\geq 0)\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta\varepsilon,\quad\text{(definition of }\varepsilon,\ \eta\geq 0)\\
&\quad=\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta\varepsilon,\quad(\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta})\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\frac{(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))^2}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}+\eta\varepsilon,\quad(\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}\geq\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0)\\
&\quad\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\frac{(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))^2}{C^2+\delta}+\eta\varepsilon,\quad(\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2\leq C^2)
\end{aligned} \tag{84}$$

We now introduce $\mathcal{I}_T$ and $\mathcal{J}_T$ as follows:

$$\begin{aligned}
\mathcal{J}_T&\triangleq\left\{t\in\{0,...,T\}:\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)-\underline{\mathrm{B}}}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}\text{ and }\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0\right\}\\
\mathcal{I}_T&\triangleq\{0,...,T\}\setminus\mathcal{I}_T
\end{aligned} \tag{85}$$

Then, by combining inequalities (82), (83) and (84), and using a telescopic sum, we obtain:

$$\begin{aligned}\|\mathbf{w}_{T+1}-\mathbf{w}_\star\|^2 \leq \|\mathbf{w}_0-\mathbf{w}_\star\|^2 &+ \sum_{t\in\mathcal{J}_T}\left(-\frac{(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))^2}{C^2+\delta}+\eta\varepsilon\right)\\ &+\sum_{t\in\mathcal{I}_T}\left(-\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\eta\varepsilon\right)\end{aligned} \tag{86}$$

Using $\|\mathbf{w}_{T+1}-\mathbf{w}_\star\|^2 \geq 0$, we obtain:

$$\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)^2+\eta\sum_{t\in\mathcal{I}_T}\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon \tag{87}$$

We now take the expectation and obtain:

$$\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\mathbb{E}\left[\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)^2\right]+\eta\sum_{t\in\mathcal{I}_T}\left(f(\mathbf{w}_t)-f_\star\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon \tag{88}$$

Since $\mathbb{E}[U]^2\leq\mathbb{E}[U^2]$ for any real-valued random variable, we can write:

$$\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2+\eta\sum_{t\in\mathcal{I}_T}\left(f(\mathbf{w}_t)-f_\star\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon \tag{89}$$

Since each $f(\mathbf{w}_t)-f_\star\geq 0$, the inequality (89) gives that:

$$\eta\sum_{t\in\mathcal{I}_T}\left(f(\mathbf{w}_t)-f_\star\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon, \tag{90}$$

and:

$$\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon. \tag{91}$$

Using the Cauchy-Schwarz inequality, we can further write:

$$\left(\sum_{t\in\mathcal{J}_T}f(\mathbf{w}_t)-f_\star\right)^2\leq|\mathcal{J}_T|\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2. \tag{92}$$

Therefore we have:

$$\begin{aligned}\sum_{t\in\mathcal{J}_T}f(\mathbf{w}_t)-f_\star &\leq \sqrt{|\mathcal{J}_T|\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2},\\ &\leq\sqrt{|\mathcal{J}_T|(C^2+\delta)\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)}.\end{aligned} \tag{93}$$

We can now put together inequalities (90) and (93) by writing:

$$\begin{aligned}&\sum_{t=0}^{T}f(\mathbf{w}_t)-f_\star\\ &=\sum_{t\in\mathcal{I}_T}f(\mathbf{w}_t)-f_\star+\sum_{t\in\mathcal{J}_T}f(\mathbf{w}_t)-f_\star\\ &\leq\frac{1}{\eta}\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)+\sqrt{|\mathcal{J}_T|(C^2+\delta)\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)}\\ &\leq\frac{1}{\eta}\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)+\sqrt{(T+1)(C^2+\delta)\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)}\end{aligned} \tag{94}$$

Dividing by $T+1$, we obtain:

$$\begin{aligned} f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star &\leq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}_t) - f_\star, \quad (f \text{ is convex}) \\ &\leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \varepsilon + \sqrt{(C^2+\delta)\left(\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1} + \eta\varepsilon\right)}, \qquad (95) \\ &\leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \varepsilon + \sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} + \eta\varepsilon\sqrt{C^2+\delta}. \end{aligned}$$

■

B.8 Theorem 7

Theorem 7. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for ($(\mathcal{P})$, $\underline{\mathrm{B}}$), and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \geq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\delta}{\beta(1-\frac{2\beta\varepsilon}{\delta})} + \frac{2\beta}{1-\frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}. \qquad (15)$$

Proof:

By using the fact that $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ rather than $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$, we can see that the inequality (42) is still valid:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \qquad (96)$$

As previously, we lower bound $\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))$ and upper bound $\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})$ individually.

We begin with $\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))$. We remark that either $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ or $\gamma_t = \eta$.

