
Loss landscape Characterization of Neural Networks without Over-Parametrization

Rustem Islamov¹ Niccoló Ajroldi² Antonio Orvieto^{2,3,4} Aurelien Lucchi¹

¹University of Basel ²Max Planck Institute for Intelligent Systems ³ ELLIS Institute Tübingen
⁴Tübingen AI Center

Abstract

Optimization methods play a crucial role in modern machine learning, powering the remarkable empirical achievements of deep learning models. These successes are even more remarkable given the complex non-convex nature of the loss landscape of these models. Yet, ensuring the convergence of optimization methods requires specific structural conditions on the objective function that are rarely satisfied in practice. One prominent example is the widely recognized Polyak-Łojasiewicz (PL) inequality, which has gained considerable attention in recent years. However, validating such assumptions for deep neural networks entails substantial and often impractical levels of over-parametrization. In order to address this limitation, we propose a novel class of functions that can characterize the loss landscape of modern deep models without requiring extensive over-parametrization and can also include saddle points. Crucially, we prove that gradient-based optimizers possess theoretical guarantees of convergence under this assumption. Finally, we validate the soundness of our new function class through both theoretical analysis and empirical experimentation across a diverse range of deep learning models.

1 Introduction

The strides in empirical progress achieved by deep neural networks over the past decade have been truly remarkable. Central to the triumph of these techniques lies the effectiveness of optimization methods, which is particularly noteworthy given the non-convex nature of the objective functions under consideration. Worst-case theoretical results point to a pessimistic view since even a degree four polynomial can be NP-hard to optimize [30] and the loss landscape of some neural networks are known to include saddle points or bad local minima [6, 70, 88].

Yet, empirical evidence has shown that gradient-based optimizers – including SGD, AdaGrad [21] and Adam [39] among many others – can effectively optimize the loss of modern deep-learning-based models. While some have pointed to the ability of gradient-based optimizers to deal with potentially complex landscapes, e.g. escaping saddle points [36, 17], another potential explanation is that the loss landscape itself is less complex than previously assumed [27, 51].

Some key factors in this success include the choice of architecture [5, 48, 40, 18], as well as the over-parametrization [73, 12, 50, 51]. In the well-known infinite-width limit [86, 35, 32], neural networks are known to exhibit simple landscapes [49]. However, practical networks operate in a finite range, which still leaves a lot of uncertainty regarding the nature of the loss landscape. This is especially important given that the convergence guarantees of gradient-based optimizers are derived by assuming some specific structure on the objective function [37, 79, 71]. Consequently, an essential theoretical endeavor involves examining the class of functions that neural networks can represent.

In this work, we present a new class of functions that satisfy a newly proposed α - β -condition (see Eq. (2)). We theoretically and empirically demonstrate that these functions effectively characterize

Table 1: Summary of existing assumptions on the problem (1) and their limitations. Here \mathcal{S} denotes the set of minimizers of f and $f_i^* := \operatorname{argmin}_x f_i(x)$. Unlike earlier conditions, the α - β -condition is specifically designed to capture local minima and saddle points. NN = Neural Network.

Condition	Definition	Comments
QCvx [28]	$\langle \nabla f(x), x - x^* \rangle \geq \theta(f(x) - f(x^*))$ for some fixed $x^* \in \mathcal{S}$	- excludes saddle points and local minima
Aiming [51]	$\langle \nabla f(x), x - \operatorname{Proj}(x, \mathcal{S}) \rangle \geq \theta f(x)$	- excludes saddle points and local minima - theoretically holds for NN in the presence of impractical over-parameterization [51] - does not always hold in practice [Fig. 1 a-b]
PL ^(a) [66]	$\ \nabla f(x)\ ^2 \geq 2\mu(f(x) - f^*)$	- excludes saddle points and local minima - theoretically holds for NN in the presence of impractical over-parameterization [50] - does not always hold in practice [Fig. 1 c-d]
α - β -condition [This work]	$\langle \nabla f_i(x), x - \operatorname{Proj}(x, \mathcal{S}) \rangle \geq \alpha(f_i(x) - f_i(\operatorname{Proj}(x, \mathcal{S}))) - \beta(f_i(x) - f_i^*)$	- might have saddles [Ex. 2] and local minima [Ex. 3] - in practice does not require over-parameterization [Ex. 5]

the loss landscape of neural networks. Furthermore, we derive theoretical convergence guarantees for commonly used gradient-based optimizers under the α - β -condition.

In summary, we make the following contributions:

1. We introduce the α - β -condition and theoretically demonstrate its applicability to a wide range of complex functions, notably those that include local saddle points and local minima.
2. We empirically validate that the α - β -condition is a meaningful assumption that captures a wide range of practical functions, including matrix factorization and neural networks (ResNet, LSTM, GNN, Transformer, and other architectures).
3. We analyze the theoretical convergence of several optimizers under α - β -condition, including vanilla SGD (Stochastic Gradient Descent), SPS_{max} (Stochastic Polyak Step-size) [53], and NGN [63] (Non-negative Gauss-Newton).
4. We provide empirical and theoretical counter-examples where the weakest assumptions, such as the PL and Aiming conditions, do not hold, but the α - β -condition does.

2 Related work

2.1 Function classes in optimization

Studying the convergence properties of gradient-based optimizers has a long history in the field of optimization and machine learning. Notably, one of the fundamental observations is the linear and sub-linear convergence exhibited by GD for *strongly convex* (SCvx) and general *convex* (Cvx) functions [61]. However, most modern Machine Learning models have non-convex loss landscapes, for which the existing convex theory is not applicable. Without assumptions on the loss functions (other than smoothness), one can only obtain weak convergence guarantees to a first-order critical point. This situation has led to the derivation of assumptions that are weaker than convexity but that are sufficient to guarantee convergence of GD-based optimizers. The list includes *error bounds* (EB) [54], *essential strong convexity* (ESC) [52], weak strong convexity (WSC) [60], the restricted secant inequality (RSI) [87], and the quadratic growth (QG) condition [1]. In the neighborhood of the minimizer set \mathcal{S} , EB, PL, and QG are equivalent if the objective is twice differentiable [68]. All of them, except QG, are sufficient to guarantee a global linear convergence of GD. However, among these less stringent conditions, the Polyak-Łojasiewicz (PL) condition stands out as particularly renowned. Initially demonstrated by Polyak [66] to ensure linear convergence, it has recently experienced a resurgence of interest, in part because it accurately characterizes the loss landscape of heavily over-parametrized neural networks [50]. It was also shown to be one of the weakest assumptions among the other known conditions outlined so far [37]. A generalized form of the PL condition for non-smooth optimization is the Kurdyka-Łojasiewicz (KL) condition [43, 10] which is satisfied for a much larger class of functions [14, 77] than PL. The KL inequality has been employed to analyze the convergence of the classic proximal-point algorithm [3, 11, 47] and other optimization methods [4, 45].

More recently, some new convex-like conditions have appeared in the literature such as *star-convex* (StarCvx) [62], *quasar-convex* (QCvx) [28], and *Aiming* [51]. These conditions are relaxations of

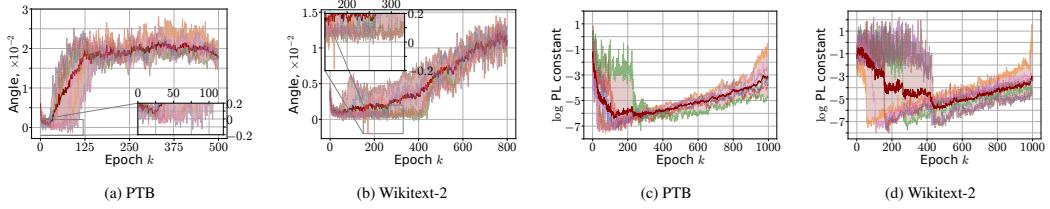


Figure 1: Training of 3 layer LSTM model that shows Aiming condition does not always hold. The term ‘‘Angle’’ in the figures refers to the angle $\angle(\nabla f(x^k), x^k - x^K)$, and it should be positive if Aiming holds, while in a-b we observe that it is negative during the first part of the training. Figures c-d demonstrate that possible constant μ in PL condition should be small which makes theoretical convergence slow.

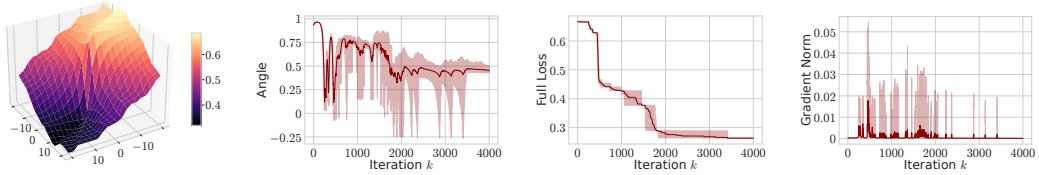


Figure 2: Training for half-space learning problem with SGD. The term ‘‘Angle’’ in the figures refers to the angle $\angle(\nabla f(x^k), x^k - x^K)$.

convexity and include non-convex functions. Within the domain of reinforcement learning, several works [84, 23] have also considered relaxations of the gradient domination condition, although these analyses are conducted specifically within the context of policy gradient methods, and therefore less relatable to StarCvx, QCvx or the Aiming condition.

We present a summary of some of these conditions in Table 1. There is no general implication between already existing assumptions such as QCvx, Aiming, PL, and the α - β -condition. However, as we will later see, the α - β -condition can more generally characterize the landscape of neural networks without requiring unpractical amounts of over-parametrization. Notably, the α - β -condition is a condition that applies globally to the loss. However, we will demonstrate that convergence guarantees can still be established for commonly-used gradient-based optimizers, although these guarantees are weaker than those derived under the PL condition, which relies on much stronger assumptions.

2.2 Limitations of existing conditions

Next, we discuss the limitations of previous conditions to characterize the loss landscape of complex objective functions such as the ones encountered when training deep neural networks.

Necessity of Over-parameterization. When considering deep models, the theoretical justification of conditions such as Aiming [51] and PL [50] require a significant amount of over-parameterization. This implies that the neural network must be considerably large, often with the minimum layer’s width scaling with the size of the dataset n . However, various studies suggest that this setup may not always accurately model real-world training dynamics [13, 2]. To the best of our knowledge, the weakest requirements on the width of a network are sub-quadratic $\tilde{\Omega}(n^{3/2})$ [20, 74], where n is the size of a dataset. This implies that, even for a small dataset such as MNIST, a network should have billions of parameters which is not a realistic setting in practice. In contrast, the α - β -condition condition does not require such an over-parametrization condition to hold (e.g., see Example 5). In Section 5 we provide empirical results showing how our condition is affected by over-parameterization.

Necessity of Invexity. One limitation of prior assumptions is their inability to account for functions containing local minima or saddle points. Indeed, many of the weakest conditions, such as QCvx, PL, KL, and Aiming, require that any point where the gradient is zero must be deemed a global minimizer. However, such conditions are not consistent with the loss landscapes observed in practical neural networks. For example, finite-size MLPs can have spurious local points or saddle points [88, 80]. Another known example is the half-space learning problem which is known to have saddle points [17]. We refer the reader to Figure 2-a that illustrates this claim (it showcases the surface of the problem

fixing all parameters except first 2), and also demonstrates that the Aiming and PL conditions fail to hold in such a setting. We present the results in Figure 2¹ where we observe that (i) the angle between the full gradient and direction to the minimizer $\angle(\nabla f(x^k), x^k - x^*)$ can be negative implying that the Aiming condition does not hold in this case (since the angle should remain positive); (ii) the gradient norm can become zero while we did not reach minimum (loss is still large) implying that the PL condition does not hold as well (since the inequality $\|\nabla f(x^k)\|^2 \geq 2\mu(f(x^k) - f^*)$ is not true for any positive μ). These observations suggest that the Aiming and PL conditions do not characterize well a landscape in the absence of invexity property. In contrast, we demonstrate in Example 2 and Figure 9 that our proposed assumption is preferable in this scenario.