Suppose $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. Then we are in the same condition as in Theorem 3, and thus the inequality (46) holds true:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2}. \qquad (97)$$

Now suppose $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. Then we have:

$$\begin{aligned} \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) &= \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\ &\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \\ &\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \\ &\qquad (\text{because } \eta \geq \frac{1}{2\beta},\ \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0). \qquad (98) \end{aligned}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. By using $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ instead of $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$, we can see that the inequality (44) is still valid, which gives:

$$\gamma_t \leq \frac{1}{2\beta} \tag{99}$$

We now use $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$ to write:

$$\begin{aligned} \gamma_t \left(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)\right) &\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \quad (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0) \\ &\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \end{aligned} \tag{100}$$

In conclusion, in all cases, it holds true that:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \tag{101}$$

By using $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$, we can remark that the inequality (47) holds true and gives:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \leq \frac{\varepsilon}{\delta}((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}. \tag{102}$$

We now put together inequalities (96), (101) and (102):

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{\delta}((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}, \\ &= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \left(\frac{1}{2\beta} - \frac{\varepsilon}{\delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon^2}{\delta}. \end{aligned} \tag{103}$$

This is exactly the same result as in the inequality (49) from the proof of Theorem 3. Therefore the rest of the proof of Theorem 3 follows and we obtain the desired result.

■

B.9 Theorem 8

Theorem 8. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for ($\mathcal{P}$), $\underline{\mathrm{B}}$), and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \leq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \frac{\delta}{2\beta} + \varepsilon. \tag{16}$$

Proof:

By using the fact that $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ rather than $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$, we can see that the inequality (42) is still valid:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \tag{104}$$

As previously, we lower bound $\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))$ and upper bound $\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})$ individually.

We begin with $\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))$. We remark that either $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ or $\gamma_t = \eta$.

Suppose $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. First we write:

$$\begin{aligned}
\gamma_t &= \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&= \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}} + \frac{\delta}{2\beta}}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} - \frac{\frac{\delta}{2\beta}}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\geq \frac{\frac{\|\ell_{z_t}(\mathbf{w}_t)\|^2}{2\beta} + \frac{\delta}{2\beta}}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} - \frac{\delta}{2\beta}\frac{1}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad \text{(Lemma 1)} \\
&= \frac{1}{2\beta} - \frac{\delta}{2\beta}\frac{1}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\geq \eta - \frac{\delta}{2\beta}\frac{1}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad (\eta \leq \frac{1}{2\beta})
\end{aligned} \tag{105}$$

Since $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$, this yields:

$$\begin{aligned}
\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) &\geq \left(\eta - \frac{\delta}{2\beta}\frac{1}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&= \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{2\beta}\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{2\beta}\frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad (\ell_{z_t}(\mathbf{w}_\star) \geq \underline{\mathrm{B}})
\end{aligned} \tag{106}$$

We now notice that since $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$, and $\gamma_t \leq \eta$, then necessarily $\frac{\ell_{z_t}(\mathbf{w}_t) - \underline{\mathrm{B}}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \leq \eta$. This gives:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta} \tag{107}$$

Now suppose $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. Then we have:

$$\begin{aligned}
\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) &= \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta}.
\end{aligned} \tag{108}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. Since $\gamma_t \leq \eta$ by definition, we have that:

$$\begin{aligned}
\gamma_t\left(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)\right) &\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \quad (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0) \\
&\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta}.
\end{aligned} \tag{109}$$

In conclusion, in all cases, it holds true that:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta} \tag{110}$$

We upper bound $\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}})$ as follows:

$$\begin{aligned}
\gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) &\leq \eta(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}) \quad (\ell_{z_t}(\mathbf{w}_\star) \geq \underline{\mathrm{B}}) \\
&\leq \eta\varepsilon \quad \text{(definition of } \varepsilon)
\end{aligned} \tag{111}$$

We combine inequalities (104), (110) and (111) and obtain:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \eta\varepsilon. \tag{112}$$

By taking the expectation and using a telescopic sum, we obtain:

$$0 \leq \|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \sum_{t=0}^{T} \left(\eta(f(\mathbf{w}_t) - f_\star) + \frac{\eta\delta}{2\beta} + \eta\varepsilon \right). \tag{113}$$

Re-arranging and using the convexity of f, we finally obtain:

$$f\left(\frac{1}{T+1} \sum_{t=0}^{T} \mathbf{w}_t \right) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \frac{\delta}{2\beta} + \varepsilon. \tag{114}$$

■

B.10 Theorem 9

Theorem 9. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for ($(\mathcal{P})$, $\underline{\mathrm{B}}$), and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \geq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \beta \exp\left(-\frac{\alpha T}{8\beta} \right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{2\delta}{\alpha} + \left(10\frac{\beta}{\alpha} + 4\frac{\beta^2}{\alpha^2} \right) \varepsilon. \tag{17}$$

Proof: Re-using inequalities (96) and (101) from the proof of Theorem 7, we obtain:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}). \tag{115}$$

This is exactly the same result as the first line of the inequality (61) in the proof of Theorem 4. Then the rest of the proof is identical to the one of Theorem 4. ■

B.11 Theorem 10

Theorem 10. *We assume that $\mathcal{X}$ is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\mathbf{w}_\star$ be an ε-interpolation for ($(\mathcal{P})$, $\underline{\mathrm{B}}$), and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \leq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \beta \exp\left(-\frac{\alpha\eta T}{4} \right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{2\delta}{\alpha} + \frac{14\varepsilon\beta}{\alpha}. \tag{18}$$