Lack of Theory. As previously mentioned, most theoretical works apply to some infinite limit or neural networks of impractical sizes. In contrast, several works [89, 27, 75] have studied the empirical properties of the loss landscape of neural networks during training. They have shown that gradient-based optimization methods do not encounter significant obstacles that impede their progress. However, these studies fall short of providing theoretical explanations for this observed phenomenon.

Lack of Empirical Evidence. Several theoretical works [49, 51] prove results on the loss landscape of neural networks without supporting their claims using experimental validation on deep learning benchmarks. We demonstrate some practical counter-examples to these conditions proved in prior work. We train LSTM-based model² with standard initialization on Wikitext-2 [57] and Penn Treebank (PTB) [58] datasets. In Figure 1 (a-b), we show that the angle between the full gradient and direction to the minimizer $\angle(\nabla f(x^k), x^k - x^*)$ can be negative in the first part of the training. This result implies that the Aiming condition does not hold in this setting (either we do not have enough over-parameterization or the initialization does not lie in the locality region where Aiming holds). Moreover, for the same setting in Figure 1 (c-d) we plot $2 \log(\|\nabla f(x^k)\|) - 2/\delta \log(f(x^k) - f(x^K))$ ³ to measure the empirical value of PL constant $\log(2\mu)$ (see derivations in Appendix A). We observe that the value of μ that might satisfy PL condition should be of order $10^{-8} - 10^{-7}$ and leads to slow theoretical convergence [37]. These observations contradict with practical results. We defer to Appendix A for a more detailed discussion. In contrast, we demonstrate in Figure 7-(g-h) that the proposed α - β -condition can be verified in this setting.

3 The proposed α - β -condition

Setting. We consider the following Empirical Risk Minimization (ERM) problem that typically appears when training machine learning models:

$$\min_{x \in \mathbb{R}^d} [f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x)]. \quad (1)$$

Here $x \in \mathbb{R}^d$ denotes the parameters of a model we aim to train, d is the number of parameters, and n is the number of samples in the training dataset. Each $f_i(x)$ is the loss associated with the i -th data point. We denote the minimum of the problem (1) by f^* and the minimum of each individual function by $f_i^* := \min_x f_i(x)$, which we assume to be finite. Besides, the set \mathcal{S} denotes the set of all minimizers of f .

A new class of functions. Next, we present a new condition that characterizes the interplay between individual losses f_i and the set of minimizers of the global loss function f .

Definition 1 (α - β -condition). Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a set and consider a function $f : \mathcal{X} \rightarrow \mathbb{R}$ as defined in (1). Then f satisfies the α - β -condition with positive parameters α and β such that $\alpha > \beta$ if for any

¹Our implementation is based on the open source repository <https://github.com/archana95in/Escaping-saddles-using-Stochastic-Gradient-Descent> from [17] with small changes to track necessary quantities.

²Our implementation is based on the open source repository <https://github.com/fhueb/parameter-agnostic-lzlo> from [33] with small changes to track necessary quantities. The detailed experiment description is given in Appendix D.2.

³We use x^K (for a large value of K) to approximate the minimizer x^* .

$x \in \mathcal{X}$ there exists $x_p \in \text{Proj}(x, \mathcal{S})$ such that for all i ,

$$\langle \nabla f_i(x), x - x_p \rangle \geq \alpha(f_i(x) - f_i(x_p)) - \beta(f_i(x) - f_i^*). \quad (2)$$

The α - β -condition recovers several existing assumptions as special cases. For example, the proposed assumption reduces to QCvx around x^* if $\alpha > 0$, $\beta = 0$, and \mathcal{S} is a singleton $\{x^*\}$. Importantly, the α - β -condition is also applicable when the set \mathcal{S} contains multiple elements.

One can easily check that the pair of parameters (α, β) in (2) can not be unique. Indeed, if the assumption is satisfied for some α and β , then due to inequality $f_i(x_p) \geq f_i^*$ it will be also satisfied for $(\alpha + \delta, \beta + \delta)$ for any $\delta \geq 0$.

3.1 Theoretical verification of the α - β -condition

To demonstrate the significance of Definition 1 as a meaningful condition for describing structural non-convex functions, we provide several examples below that satisfy (2). We do not aim to provide the tightest possible choice of α and β such that Definition 1 holds. Instead, this section aims to offer a variety of examples that demonstrate specific desired characteristics when α - β -condition holds, encompassing a broad range of functions.

The initial example illustrates that \mathcal{S} could potentially be infinite for the class of functions satisfying Definition 1.

Example 1. Let $f, f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f = \frac{1}{2}(f_1 + f_2) \quad \text{with} \quad f_1(x, y) = \frac{(x+y)^2}{(x+y)^2+1}, \quad f_2(x, y) = \frac{(x+y+1)^2}{(x+y+1)^2+1}, \quad (3)$$

then Definition 1 holds with $\alpha \in [5/2, +\infty)$ and $\beta \in [4\alpha/5, \alpha)$.

Next, we provide an example where f satisfies Definition 1 even in the presence of saddle points.

Example 2. Let $f, f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f = \frac{1}{2}(f_1 + f_2) \quad \text{with} \quad f_1(x, y) = 1 - e^{-x^2-y^2}, \quad f_2(x, y) = 1 - e^{-(x-2)^2-(y-2)^2}, \quad (4)$$

then Definition 1 holds for some α and $\beta = \alpha - 8$.

Remark 1. Examples 1 and 2 can be generalized for any number of functions n and dimension d as follows

$$f_i(x) = \frac{(\sum_{j=1}^d x_j + a_i)^2}{(\sum_{j=1}^d x_j + a_i)^2 + 1}, \quad f_i(x) = \frac{\sum_{j=1}^d (x_j - b_{ij})^2}{1 + \sum_{j=1}^d (x_j - b_{ij})^2}, \quad (5)$$

for some properly chosen $\{a_i\}$ and $\{b_{ij}\}$, $i \in [n], j \in [d]$.

Example 3. Let $f, f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f = \frac{1}{2}(f_1 + f_2) \quad \text{with} \quad f_1(x, y) = \frac{1+x^2+y^2}{4+x^2+y^2}, \quad f_2(x, y) = \frac{(x-2.5)^2+(y-2.5)^2}{4+(y-2.5)^2+(y-2.5)^2}, \quad (6)$$

then Definition 1 holds for some α and $\beta = \alpha - 1$.

The three examples above demonstrate that functions satisfying Definition 1 can potentially be non-convex with an unbounded set of minimizers \mathcal{S} (Example 1) and can have saddle points (Example 2) and local minima (Example 3). In contrast, the PL and Aiming conditions are not met in cases where a problem exhibits saddle points. For illustration purposes, we plot the loss landscapes of f in Figure 3.

So far, we have presented simple examples to verify Definition 1. Next, we turn our attention to more practical examples in the field of machine learning. We start with the matrix factorization problem that is known to have saddle points [76] but can be shown to be PL after a sufficiently large number of iterations of alternating gradient descent and under a specific random initialization [81].

Example 4. Let f_i, f_{ij} be such that

$$f(W, S) = \frac{1}{2nm} \|X - W^\top S\|_F^2 = \frac{1}{2nm} \sum_{i,j} (X_{ij} - w_i^\top s_j)^2, \quad f_{ij}(W, S) = \frac{1}{2} (X_{ij} - w_i^\top s_j)^2, \quad (7)$$

where $X \in \mathbb{R}^{n \times m}$, $W = (w_i)_{i=1}^n \in \mathbb{R}^{k \times n}$, $S = (s_j)_{j=1}^m \in \mathbb{R}^{k \times m}$, and $\text{rank}(X) = r \geq k$. We assume that X is generated using matrices W^* and S^* with non-zero additive noise that minimize empirical loss, namely, $X = (W^*)^\top S^* + (\varepsilon_{ij})_{i \in [n], j \in [m]}$ where $W^*, S^* = \text{argmin}_{W, S} f(W, S)$. Let \mathcal{X} be any bounded set that contains \mathcal{S} . Then Definition 1 is satisfied with $\alpha = \beta + 1$ and some $\beta > 0$.

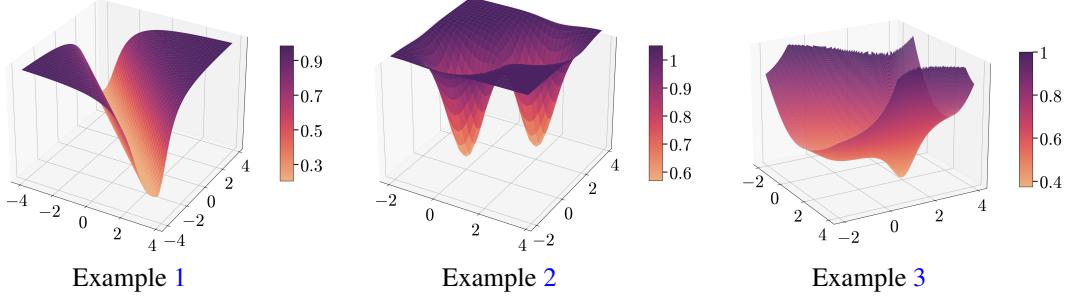


Figure 3: Loss landscape of f that satisfy Definition 1. The analytical form of f_i is given in Section 3.1. These examples demonstrate that the problem (1) that satisfies α - β -condition might have an unbounded set of minimizers \mathcal{S} (Example 1), a saddle point (Example 2), and local minima (Example 3) in contrast to the PL and Aiming conditions. The contour plots are presented in ??.

Example 5. Consider training a two-layer neural network with a logistic loss

$$f = \frac{1}{n} \sum_{i=1}^n f_i, \quad f_i(W, v) = \phi(y_i \cdot v^\top \sigma(Wx_i)) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 \quad (8)$$

for a classification problem where $\phi(t) := \log(1 + \exp(-t))$, $W \in \mathbb{R}^{k \times d}$, $v \in \mathbb{R}^k$, σ is a ReLU function applied coordinate-wise, $y_i \in \{-1, +1\}$ is a label and $x_i \in \mathbb{R}^d$ is a feature vector. Let \mathcal{X} be any bounded set that contains \mathcal{S} . Then the α - β -condition holds in \mathcal{X} for some $\alpha \geq 1$ and $\beta = \alpha - 1$.

Remark 2. The previous examples can be extended to any positive and convex function ϕ (e.g., square loss) with the additional assumption that each individual loss f_i does not have minimizers in \mathcal{S} , i.e. $\#(W^*, v^*) \in \mathcal{S}$ such that $f_i(W^*, v^*) = f_i^*$ for some $i \in [n]$.

We highlight that Example 5 is applicable for a bounded set \mathcal{X} of an *arbitrary* size. Moreover, in practice, we typically add ℓ_1 or ℓ_2 regularization which can be equivalently written as a constrained optimization problem, and therefore, Example 5 holds in this scenario. In comparison, the results from previous works do not hold for an arbitrary bounded set around \mathcal{S} requiring initialization to be close enough to the solution set [51, 50]. The proofs for all examples are given in Appendix C.1.

4 Theoretical convergence of algorithms

We conduct our analysis under the following smoothness assumption that is standard in the optimization literature.

Assumption 1. We assume that each f_i is L -smooth, i.e. for all $x, y \in \mathbb{R}^d$ it holds $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$.

The next assumption, which is sometimes called functional dissimilarity [56], is standard in the analysis of SGD with adaptive stepsizes [53, 26, 24].

Assumption 2. We assume that the interpolation error $\sigma_{\text{int}}^2 := \mathbb{E}_i[f^* - f_i^*]$ is finite, where the expectation is taken concerning the randomness of indices i for a certain algorithm.

4.1 Convergence under the α - β -condition

Now we demonstrate that the α - β -condition is sufficient for common optimizers to converge up to a neighbourhood of the set of minimizers \mathcal{S} . We provide convergence guarantees for SGD-based algorithms with fixed and adaptive stepsize (i.e., the update direction is of the form $-\gamma_k \nabla f_{i_k}(x^k)$). In this section, we only present the main statements about convergence while the algorithms' description and the proofs are deferred to Appendix C.2.

Convergence of SGD. We start with the results for vanilla SGD with constant stepsize.

Theorem 1. Assume that Assumptions 1-2 hold. Then the iterates of SGD (Alg. 1) with stepsize $\gamma \leq \frac{\alpha-\beta}{2L}$ satisfy

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{K} \frac{1}{\gamma(\alpha - \beta)} + \frac{2L\gamma}{\alpha - \beta} \sigma_{\text{int}}^2 + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2. \quad (9)$$

Table 2: Summary of how the non-vanishing term $\beta\sigma_{\text{int}}^2$ (as appearing e.g. in Eq. (9)) increases (\nearrow) or decreases (\searrow) as a function of specific quantities of interest.

	Model's width \nearrow	Model's depth \nearrow	Batch-size \nearrow
Change in $\beta\sigma_{\text{int}}^2$	\searrow	\searrow	\searrow

Theorem 1 shows that under the α - β -condition, SGD converges with a rate $\mathcal{O}(K^{-1/2})$ (the same rate obtained by SGD for convex functions [24]) up to a ball of size $\mathcal{O}(\beta\sigma_{\text{int}}^2)$. We argue that the non-vanishing term $\mathcal{O}(\beta\sigma_{\text{int}}^2)$ must appear in the convergence rate for non-convex optimization for several reasons: (i) This term arises directly from the use of the α - β condition in the analysis, without resorting to additional upper bounds or approximations. It reflects the potential existence of local minima that the α - β condition is designed to model. In the worst-case scenario, if SGD is initialized near local minima and uses a sufficiently small step size, it may fail to converge to the exact minimizer and can become trapped in suboptimal minima. This sub-optimality is modeled in the upper bound by the stepsize-independent quantity $\mathcal{O}(\beta\sigma_{\text{int}}^2)$ since we provide convergence guarantees for the function value sub-optimality rather than the squared gradient norm, which is more typical in the non-convex setting. (ii) We also observe that the last term in (9) shrinks as a model becomes more over-parameterized (which is consistent with prior works such as [50] that require large amounts of over-parametrization); see Sections 5.1, 5.2, and D.7 for experimental validation of this claim. Further empirical observations are summarized in Table 2 and will be discussed in Section 5. Theoretically, if a model is sufficiently over-parameterized such that the interpolation condition $f_i^* = f^*$ holds, then the non-vanishing term is not present in the bound. (iii) The presence of a non-vanishing error term in the rate with the α - β condition is consistent with empirical observation as it is frequently observed when training neural networks (see Figure 8.1 in [25]). This is also observed during the training of language models where the loss is significantly larger than 0 (see Figure 19). This phenomenon suggests that reaching a critical point, which is a global minimizer, is not commonly observed practically. (iv) Finally, we note a potential similarity with prior works that propose other conditions to describe the loss landscape of deep neural networks (e.g. gradient confusion [71]), and also obtain a non-vanishing term in the convergence rate (see Theorem 3.2 in [71]).

Convergence of SPS_{max}. Next, we consider the SPS_{max} algorithm. SPS_{max} stepsize is given by $\gamma_k := \min \left\{ \frac{f_{i_k}(x^k) - f_{i_k}^*}{c \|\nabla f_{i_k}(x^k)\|^2}, \gamma_b \right\}$ where c and γ_b are the stepsize hyperparameters.

Theorem 2. *Assume that Assumptions 1-2 hold. Then the iterates of SPS_{max} (Alg. 2) with a stepsize hyperparameter $c > \frac{1}{2(\alpha-\beta)}$ satisfy*

$$\min_{0 \leq k \leq K} \mathbb{E} [f(x^k) - f^*] \leq \frac{c_1}{K} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] + 2\alpha c_1 \gamma_b \sigma_{\text{int}}^2, \quad (10)$$

where $\gamma_{\min} := \min\{1/2cL, \gamma_b\}$ and $c_1 := \frac{c}{\gamma_{\min}(2(\alpha-\beta)c-1)}$.

In the convex case, i.e. α - β -condition holds with $\alpha = 1, \beta = 0$, we recover the rate of Loizou et al. [53].

Convergence of NGN. Finally, we turn to the analysis of NGN. This algorithm is proposed for minimizing positive functions f_i which is typically the case for many practical choices. Its stepsize $\gamma_k := \frac{\gamma}{1 + \frac{\gamma}{2f_{i_k}(x^k)} \|\nabla f_{i_k}(x^k)\|^2}$ where γ is the stepsize hyperparameter. NGN stepsize differs from that of SPS_{max} by replacing min operator by softer harmonic averaging of SPS stepsize and a constant γ . In addition to the already mentioned assumptions, we make a mild assumption that the positivity error $\sigma_{\text{pos}}^2 := \mathbb{E}[f_i^*]$ is finite.

Theorem 3. *Assume that Assumptions 1 with $\alpha \geq \beta + 1$ and 1-2 hold. Assume that each function f_i is positive and $\sigma_{\text{pos}}^2 < \infty$. Then the iterates of NGN (Alg. 3) with a stepsize parameter $\gamma > 0$ satisfy*

$$\begin{aligned} \min_{0 \leq k \leq K-1} \mathbb{E} [f(x^k) - f^*] &\leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{2\gamma K} \frac{(1 + 2\gamma L)^2}{c_2} + \frac{3L\gamma\alpha(1 + \gamma L)\sigma_{\text{int}}^2}{c_2} \\ &\quad + \frac{\gamma L}{a} \max \{2\gamma L - 1, 0\} \sigma_{\text{pos}}^2 + \frac{2\beta\sigma_{\text{int}}^2}{c_2}, \end{aligned} \quad (11)$$

where $c_2 := 2\gamma L(\alpha - \beta - 1) + \alpha - \beta$.

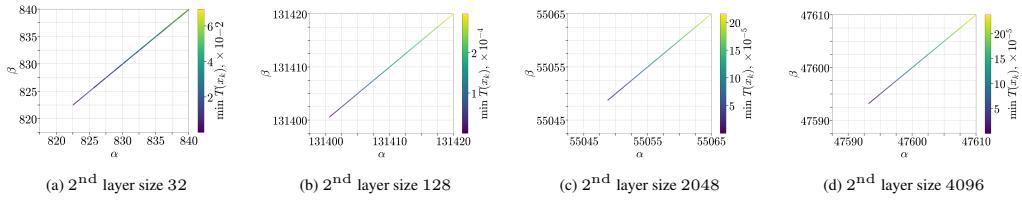


Figure 4: α - β -condition in the training of 3 layer MLP model on Fashion-MNIST dataset varying the size of the second layer. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for given pair of (α, β) .

One of the main properties of NGN is its robustness to the choice of stepsize γ . Theorem 3 can be seen as an extension of this feature from the set of convex functions originally analyzed in [63] to the class of structured non-convex satisfying α - β -condition.

Comparing the results of Theorems 2 and 3 we highlight several important differences. (i) There is no restriction on the stepsize parameter γ for NGN. Conversely, SPS_{\max} requires c to be lower bounded. (ii) Both algorithms converge to a neighborhood of the solution with a fixed stepsize hyperparameter. However, the neighborhood size of SPS_{\max} is not controllable by the stepsize hyperparameter and remains constant even in the convex setting when $\beta = 0$. In contrast, NGN converges to a ball whose size can be made smaller by choosing a small stepsize parameter, and the “non-vanishing” term disappears in the convex setting $\beta = 0$.

We note that our goal was not to achieve the tightest convergence guarantees for each algorithm, but rather to underscore the versatility of the α - β -condition in deriving convergence guarantees for SGD-type algorithms, both for constant or adaptive stepsizes. In addition to the results of this section, we demonstrate the convergence guarantees for SGD, SPS_{\max} , and NGN with decreasing with k stepsizes in Appendix C.2. Besides, in Appendix C.2.4 we present a convergence of a slightly modified version of Adagrad-norm method [82] under α - β -condition.

5 Experimental validation of the α - β -condition

In this section, we provide extensive numerical results supporting that the α - β -condition does hold in many practical applications for various tasks, model architectures, and datasets. The detailed experimental setting is described in Appendix D.

In all cases, we approximate $\text{Proj}(x^k, \mathcal{S})$ as the last iterate x^K in a run. After finding such an approximation, we start a second training run with the same random seed to measure all necessary quantities. To guarantee that the second training trajectory follows the same path as the first run, we disable non-deterministic CUDA operations while training on a GPU. For each task, we demonstrate possible values of pairs of (α, β) that work across all runs (might differ from one experiment to another) with different random seeds and satisfy $\alpha \geq \beta + 0.1$.

5.1 MLP architecture

First, we test MLP neural networks with 3 fully connected layers on Fashion-MNIST [83] dataset. We fix the second layer of the network to be a square matrix and vary its dimension layer to investigate the effect of over-parameterization on α - β -condition. We test it for dimensions $\{32, 128, 2049, 4096\}$, and for each case, we run experiments for 4 different random seeds. In Figure 4 we demonstrate possible values of pairs of (α, β) that work across all 4 runs. We observe that minimum possible values of α and β increase from small size to medium, and then tend to decrease again as the model becomes more over-parameterized. We defer more experimental results for MLP to Appendix D.3 to showcase this phenomenon. This observation leads to the fact that the neighborhood of convergence $\mathcal{O}(\beta\sigma_{\text{int}}^2)$ of SGD eventually becomes smaller with the size of the model as we expect (since it becomes more over-parameterized).

5.2 CNN architecture

In our next experiment, we test convolutional neural networks with 2 convolution layers and 1 fully connected layer on CIFAR10 dataset [41]. We vary the number of convolutions in the second convolution layer to investigate the effect of over-parameterization on α - β -condition. We test it

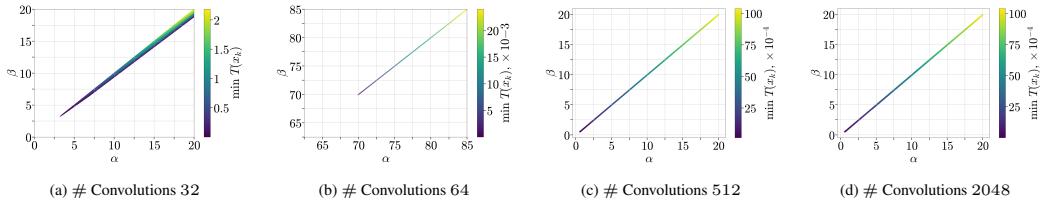


Figure 5: α - β -condition in the training of CNN model on CIFAR10 dataset varying the number of convolutions in the second layer. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for a given pair of (α, β) .