Proof: Re-using inequalities (104) and (110) from the proof of Theorem 8, we can write:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \gamma_t(\ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}}), \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \eta\varepsilon \quad (\gamma_t \leq \eta,\, 0 \leq \ell_{z_t}(\mathbf{w}_\star) - \underline{\mathrm{B}} \leq \varepsilon). \end{aligned} \tag{116}$$

Furthermore, the inequality (62) gives:

$$\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq \frac{\alpha}{4}\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\varepsilon \tag{117}$$

Therefore, we can write:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{\alpha\eta}{4}\|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \frac{\eta\delta}{2\beta} + 3\eta\varepsilon, \\ &= \left(1 - \frac{\alpha\eta}{4}\right) \|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \frac{\eta\delta}{2\beta} + 3\eta\varepsilon. \end{aligned} \tag{118}$$

Then a trivial induction gives that:

$$\begin{aligned}\|\mathbf{w}_{T+1}-\mathbf{w}_\star\|^2 &\leq \left(1-\frac{\alpha\eta}{4}\right)^T\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\left(\frac{\eta\delta}{2\beta}+3\eta\varepsilon\right)\sum_{t=0}^{T}\left(1-\frac{\alpha\eta}{4}\right)^t,\\ &\leq \left(1-\frac{\alpha\eta}{4}\right)^T\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\left(\frac{\eta\delta}{2\beta}+3\eta\varepsilon\right)\sum_{t=0}^{\infty}\left(1-\frac{\alpha\eta}{4}\right)^t,\\ &= \left(1-\frac{\alpha\eta}{4}\right)^T\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\left(\frac{\eta\delta}{2\beta}+3\eta\varepsilon\right)\frac{1}{1-\left(1-\frac{\alpha\eta}{4}\right)},\\ &= \left(1-\frac{\alpha\eta}{4}\right)^T\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{2\delta}{\alpha\beta}+\frac{12\varepsilon}{\alpha}.\end{aligned} \tag{119}$$

We now re-use the inequality (65) to write:

$$\begin{aligned}f(\mathbf{w}_{T+1})-f_\star &\leq \beta\|\mathbf{w}_{T+1}-\mathbf{w}_\star\|^2+\frac{2\beta\varepsilon}{\alpha},\\ &\leq \beta\left(1-\frac{\alpha\eta}{4}\right)^T\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{2\delta}{\alpha}+\frac{14\varepsilon\beta}{\alpha},\\ &\leq \beta\exp\left(\frac{-\alpha\eta T}{4}\right)\|\mathbf{w}_0-\mathbf{w}_\star\|^2+\frac{2\delta}{\alpha}+\frac{14\varepsilon\beta}{\alpha}.\end{aligned} \tag{120}$$

■

C Cross-Validation

We refer to the $\mathcal{L}_2$ regularization hyper-parameter as λ when the regularization is part of the objective and as ρ when it is incorporated as a constraint. The hyper-parameter η refers to the initial or maximal learning rate for SGD, ALI-G and adaptive gradient methods. We remind that we use baselines from [BZK19], and the results of their cross-validation can be found in Appendix B of their work.

C.1 SVHN

	Optimal Hyper-Parameter		
Optimizer	η or α	λ	ρ
Adagrad	0.01	5e-4	-
Adam	0.00001	5e-4	-
AMSGrad	0.0001	5e-4	-
BPGrad	0.01	5e-4	-
DFW	0.01	1e-4	-
L4Adam	0.15	1e-4	-
L4Mom	0.0015	5e-4	-
ALI-G	0.01	-	50

Table 5: *Cross-validation results on SVHN.*

C.2 SNLI

CE Loss	
	Optimal Hyper-Parameter
Optimizer	η or α
L4Adam	0.15
L4Mom	0.015
ALI-G	1

SVM Loss	
	Optimal Hyper-Parameter
Optimizer	η or α
L4Adam	0.015
L4Mom	0.015
ALI-G	1

Table 6: *Cross-validation results on SNLI.*

C.3 CIFAR

WRN on CIFAR-10				
	Optimal Hyper-Parameter			
Optimizer	η or α	λ	ρ	batch-size
L4Adam	0.015	1e-4	-	128
L4Mom	0.15	1e-4	-	128
ALI-G	1	-	100	256

DN on CIFAR-10				
	Optimal Hyper-Parameter			
Optimizer	η or α	λ	ρ	batch-size
L4Adam	0.15	1e-4	-	256
L4Mom	0.15	1e-4	-	64
ALI-G	0.1	-	75	64

WRN on CIFAR-100				
	Optimal Hyper-Parameter			
Optimizer	η or α	λ	ρ	batch-size
L4Adam	0.015	1e-4	-	512
L4Mom	0.015	5e-4	-	512
ALI-G	0.1	-	100	128

DN on CIFAR-100				
	Optimal Hyper-Parameter			
Optimizer	η or α	λ	ρ	batch-size
L4Adam	0.015	1e-4	-	256
L4Mom	0.015	1e-4	-	256
ALI-G	0.1	-	75	256

Table 7: *Cross-validation results on CIFAR.*