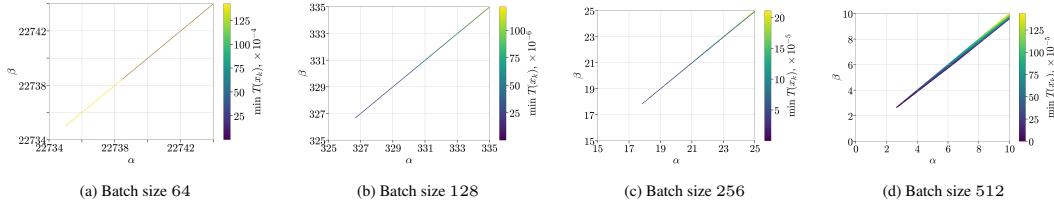


Figure 6: α - β -condition in the training of Resnet9 model on CIFAR100 dataset varying the batch size. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for a given pair of (α, β) .

for $\{32, 128, 512, 2048\}$ number of convolutions in the second layer, and for each case, we run experiments for 4 different random seeds. In Figure 5, we observe that the smallest possible values of (α, β) increase till 64 convolutions, and then decrease back. Second, the difference $\alpha - \beta$ for possible choice of α and β decreases from Figure 5-a to Figure 5-b, but then it increases again.

5.3 Resnet architecture

Next, we switch to the Resnet architecture [29] with batch sizes in $\{64, 128, 256, 512\}$ trained on CIFAR100 [41]. For each batch size, we run experiments for 4 different random seeds. In Figure 6, we plot the possible choice of pairs (α, β) that works across all runs. We observe that α - β -condition holds in all cases. Besides, there is a tendency for the minimum possible choice of α and β to decrease with batch size. Moreover, for larger batches, the difference between α and β also increases. From Theorem 1, this result suggests that we can use bigger stepsizes with larger batches.

5.4 Training of AlgoPerf workloads and transformers for language modeling

We are now interested in assessing the validity of α - β -condition on modern Deep Learning architectures. Thereby, we consider tasks from the AlgoPerf benchmarking suite [16]. We consider four workloads from the competition: (i) DLRMsmall model [59] on Criteo 1TB dataset [44]; (ii) U-Net model [69] on FastMRI dataset [85] (iii) GNN model [7] on OGBG dataset [31]; (iv) Transformer-big [78] on WMT dataset [9]. To train the models, we use NAdamW optimizer [19]⁴, which achieves state-of-the-art performances on the current version of the benchmark. The hyperparameters of the optimizer are chosen to reach the validation target threshold set by the original competition. Moreover, we consider the pretraining of a decoder-only transformer architecture [78] for causal language modeling. We conduct our evaluation on two publicly available Pythia models [8], of sizes 70M and 160M, trained on 1.25B and 2.5B tokens respectively. For this study, we use the SlimPajama [72] dataset. Following the original Pythia recipes, we fix a sequence length of 2048 and train the language model to predict the next token in a self-supervised fashion. We refer to section Appendices D and D.8 for additional details.

The results are presented in Figure 7. We observe that α - β -condition holds for a wide range of values of α and β which demonstrates that our condition can be seen as a good characterization of the training of modern large models as well. One can notice that the possible values of α and β for AlgoPerf workloads are higher than those for smaller models discussed in previous sections.

⁴Our implementation is based on open source AlgoPerf code <https://github.com/mlcommons/algorithmic-efficiency> with minimal changes to track necessary quantities.

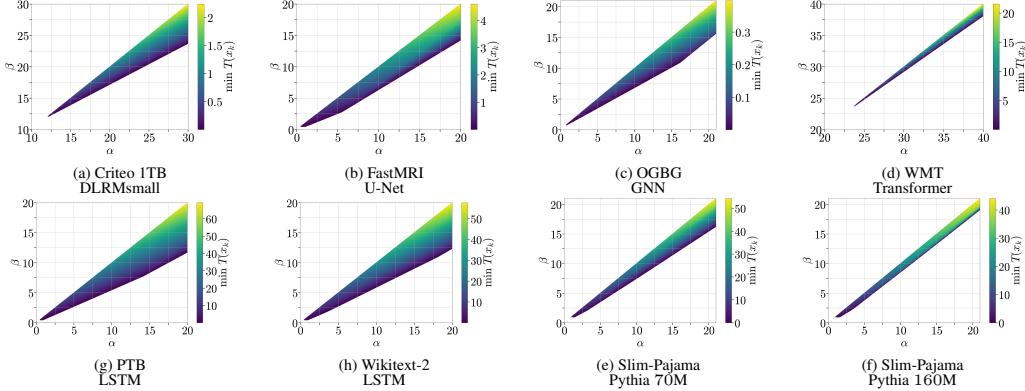


Figure 7: α - β -condition in the training of some large models from AlgoPerf, 3-layer LSTM, and Transformers for language modeling. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for a given pair of (α, β) .

This difference is attributable to smaller models interpolating the training data more effectively, resulting in significantly lower training losses compared to those observed in AlgoPerf experiments (see Appendix D for detailed results). However, we highlight that the convergence guarantees of the optimizers depend on a term $\mathcal{O}(\beta\sigma_{\text{int}}^2)$ which is stable across all experiments we provide.

5.5 Additional experiments

We defer the verification of α - β -condition by Adam and SGDM to Appendix D.6. The results in Figure 16 suggest that Adam explores the part of the landscape with smaller values of α and β than those for SGDM. The same conclusions can be drawn when comparing SGDM and SGD. These observations might be of the reasons why diagonal preconditioning and momentum are helpful in the training of DL models. The verification of the proposed condition varying the depth of Resnet model can be found in Appendix D.7. The results from Figure 17 demonstrate that the values of α and β decrease with the depth of Resnet model, i.e. the model becomes more over-parametrized.

6 Conclusion, potential extensions, and limitations.

In this work, we introduce a new class of functions that more accurately characterize loss landscapes of neural networks. In particular, we provide several examples that satisfy the proposed condition, including 2-layer ReLU-neural networks. Additionally, we prove that several optimization algorithms converge under our condition. Finally, we provide extensive empirical verification showing that the proposed α - β -condition holds along the optimization trajectories of various large deep learning models.

It is also possible to further expand convergence guarantees upon the ones presented in Section 4, for instance, by considering momentum [67] which is widely used in practice. However, we defer the exploration of other possible extensions to future research endeavors.

One of the limitations of this work is the empirical validation of the α - β -condition on neural networks. We are only able to verify this condition along the trajectories of specific optimizers. Even when performing checks with many random seeds, we cannot fully observe the entire loss landscape. Additionally, while our theoretical examples demonstrate that the proposed condition holds, we do not provide the most precise theoretical values of α and β that satisfy Definition 1. Therefore, in future work, we aim to obtain stronger theoretical guarantees demonstrating that the α - β -condition holds for neural networks with more precise values of α and β . We also intend to explore how the α - β -condition varies when changing the architecture or the number of parameters (i.e., theoretical exploration of over-parameterization).

Acknowledgement

Rustem Islamov and Aurelien Lucchi acknowledge the financial support of the Swiss National Foundation, SNF grant No 207392. Antonio Orvieto acknowledges the financial support of the Hector Foundation. The authors thank the anonymous reviewers for their valuable comments and suggestions on improving the paper.

References

- [1] Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 2000. (Cited on page 2)
- [2] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019. (Cited on page 3)
- [3] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009. (Cited on page 2)
- [4] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 2013. (Cited on page 2)
- [5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv: 1607.06450*, 2016. (Cited on page 1)
- [6] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 1989. (Cited on page 1)
- [7] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv: 1806.01261*, 2018. (Cited on page 9)
- [8] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023. (Cited on pages 9 and 41)
- [9] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. Findings of the 2017 conference on machine translation (wmt17). Association for Computational Linguistics, 2017. (Cited on page 9)
- [10] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of lojasiewicz inequalities and applications. *arXiv preprint arXiv:0802.0826*, 2008. (Cited on page 2)
- [11] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 2017. (Cited on page 2)
- [12] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018. (Cited on page 1)
- [13] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS 2019)*, 2019. (Cited on page 3)

[14] Michel Coste. *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000. (Cited on page 2)

[15] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. (Cited on page 37)

[16] George E. Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, Juhan Bae, Justin Gilmer, Abel L. Peirson, Bilal Khan, Rohan Anil, Mike Rabbat, Shankar Krishnan, Daniel Snider, Ehsan Amid, Kongtao Chen, Chris J. Maddison, Rakshith Vasudev, Michal Badura, Ankush Garg, and Peter Mattson. Benchmarking neural network training algorithms. *arXiv preprint arXiv: 2306.07179*, 2023. (Cited on pages 9 and 40)

[17] Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. (Cited on pages 1, 3, 4, and 37)

[18] Hadi Daneshmand, Amir Joudaki, and Francis Bach. Batch normalization orthogonalizes representations in deep random networks. In *Advances in Neural Information Processing Systems*, 2021. (Cited on page 1)

[19] Timothy Dozat. Incorporating nesterov momentum into adam. 2016. URL <https://openreview.net/pdf?id=0M0jvwB8jIp57ZJjtNEZ>. (Cited on page 9)

[20] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. (Cited on page 3)

[21] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011. (Cited on page 1)

[22] Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global kurdyka-lojasiewicz inequality. In *Advances in Neural Information Processing Systems*, 2022. (Cited on page 19)

[23] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, 2023. (Cited on page 3)

[24] Guillaume Garrigos and Robert Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv: 2301.11235*, 2024. (Cited on pages 6, 7, 25, and 34)

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. (Cited on page 7)

[26] Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021. (Cited on page 6)

[27] Charles Guille-Escuret, Hiroki Naganuma, Kilian Fatras, and Ioannis Mitliagkas. No wrong turns: The simple geometry of neural networks optimization paths. *arXiv preprint arXiv: 2306.11922*, 2023. (Cited on pages 1 and 4)

[28] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 2018. (Cited on page 2)

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. (Cited on page 9)

[30] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 2013. (Cited on page 1)

[31] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 2020. (Cited on page 9)

[32] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Proceedings of Advances in Neural Information Processing Systems*, 2019. (Cited on page 1)

[33] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024. (Cited on pages 4 and 37)

[34] Rustem Islamov, Mher Safaryan, and Dan Alistarh. AsGrad: A sharp unified analysis of asynchronous-SGD algorithms. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 649–657. PMLR, 2024. (Cited on page 19)

[35] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 2018. (Cited on page 1)

[36] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, 2017. (Cited on page 1)

[37] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. *Machine Learning and Knowledge Discovery in Databases*, 2016. (Cited on pages 1, 2, 4, and 19)

[38] Sarit Khirirat, Eduard Gorbunov, Samuel Horváth, Rustem Islamov, Fakhri Karray, and Peter Richtárik. Clip21: Error feedback for gradient clipping. *arXiv preprint: arXiv 2305.18929*, 2023. (Cited on page 19)

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv: 1412.6980*, 2017. (Cited on page 1)

[40] Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019. (Cited on page 1)

[41] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Lecture Notes*, 2009. (Cited on pages 8, 9, 38, and 39)

[42] Shailesh Kumar. Classifying cifar100 images using resnets, regularization and data augmentation in pytorch. <https://jovian.com/kumar-shailesh1597/cifar100-resnet18>, 2023. (Cited on page 39)

[43] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998. (Cited on page 2)

[44] Criteo A. I. Lab. Criteo 1tb click logs dataset. 2014. URL <https://ailab.criteo.com/download-criteo-1tb-click-logs-dataset/>. (Cited on page 9)

[45] Christian Lageman. *Convergence of gradient-like dynamical systems and optimization algorithms*. PhD thesis, Universität Würzburg, 2007. (Cited on page 2)

[46] Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. (Cited on page 35)

[47] Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 2018. (Cited on page 2)

[48] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. (Cited on page 1)

[49] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 2020. (Cited on pages 1 and 4)

[50] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022. (Cited on pages 1, 2, 3, 6, and 7)

[51] Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of SGD for overparametrized problems. In *Proceedings of Advances in Neural Information Processing Systems*, 2023. (Cited on pages 1, 2, 3, 4, and 6)

[52] Ji Liu, Steve Wright, Christopher Re, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, 2014. (Cited on page 2)

[53] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021. (Cited on pages 2, 6, 7, and 26)

[54] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 1993. (Cited on page 2)

[55] Maksim Makarenko, Elnur Gasanov, Abdurakhmon Sadiev, Rustem Islamov, and Peter Richtárik. Adaptive compression for communication-efficient distributed training. *Transactions on Machine Learning Research*, 2023. (Cited on page 19)

[56] Grigory Malinovsky, Konstantin Mishchenko, and Peter Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization. *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 2022. (Cited on page 6)

[57] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *Proceedings of 4th International Conference on Learning Representations (ICLR 2016)*, 2016. (Cited on page 4)

[58] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010. (Cited on page 4)

[59] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019. (Cited on page 9)

[60] I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 2019. (Cited on page 2)

[61] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004. (Cited on page 2)

[62] Yurii Nesterov and Boris T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 2006. (Cited on page 2)

[63] Antonio Orvieto and Lin Xiao. An adaptive stochastic gradient method with non-negative gauss-newton stepsizes. *arXiv preprint arXiv:2407.04358*, 2024. (Cited on pages 2, 8, 30, and 32)

[64] Antonio Orvieto, Simon Lacoste-Julien, and Nicolas Loizou. Dynamics of sgd with stochastic polyak stepsizes: Truly adaptive variants and convergence to exact solution. In *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 27, 29, and 34)

[65] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems*, 2019. (Cited on page 36)

[66] Boris T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 1963. (Cited on page 2)

[67] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical*, 1964. (Cited on page 10)

[68] Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for c_2 functions. *arXiv preprint arXiv:2303.00096*, 2023. (Cited on page 2)

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015. (Cited on page 9)

[70] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. (Cited on page 1)

[71] Karthik A. Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *Proceedings of the 37st International Conference on Machine Learning*, 2020. (Cited on pages 1 and 7)

[72] Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*, 2023. (Cited on pages 9 and 41)

[73] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*, 2015. (Cited on page 1)

[74] Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. In *Advances in Neural Information Processing Systems*, 2021. (Cited on page 3)

[75] Hoang Tran, Qinzi Zhang, and Ashok Cutkosky. Empirical tests of optimization assumptions in deep learning. *arXiv preprint arXiv:2407.01825*, 2024. (Cited on page 4)

[76] Hossein Valavi, Sulin Liu, and Peter Ramadge. Revisiting the landscape of matrix factorization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020. (Cited on pages 5 and 22)

[77] Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. 1996. (Cited on page 2)

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. (Cited on page 9)

[79] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models (and an accelerated perceptron). In *Proceedings of 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, 2019. (Cited on page 1)

[80] Luca Venturi, Afonso S. Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 2019. (Cited on page 3)

[81] Rachel Ward and Tamara Kolda. Convergence of alternating gradient descent for matrix factorization. *Advances in Neural Information Processing Systems*, 36:22369–22382, 2023. (Cited on page 5)

- [82] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. (Cited on page 8)
- [83] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv: 1708.07747*, 2017. (Cited on pages 8 and 37)
- [84] Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, 2022. (Cited on page 3)
- [85] Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael G. Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastmri: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv: 1811.08839*. (Cited on page 9)
- [86] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 2016. (Cited on page 1)
- [87] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv: 1303.4645*, 2013. (Cited on page 2)
- [88] Yi Zhou and Yingbin Liang. Critical points of linear neural networks: Analytical forms and landscape properties. In *Proceedings of International Conference on Learning Representations (ICLR 2018)*, 2018. (Cited on pages 1 and 3)
- [89] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *Proceedings of International Conference on Learning Representations (ICLR 2019)*, 2019. (Cited on page 4)

Contents

1	Introduction	1
2	Related work	2
2.1	Function classes in optimization	2
2.2	Limitations of existing conditions	3
3	The proposed α-β-condition	4
3.1	Theoretical verification of the α - β -condition	5
4	Theoretical convergence of algorithms	6
4.1	Convergence under the α - β -condition	6
5	Experimental validation of the α-β-condition	8
5.1	MLP architecture	8
5.2	CNN architecture	8
5.3	Resnet architecture	9
5.4	Training of AlgoPerf workloads and transformers for language modeling	9
5.5	Additional experiments	10
6	Conclusion, potential extensions, and limitations.	10
A	Additional explanation on PL assumption	19
B	Additional examples	19
C	Missing proofs	19
C.1	Proofs of examples satisfying definition 1	19
C.1.1	Proof of example 1	20
C.1.2	Proof of example 2	20
C.1.3	Proof of Example 3	21
C.1.4	Proof of example 6	22
C.1.5	Proof of example 4	22
C.1.6	Proof of example 5	23
C.2	Convergence of optimization algorithms under α - β -condition	24
C.2.1	Convergence of SGD	24
C.2.2	Convergence of SGD with Polyak Stepsize	26
C.2.3	Convergence of NGN	30
C.2.4	Convergence of AdaGrad-norm-max	34
D	Additional experiments	36
D.1	Half space learning	37

D.2	Experiment setup from section 2.2	37
D.3	MLP architecture	37
D.4	CNN architecture	38
D.5	Resnet architecture	39
D.6	Verification of α - β -condition by different optimizers	39
D.7	Increasing the depth of Resnet architecture	40
D.8	AlgoPerf experiments	40
D.9	Pythia experiments	41
E	Checklist	42

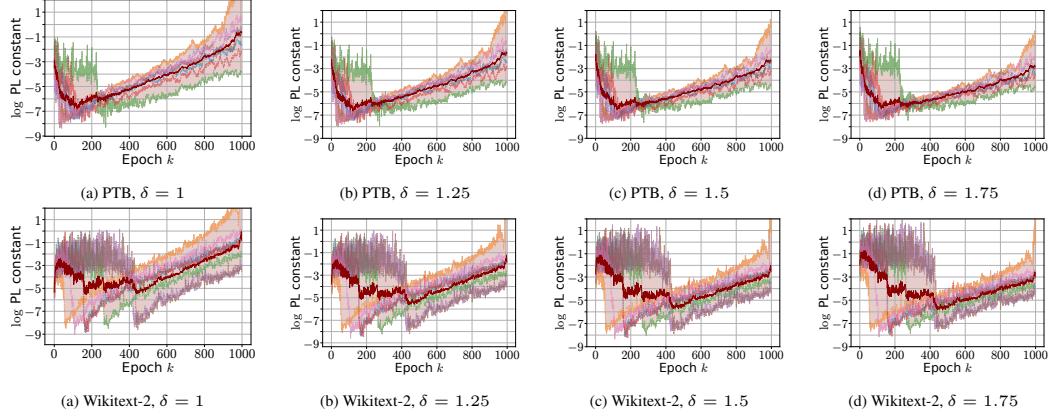


Figure 8: Training of 3 layer LSTM model that shows Aiming condition does not always hold. We plot possible values of PL constant for different powers δ in (12). We observe that possible values of μ are of order $10^{-9} - 10^{-7}$ which implies slow theoretical convergence and contradicts practical observations.

A Additional explanation on PL assumption

Let us consider the relaxation of PL condition for some $\delta \in [1, 2]$ and $\mu > 0$ (Assumption 3 in [22])

$$\|\nabla f(x)\|^\delta \geq (2\mu)^{\delta/2}(f(x) - f^*) \quad \text{for all } x \in \mathbb{R}^d. \quad (12)$$

Taking logarithms of both sides we continue

$$2 \log(\|\nabla f(x)\|) - \frac{2}{\delta} \log(f(x) - f^*) \geq \log(2\mu). \quad (13)$$

To satisfy PL condition for some δ , we need to find μ that satisfies (13) for all iterations. In Figure 8 we plot LHS of (13). We observe that the values of μ that satisfy (13) for all iterations should be of order $10^{-9} - 10^{-7}$ that leads to slow theoretical convergence [37]. Therefore, we claim that PL condition does not hold globally for neural networks. Nevertheless, it might hold locally around the solution (the value of μ closer to the end of the training are large enough).

B Additional examples

In this section, we list additional examples that satisfy Definition 1 and might have non-optimal stationary points. The functions of this form are typically used to simulate non-convex optimization problems and test the performance of algorithms on small or synthetic datasets [34, 55, 38].

Example 6. Let $f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f_1(x, y) = \frac{x^2 + y^2}{1 + x^2 + y^2}, \quad f_2(x, y) = \frac{(x-1)^2 + (y-1)^2}{1 + (x-1)^2 + (y-1)^2}, \quad (14)$$

then Definition 1 holds with $\alpha \geq \frac{2}{\min_i \min_{(z,t) \in \mathcal{S}} f_i(z,t)} \approx 41.369325$ and $\beta = \alpha - 1$. Besides, Definition 1 does not hold with $\beta = 0$ not satisfy Definition 1 with $\beta = 0$.

C Missing proofs

C.1 Proofs of examples satisfying definition 1

We highlight that we only show that α - β -condition holds for some α, β without giving all possible values of them.

C.1.1 Proof of example 1

Example 1. Let $f, f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f = \frac{1}{2}(f_1 + f_2) \quad \text{with} \quad f_1(x, y) = \frac{(x+y)^2}{(x+y)^2+1}, \quad f_2(x, y) = \frac{(x+y+1)^2}{(x+y+1)^2+1}, \quad (3)$$

then Definition 1 holds with $\alpha \in [5/2, +\infty)$ and $\beta \in [4\alpha/5, \alpha]$.

Proof. Let $(z, t) = \text{Proj}((x, y), \mathcal{S})$. One can show that the set of minimizers of f in this example is $\mathcal{S} = \{(x, y) : y = -x - 1/2\}$. Moreover, $f_1^* = 0$ for $x = -y$. We will show that α - β -condition holds for $i = 1$. For $i = 2$ it can be shown in similar way with change of variables. We have

$$\partial_x f_1(x, y) = \partial_y f_1(x, y) = \frac{2(x+y)}{(1+(x+y)^2)^2}.$$

Therefore, we need to show that

$$\frac{2(x+y)}{(1+(x+y)^2)}(x-z+y-t) \geq (\alpha-\beta) \frac{(x+y)^2}{1+(x+y)^2} - \alpha f_i(z, t).$$

Note that $f_i(z, t) = \frac{(z+t)^2}{1+(z+t)^2} = \frac{(-1/2)^2}{1+(-1/2)^2} = \frac{1}{5}$. Moreover, we can compute the projection operator onto \mathcal{S} . After simple derivations, the projection can be expressed as

$$\text{Proj}((x, y), \mathcal{S}) = \frac{1}{2} \begin{pmatrix} x - y - 1/2 \\ -x + y - 1/2 \end{pmatrix}.$$

Therefore, we need to satisfy

$$\frac{2(x+y)}{(1+(x+y)^2)^2}(x+y+1/2) \geq (\alpha-\beta) \frac{(x+y)^2}{1+(x+y)^2} - \frac{\alpha}{5}.$$

This is equivalent to

$$\begin{aligned} 2(x+y)(x+y+1/2) &\geq (\alpha-\beta)(x+y)^2(1+(x+y)^2) - \frac{\alpha}{5}(1+(x+y)^2)^2 \\ \Leftrightarrow 2(x+y)^2 + \frac{\alpha}{5}(1+2(x+y)^2+(x+y)^4) &\geq (\alpha-\beta)((x+y)^2+(x+y)^4) - (x+y). \end{aligned} \quad (15)$$

We should satisfy the above for all values of $x+y$. First, we need the coefficient next to $(x+y)^4$ in LHS to be larger or equal of the corresponding coefficient in RHS. This gives us $\frac{\alpha}{5} \geq \alpha - \beta$. This shows that β can not be zero. Using $2ab \leq a^2 + b^2$, (15) is satisfied as long as we have

$$2(x+y)^2 + \frac{\alpha}{5}(1+2(x+y)^2+(x+y)^4) \geq (\alpha-\beta)((x+y)^2+(x+y)^4) + \frac{1}{2}(x+y)^2 + \frac{1}{2}.$$

Therefore, we should satisfy

$$\begin{cases} \text{coefficient next to } (x+y)^4 : \frac{\alpha}{5} \geq \alpha - \beta \\ \text{coefficient next to } (x+y)^2 : 2 + \frac{2\alpha}{5} \geq \alpha - \beta + \frac{1}{2} \\ \text{coefficient next to } 1 : \frac{\alpha}{5} \geq \frac{1}{2} \end{cases}$$

Note that if the first inequality holds, then the second one is true as well. Therefore, from the first and third inequalities we get that the solution of the system of inequalities is $\alpha \geq \frac{5}{2}$ and $\beta \in [4\alpha/5, \alpha]$. \square

C.1.2 Proof of example 2

Example 2. Let $f, f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f = \frac{1}{2}(f_1 + f_2) \quad \text{with} \quad f_1(x, y) = 1 - e^{-x^2-y^2}, \quad f_2(x, y) = 1 - e^{-(x-2)^2-(y-2)^2}, \quad (4)$$

then Definition 1 holds for some α and $\beta = \alpha - 8$.

Proof. Let $(z, t) = \text{Proj}((x, y), \mathcal{S})$. Again, we show that α - β -condition holds for f_1 ; for f_2 it can be proved with the same arguments with change of variables. Note that $f_i^* = 0$ and $f^* > 0$. We also observe that \mathcal{S} contains two points in total, and both of them of the form (t, t) where for one t is close to 0 (it is around $t \approx 0.00067$) and for second t is close to 2 (it is around $t \approx 1.99932$). Besides, note that $c := \min_i \min_{(t,t) \in \mathcal{S}} f_i(t, t) \in (0, 1)$ and $c \approx 9 \cdot 10^{-7}$, which means that we are close to interpolation.

We have in this case

$$\partial_x f_1(x, y) = 2x \exp(-x^2 - y^2), \quad \partial_y f_1(x, y) = 2y \exp(-x^2 - y^2).$$

Therefore, we need to satisfy

$$\exp(-x^2 - y^2)(2x(x - t) + 2y(y - t)) \geq (\alpha - \beta)(1 - \exp(-x^2 - y^2)) - \alpha f_1(z, t). \quad (16)$$

Note that if $\beta = 0$ and $x, y \rightarrow +\infty$, then the inequality is obviously can not be satisfied for all $x, y \in \mathbb{R}$. Indeed, assume $\beta = 0$, then since $f_1(z, t) < 1$ and LHS goes to 0 while RHS to $\alpha - \alpha f_1(z, t) > 0$.

Rearranging terms in (16), we should satisfy

$$\exp(-x^2 - y^2)(2x^2 + 2y^2 + \alpha - \beta) + \alpha f_1(z, t) \geq \exp(-x^2 - y^2)(2xt + 2yt) + (\alpha - \beta).$$

Using $2ab \leq a^2 + b^2$, the above is satisfied if the following inequality holds

$$\begin{aligned} & \exp(-x^2 - y^2)(2x^2 + 2y^2 + \alpha - \beta) + \alpha c \geq \exp(-x^2 - y^2)(x^2 + y^2 + 2t^2) + (\alpha - \beta) \\ \Leftrightarrow & \exp(-x^2 - y^2)(x^2 + y^2 + \alpha - \beta) + \alpha c \geq \exp(-x^2 - y^2) \cdot 2t^2 + (\alpha - \beta). \end{aligned}$$

Since $t \in (0, 2)$ we can choose $\alpha - \beta = 8$, so that $x^2 + y^2 - 2t^2 + \alpha - \beta \geq x^2 + y^2 \geq 0$. Finally, we choose $\alpha \geq \frac{8}{c}$, so that $\alpha c \geq \alpha - \beta = 8$. We can estimate that $\alpha \gtrsim 72 \cdot 10^7$. \square

C.1.3 Proof of Example 3

Example 3. Let $f, f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f = \frac{1}{2}(f_1 + f_2) \quad \text{with} \quad f_1(x, y) = \frac{1+x^2+y^2}{4+x^2+y^2}, \quad f_2(x, y) = \frac{(x-2.5)^2+(y-2.5)^2}{4+(y-2.5)^2+(y-2.5)^2}, \quad (6)$$

then Definition 1 holds for some α and $\beta = \alpha - 1$.

Proof. Let $(z, t) = \text{Proj}((x, y), \mathcal{S})$. For this example, we have $f_1^* = \frac{1}{4}$ (achieved at $(0, 0)$), $f_2^* = 0$ (achieved at $(2.5, 2.5)$), and $f^* \approx 0.408$ (achieved at $(2.471, 2.471)$). Besides, we have $f(0, 0) = 0.587963$ and $f(2.5, 2.5) = 0.409091$. Besides, $c := \min_i \min_{(z,t) \in \mathcal{S}} f_i(z, t) = f_2(2.471, 2.471) = 0.00167918$. Let us show that α - β -condition holds for f_1 ; for f_2 it can be shown similarly. We have

$$\partial_x f_1(x, y) = \frac{6x}{(4 + x^2 + y^2)^2}, \quad \partial_y f_1(x, y) = \frac{6y}{(4 + x^2 + y^2)^2}.$$

Therefore, we need to satisfy for some α and β

$$\begin{aligned} & \frac{6}{(4 + x^2 + y^2)^2} (x(x - z) + y(y - t)) \geq (\alpha - \beta) \frac{1 + x^2 + y^2}{4 + x^2 + y^2} - f_1(z, t) \\ \Leftrightarrow & 6x^2 + 6y^2 - 6xz - 6yt \geq (\alpha - \beta)(1 + x^2 + y^2)(4 + x^2 + y^2) - \alpha f_1(z, t)(4 + x^2 + y^2)^2 \\ \Leftrightarrow & 6(x^2 + y^2) + \alpha f_1(z, t)(16 + 8(x^2 + y^2) + (x^2 + y^2)^2) \geq 6xz + 6yt \\ & \quad + (\alpha - \beta)(4 + 5(x^2 + y^2) + (x^2 + y^2)^2). \end{aligned}$$

The above is satisfied for all $x, y \in \mathbb{R}$ and $\alpha - \beta = 1$ if we have

$$\begin{aligned} 6(x^2 + y^2) + \alpha f_1(z, t)(16 + 8(x^2 + y^2) + (x^2 + y^2)^2) & \geq 3(x^2 + y^2) + 3z^2 + 3t^2 \\ & \quad + (4 + 5(x^2 + y^2) + (x^2 + y^2)^2). \end{aligned}$$

To satisfy the above for all $x, y \in \mathbb{R}$ we should have

$$\begin{cases} 16\alpha f_1(z, t) & \geq 4 + 3(z^2 + t^2) \\ 6 + 8\alpha f_1(z, t) & \geq 3 + 5 \\ \alpha f_1(z, t) & \geq 1, \end{cases}$$

where $z = t = 2.471$, and $f_1(z, t) = 0.0016718$. We can satisfy α - β -condition with $\alpha \gtrsim 1512.4586$. \square

C.1.4 Proof of example 6

Example 6. Let $f_1, f_2: \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that

$$f_1(x, y) = \frac{x^2 + y^2}{1 + x^2 + y^2}, \quad f_2(x, y) = \frac{(x-1)^2 + (y-1)^2}{1 + (x-1)^2 + (y-1)^2}, \quad (14)$$

then Definition 1 holds with $\alpha \geq \frac{2}{\min_i \min_{(z,t) \in \mathcal{S}} f_i(z,t)} \approx 41.369325$ and $\beta = \alpha - 1$. Besides, Definition 1 does not hold with $\beta = 0$ not satisfy Definition 1 with $\beta = 0$.

Proof. Let $(z, t) = \text{Proj}((x, y), \mathcal{S})$. We show that α - β -condition holds for f_1 ; for f_2 the proof is similar with change of variables. In this case, $f_i^* = 0$, while $f^* > 0$ ($f^* \approx 0.316988$ at points $(0.159375, 0.159375)$ and $(0.840625, 0.840625)$). Besides, we have $c := \min_i \min_{(z,t) \in \mathcal{S}} f_i(z, t) \approx 0.048345 > 0$. Moreover, We have

$$\partial_x f_1(x, y) = \frac{2x}{(1 + x^2 + y^2)^2}, \quad \partial_y f_1(x, y) = \frac{2y}{(1 + x^2 + y^2)^2}.$$

Therefore, we need to satisfy for some α and β

$$\begin{aligned} & \frac{1}{(1 + x^2 + y^2)^2} (2x(x-z) + 2y(y-t)) \geq (\alpha - \beta) \frac{x^2 + y^2}{1 + x^2 + y^2} - \alpha f_1(z, t) \\ \Leftrightarrow & 2x^2 + 2y^2 - 2xz - 2yt \geq (\alpha - \beta)(x^2 + y^2)(1 + x^2 + y^2) - \alpha f_1(z, t)(1 + x^2 + y^2)^2 \\ \Leftrightarrow & 2x^2 + 2y^2 + \alpha f_1(z, t)(1 + 2(x^2 + y^2)) + (x^2 + y^2)^2 \geq 2xz + 2yt + \\ & (\alpha - \beta)(x^2 + y^2)(1 + x^2 + y^2). \end{aligned}$$

Note that if we take $\beta = 0$ and $x, y \rightarrow +\infty$, then LHS grows as $\alpha f_1(z, t)(x^2 + y^2)^2$ and RHS grows as $\alpha(x^2 + y^2)^2$, therefore in the limit the RHS is larger. This implies that $\beta \neq 0$. Using $2ab \leq a^2 + b^2$ the above is satisfied if we have

$$(2 + 2\alpha f_1(z, t))(x^2 + y^2) + \alpha f_1(z, t)(x^2 + y^2)^2 + \alpha f_1(z, t) \geq (x^2 + y^2)(1 + \alpha - \beta) + z^2 + t^2 + (\alpha - \beta)(x^2 + y^2)^2. \quad (17)$$

Let us take $\alpha - \beta = 1$ and $\alpha \geq \frac{2}{\min_i \min_{(z,t) \in \mathcal{S}} f_i(z,t)} \approx 41.369325$. With this choice, we have

$$\begin{aligned} z^2 + t^2 & \leq 2 \leq \alpha f_1(z, t), \\ 1 + \alpha - \beta & = 2 \leq 2 + 2\alpha f_1(z, t), \\ \alpha - \beta & = 1 < 2 \leq \alpha f_1(z, t). \end{aligned}$$

Therefore, LHS is always larger than RHS in (17). \square

C.1.5 Proof of example 4

Example 4. Let f_i, f_{ij} be such that

$$f(W, S) = \frac{1}{2nm} \|X - W^\top S\|_F^2 = \frac{1}{2nm} \sum_{i,j} (X_{ij} - w_i^\top s_j)^2, \quad f_{ij}(W, S) = \frac{1}{2} (X_{ij} - w_i^\top s_j)^2, \quad (7)$$

where $X \in \mathbb{R}^{n \times m}$, $W = (w_i)_{i=1}^n \in \mathbb{R}^{k \times n}$, $S = (s_j)_{j=1}^m \in \mathbb{R}^{k \times m}$, and $\text{rank}(X) = r \geq k$. We assume that X is generated using matrices W^* and S^* with non-zero additive noise that minimize empirical loss, namely, $X = (W^*)^\top S^* + (\varepsilon_{ij})_{i \in [n], j \in [m]}$ where $W^*, S^* = \text{argmin}_{W,S} f(W, S)$. Let \mathcal{X} be any bounded set that contains \mathcal{S} . Then Definition 1 is satisfied with $\alpha = \beta + 1$ and some $\beta > 0$.

Proof. Since $k \leq r$, then the matrix factorization problem has a unique solution which can be found from SVD decomposition of X [76].

We have

$$\nabla_{w_i} f_{ij}(W, S) = (w_i^\top s_j - X_{ij})s_j, \quad \nabla_{s_j} f_{ij}(W, S) = (w_i^\top s_j - X_{ij})w_i.$$

Therefore, we need to show that

$$\begin{aligned} \langle (w_i^\top s_j - X_{ij})s_j, w_i - w_i^* \rangle + \langle (w_i^\top s_j - X_{ij})w_i, s_j - s_j^* \rangle &\geq \frac{\alpha - \beta}{2}(X_{ij} - w_i^\top s_j)^2 \\ &\quad - \frac{\alpha}{2}(X_{ij} - (w_i^*)^\top s_j^*)^2, \end{aligned}$$

since $f_{ij}^* = 0$. Since f_{ij} is convex w.r.t. w_i , the above holds if we have

$$\begin{aligned} \frac{1}{2}(w_i^\top s_j - X_{ij})^2 - \frac{1}{2}(s_j^\top w_i^* - X_{ij})^2 + \frac{1}{2}(w_i^\top s_j - X_{ij})^2 - \frac{1}{2}(w_i^\top s_j^* - X_{ij})^2 &\geq \\ \frac{\alpha - \beta}{2}(X_{ij} - w_i^\top s_j)^2 - \frac{\alpha}{2}(X_{ij} - (w_i^*)^\top s_j^*)^2. \end{aligned}$$

Let us take $\alpha = \beta + 1$, then we can simplify the above as follows

$$\frac{1}{2}(w_i^\top s_j - X_{ij})^2 + \frac{\alpha}{2}(X_{ij} - (w_i^*)^\top s_j^*)^2 \geq \frac{1}{2}(s_j^\top w_i^* - X_{ij})^2 + \frac{1}{2}(w_i^\top s_j^* - X_{ij})^2 \quad (18)$$

Since we consider \mathcal{X} to be bounded, then there exist $r \geq 0$ such that $\|s_j\|, \|w_i\| \leq r$ for all i and j . Therefore, the RHS in (18) is bounded by some constant C . From the data generation, we have that $c = \min_{ij}(X_{ij} - (w_i^*)^\top s_j^*)^2 = \min_{ij} \varepsilon_{ij}^2 > 0$. Therefore, we can take $\alpha \geq \frac{2C}{c}$ to verify (18).

□

C.1.6 Proof of example 5

Example 5. Consider training a two-layer neural network with a logistic loss

$$f = \frac{1}{n} \sum_{i=1}^n f_i, \quad f_i(W, v) = \phi(y_i \cdot v^\top \sigma(Wx_i)) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 \quad (8)$$

for a classification problem where $\phi(t) := \log(1 + \exp(-t))$, $W \in \mathbb{R}^{k \times d}$, $v \in \mathbb{R}^k$, σ is a ReLU function applied coordinate-wise, $y_i \in \{-1, +1\}$ is a label and $x_i \in \mathbb{R}^d$ is a feature vector. Let \mathcal{X} be any bounded set that contains \mathcal{S} . Then the α - β -condition holds in \mathcal{X} for some $\alpha \geq 1$ and $\beta = \alpha - 1$.

Proof. Let $(Z, z) := \text{Proj}((W, v), \mathcal{S})$. We have the following derivations for gradients

$$\begin{aligned} \mathbb{R}^{k \times d} \ni \nabla_W f_i(W, v) &= \phi'(y_i v^\top \sigma(Wx_i)) \cdot y_i (v \circ \mathbb{1}_{Wx_i \geq 0}) x_i^\top + 2\lambda_2 W, \\ \mathbb{R}^k \ni \nabla_v f_i(W, v) &= \phi'(y_i v^\top \sigma(Wx_i)) \cdot y_i \sigma(Wx_i) = \phi'(y_i v^\top \sigma(Wx_i)) \cdot y_i (Wx_i) \circ \mathbb{1}_{Wx_i \geq 0} \\ &\quad + 2\lambda_1 v \\ &= \phi'(y_i v^\top \sigma(Wx_i)) \cdot y_i (Wx_i) \circ e_w + 2\lambda_1 v, \end{aligned}$$

where we denote $e_w := \mathbb{1}_{Wx_i \geq 0} \in \mathbb{R}^k$. Besides, we denote $e_z := \mathbb{1}_{Zx_i \geq 0} \in \mathbb{R}^k$. Note that the optimal value of $f_i^* > 0$ because of the L2 regularization.

Note that we have the following relations

$$v^\top \sigma(Wx_i) = \langle v, (Wx_i) \circ \mathbb{1}_{Wx_i \geq 0} \rangle = \langle v \circ \mathbb{1}_{Wx_i \geq 0}, Wx_i \rangle = \langle v \circ e_w, Wx_i \rangle. \quad (19)$$

Note that \mathcal{S} is bounded because of the L2 regularization. Since we assume that the interpolation does not hold, then f is always strictly larger than f_i^* in \mathcal{X} , and due to continuity there exists $c := \min_{i \in [n]} \min_{(Z, z) \in \mathcal{S}} f_i(Z, z) > 0$.

We need to show that there exist some α and β such that

$$\begin{aligned} &\langle \nabla_W f_i(W, v), W - Z \rangle_F + \langle \nabla_v f_i(W, v), v - z \rangle \\ &\geq \alpha(f_i(W, v) - f_i(Z, z)) - \beta(f_i(W, v) - f_i^*) \\ \Leftrightarrow &\phi'(y_i(v \circ e_w)^\top Wx_i) y_i (\langle (v \circ e_w) x_i^\top, W - Z \rangle + \langle (Wx_i) \circ e_w, v - z \rangle) + 2\lambda_1 \langle v, v - z \rangle \\ &\quad + 2\lambda_2 \langle W, W - Z \rangle \\ \geq &\alpha [\phi(y_i(v \circ e_w)^\top Wx_i) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 - \phi(y_i(z \circ e_z)^\top Zx_i) - \lambda_1 \|z\|^2 - \lambda_2 \|Z\|_F^2] \\ &\quad - \beta [\phi(y_i(v \circ e_w)^\top Wx_i) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 - f_i^*] \\ \Leftrightarrow &\phi'(y_i(v \circ e_w)^\top Wx_i) ([y_i(v \circ e_w)^\top Wx_i - y_i(v \circ e_w)^\top Zx_i] + [y_i(v \circ e_w)^\top Wx_i - y_i(z \circ e_z)^\top Wx_i]) \\ &\quad + 2\lambda_1 \langle v, v - z \rangle + 2\lambda_2 \langle W, W - Z \rangle \\ \geq &\alpha [\phi(y_i(v \circ e_w)^\top Wx_i) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 - \phi(y_i(z \circ e_z)^\top Zx_i) - \lambda_1 \|z\|^2 - \lambda_2 \|Z\|_F^2] \\ &\quad - \beta [\phi(y_i(v \circ e_w)^\top Wx_i) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 - f_i^*], \end{aligned} \quad (20)$$

Algorithm 1 SGD with constant stepsize

- 1: **Input:** Stepsize γ
- 2: **for** $k = 0, 1, 2, \dots, K - 1$ **do**
- 3: Sample $i_k \sim \text{Unif}[n]$ and compute $\nabla f_{i_k}(x^k)$
- 4: Update model

$$x^{k+1} = x^k - \gamma \nabla f_{i_k}(x^k)$$

- 5: **end for**

where we use (19). Since ϕ is convex, then we have $\phi'(x)(x - y) \geq \phi(x) - \phi(y)$. Therefore, using (19) again, we get that (20) is satisfied if we have

$$\begin{aligned} & \phi(y_i(v \circ e_w)^\top W x_i) - \phi(y_i(v \circ e_w)^\top Z x_i) - \phi(y_i(z \circ e_w)^\top W x_i) + \lambda_1 \langle v, v - z \rangle + \lambda_2 \langle W, W - Z \rangle \\ & \geq \alpha [\phi(y_i(v \circ e_w)^\top W x_i) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 - \phi(y_i(z \circ e_z)^\top Z x_i) - \lambda_1 \|z\|^2 - \lambda_2 \|Z\|_F^2] \\ & \quad - \beta [\phi(y_i(v \circ e_w)^\top W x_i) + \lambda_1 \|v\|^2 + \lambda_2 \|W\|_F^2 - f_i^*]. \end{aligned}$$

Now we take $\alpha = \beta + 1$ and simplify the above as follows

$$\begin{aligned} & \phi(y_i(v \circ e_w)^\top W x_i) + \alpha \phi(y_i(z \circ e_w)^\top Z x_i) + \lambda_1 (2 \langle v, v - z \rangle - \|v\|^2 - \|z\|^2) \\ & \quad + \lambda_2 (2 \langle W, W - Z \rangle - \|W\|_F^2 - \|Z\|_F^2) + (\alpha - 1) (\lambda_1 \|z\|^2 + \lambda_2 \|Z\|_F^2) \\ & \geq \phi(y_i(v \circ e_w)^\top Z x_i) + \phi(y_i(z \circ e_w)^\top W x_i) + (\alpha - 1) f_i^*. \end{aligned} \quad (21)$$

The above is satisfied if we have

$$\begin{aligned} & \phi(y_i(v \circ e_w)^\top W x_i) + \alpha c + \lambda_1 \|v - z\|^2 + \lambda_2 \|W - Z\|_F^2 + (\alpha - 1) (\|z\|^2 + \|Z\|_F^2) \\ & \geq \phi(y_i(v \circ e_w)^\top Z x_i) + \phi(y_i(z \circ e_w)^\top W x_i) + (\alpha - 1) f_i^*, \end{aligned} \quad (22)$$

because we also have $\phi(y_i(z \circ e_z)^\top Z x_i) \geq c$ for all i and $(Z, z) \in \mathcal{S}$. Since $(W, v) \in \mathcal{X}$, there exist constants $R, r \geq 0$ such that $\|v\| \leq r$ and $\|W\| \leq R$, and RHS in (22) is bounded by

$$2 \max_{i \in [n]} \log(1 + \exp(Rr\|x_i\|)) \geq \phi(y_i(v \circ e_w)^\top Z x_i) + \phi(y_i(z \circ e_w)^\top W x_i).$$

Therefore, we can take $\alpha \geq \left\{ \frac{2 \max_{i \in [n]} \log(1 + \exp(Rr\|x_i\|))}{c - f_i^*}, 1 \right\}$ and $\beta = \alpha - 1$. \square

Remark 3. The result suggest that the choice $\beta = 0$ is possible only if constants r and R are small, i.e. locally around \mathcal{S} only. In order to satisfy α - β -condition in larger set \mathcal{X} , one needs to choose $\beta > 0$.

C.2 Convergence of optimization algorithms under α - β -condition

C.2.1 Convergence of SGD

Constant stepsize. In this section, we present the proof of convergence of SGD with constant stepsize under α - β -condition for completeness of the presentation.

Theorem 1. Assume that Assumptions 1-2 hold. Then the iterates of SGD (Alg. 1) with stepsize $\gamma \leq \frac{\alpha - \beta}{2L}$ satisfy

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{K} \frac{1}{\gamma(\alpha - \beta)} + \frac{2L\gamma}{\alpha - \beta} \sigma_{\text{int}}^2 + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2. \quad (9)$$

Proof. Let $x_p^k \in \text{Proj}(x^k, \mathcal{S})$ satisfies Definition 1. Using smoothness we have

$$\begin{aligned} \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] & \leq \mathbb{E}_k [\|x^{k+1} - x_p^k\|^2] \\ & = \text{dist}(x^k, \mathcal{S})^2 - 2\gamma \mathbb{E}_k [\langle \nabla f_{i_k}(x^k), x^k - x_p^k \rangle] + \gamma^2 \mathbb{E}_k [\|\nabla f_{i_k}(x^k)\|^2] \\ & \stackrel{(i)}{\leq} \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\gamma \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] + 2\beta\gamma \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}^*] \\ & \quad + 2L\gamma^2 \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}^*] \\ & = \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\gamma \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\ & \quad + 2\gamma(\beta + L\gamma) \mathbb{E}_k [(f_{i_k}(x^k) - f_{i_k}(x_p^k)) + (f_{i_k}(x_p^k) - f_{i_k}^*)], \end{aligned}$$

where (i) holds because of the α - β -condition and smoothness. Now we need to choose a stepsize $\gamma \leq \frac{\alpha-\beta}{2L}$ to get

$$\begin{aligned}\mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - \gamma(\alpha - \beta)\mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\ &\quad + 2\gamma(\beta + L\gamma)\mathbb{E}_k [f_{i_k}(x_p^k) - f_{i_k}^*].\end{aligned}$$

Taking full expectation, noting that x_p^k is independent of the randomness of i_k , and performing simple derivations, we get

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{K} \frac{1}{\gamma(\alpha - \beta)} + \frac{2L\gamma}{\alpha - \beta} \sigma_{\text{int}}^2 + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2. \quad (23)$$

□

Decreasing stepsize. Now we present the results with decreasing stepsize.

Theorem 4. *Assume that Assumptions 1-2 hold. Then the iterates of SGD with decreasing stepsize $\gamma_k = \frac{\gamma_0}{\sqrt{k+1}}$ where $\gamma_0 \leq \frac{\alpha-\beta}{2L}$ satisfy*

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{5\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{4(\alpha - \beta)\gamma_0\sqrt{K}} + \frac{5\gamma_0 L\sigma_{\text{int}}^2}{\alpha - \beta} \frac{\log(K+1)}{\sqrt{K}} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2 \quad (24)$$

$$= \tilde{\mathcal{O}} \left(\frac{1}{\sqrt{K}} + \beta \sigma_{\text{int}}^2 \right). \quad (25)$$

Proof. Similarly to the proof of constant stepsize SGD we can obtain

$$\begin{aligned}\mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - \gamma_k(\alpha - \beta)\mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\ &\quad + 2\gamma_k(\beta + L\gamma_k)\mathbb{E}_k [f_{i_k}(x_p^k) - f_{i_k}^*],\end{aligned}$$

since $\gamma_k \leq \gamma_0 \leq \frac{\alpha-\beta}{2L}$ for all k . Now we follow standard proof techniques [24] for decreasing stepsize. Taking full expectation and dividing both sides by $\alpha - \beta$ we get

$$\gamma_k \mathbb{E} [f(x^k) - f^*] \leq \frac{\mathbb{E} [\text{dist}(x^k, \mathcal{S})^2] - \mathbb{E} [\text{dist}(x^{k+1}, \mathcal{S})^2]}{\alpha - \beta} + \frac{2\beta}{\alpha - \beta} \gamma_k \sigma_{\text{int}}^2 + \frac{2L}{\alpha - \beta} \gamma_k^2 \sigma_{\text{int}}^2.$$

Summing the above from iteration 0 to $K - 1$ leads to

$$\sum_{k=0}^{K-1} \gamma_k \mathbb{E} [f(x^k) - f^*] \leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{\alpha - \beta} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2 \sum_{k=0}^{K-1} \gamma_k + \frac{2L}{\alpha - \beta} \sigma_{\text{int}}^2 \sum_{k=0}^{K-1} \gamma_k^2.$$

Therefore, we get

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{(\alpha - \beta) \sum_{k=0}^{K-1} \gamma_k} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2 + \frac{2L}{\alpha - \beta} \sigma_{\text{int}}^2 \frac{\sum_{k=0}^{K-1} \gamma_k^2}{\sum_{k=0}^{K-1} \gamma_k}.$$

Using the results of Theorem 5.7 [24] we get

$$\sum_{k=0}^{K-1} \gamma_k^2 \leq 2\gamma_0^2 \log(K+1), \quad \sum_{k=0}^{K-1} \gamma_k \geq \frac{4\gamma_0}{5} \sqrt{K}.$$

Therefore, the final rate we get is

$$\begin{aligned}\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{(\alpha - \beta) \frac{4}{5} \gamma_0 \sqrt{K}} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2 + \frac{2L}{\alpha - \beta} \sigma_{\text{int}}^2 \frac{2\gamma_0^2 \log(K+1)}{\frac{4\gamma_0}{5} \sqrt{K}} \\ &= \frac{5\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{4(\alpha - \beta) \gamma_0 \sqrt{K}} + \frac{5\gamma_0 L\sigma_{\text{int}}^2}{\alpha - \beta} \frac{\log(K+1)}{\sqrt{K}} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2.\end{aligned}$$

□

Algorithm 2 SPS_{\max} : Stochastic Polyak Stepsize

- 1: **Input:** Stepsize parameter c and stepsize upper bound γ_b
- 2: **for** $k = 0, 1, 2, \dots, K - 1$ **do**
- 3: Sample $i_k \sim \text{Unif}[n]$ and compute $\nabla f_{i_k}(x^k)$
- 4: Compute Polyak stepsize

$$\gamma_k := \min \left\{ \frac{f_{i_k}(x^k) - f_{i_k}^*}{c \|\nabla f_{i_k}(x^k)\|^2}, \gamma_b \right\}$$

- 5: Update model

$$x^{k+1} = x^k - \gamma_k \nabla f_{i_k}(x^k)$$

- 6: **end for**

C.2.2 Convergence of SGD with Polyak Stepsize

Constant stepsize parameter. In this section, we present the proof of convergence of SGD with Polyak stepsize (with constant stepsize parameters) under α - β -condition for completeness of the presentation.

Lemma 1 (Lemma from [53]). The SPS_{\max} stepsize satisfy

$$\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2 \leq \frac{\gamma_k}{c} (f_{i_k}(x^k) - f_{i_k}^*). \quad (26)$$

Lemma 2 (Lemma from [53]). Assume each f_i is L -smooth, then SPS_{\max} stepsize satisfy

$$\gamma_{\min} := \min \left\{ \frac{1}{2cL}, \gamma_b \right\} \leq \gamma_k \leq \gamma_b. \quad (27)$$

Now we present the proof of Theorem 2 with constant stepsize parameters.

Theorem 2. Assume that Assumptions 1-2 hold. Then the iterates of SPS_{\max} (Alg. 2) with a stepsize hyperparameter $c > \frac{1}{2(\alpha-\beta)}$ satisfy

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{c_1}{K} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] + 2\alpha c_1 \gamma_b \sigma_{\text{int}}^2, \quad (10)$$

where $\gamma_{\min} := \min\{1/2cL, \gamma_b\}$ and $c_1 := \frac{c}{\gamma_{\min}(2(\alpha-\beta)c-1)}$.

Proof. Let $x_p^k \in \text{Proj}(x^k, \mathcal{S})$ that satisfies Definition 1, then we have

$$\begin{aligned} \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \mathbb{E}_k [\|x^{k+1} - x_p^k\|^2] \\ &= \|x^k - x_p^k\|^2 - 2\mathbb{E}_k [\gamma_k \langle \nabla f_{i_k}(x^k), x^k - x_p^k \rangle] + \mathbb{E}_k [\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2] \\ &\stackrel{(i)}{\leq} \text{dist}(x^k, \mathcal{S})^2 - 2\alpha \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}(x_p^k))] \\ &\quad + 2\beta \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] + \frac{1}{c} \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] \\ &= \text{dist}(x^k, \mathcal{S})^2 - 2\alpha \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*) - (f_{i_k}(x_p^k) - f_{i_k}^*)] \\ &\quad + 2\beta \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] + \frac{1}{c} \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] \\ &= \text{dist}(x^k, \mathcal{S})^2 - \mathbb{E}_k \left[\gamma_k \left(2\alpha - 2\beta - \frac{1}{c} \right) (f_{i_k}(x^k) - f_{i_k}^*) \right] \\ &\quad + 2\alpha \mathbb{E}_k [\gamma_k (f_{i_k}(x_p^k) - f_{i_k}^*)], \end{aligned}$$

where (i) follows from Lemma 1 and α - β -condition. Now, since $c > \frac{1}{2(\alpha-\beta)}$, we get that $2\alpha - 2\beta - 1/c > 0$. Therefore, using Lemma 2 and the fact that $f_{i_k}^* \leq f_{i_k}(x_p^k)$ and $f_{i_k}^* \leq f_{i_k}(x^k)$ we get

$$\begin{aligned} \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - \gamma_{\min} \left(2\alpha - 2\beta - \frac{1}{c} \right) \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}^*] \\ &\quad + 2\alpha\gamma_b \mathbb{E}_k [(f_{i_k}(x_p^k) - f_{i_k}^*)] \\ &= \text{dist}(x^k, \mathcal{S})^2 - \gamma_{\min} \left(2\alpha - 2\beta - \frac{1}{c} \right) \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\ &\quad - \gamma_{\min} \left(2\alpha - 2\beta - \frac{1}{c} \right) \mathbb{E}_k [f_{i_k}(x_p^k) - f_{i_k}^*] + 2\alpha\gamma_b \mathbb{E}_k [(f_{i_k}(x_p^k) - f_{i_k}^*)] \\ &\leq \text{dist}(x^k, \mathcal{S})^2 - \gamma_{\min} \left(2\alpha - 2\beta - \frac{1}{c} \right) \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\ &\quad + 2\alpha\gamma_b \mathbb{E}_k [(f_{i_k}(x_p^k) - f_{i_k}^*)]. \end{aligned}$$

Therefore, noticing that $\mathbb{E}_k [f_{i_k}(x_p^k)] = f^*$ since x_p^k is independent of i_k , we have

$$\gamma_{\min} \left(2\alpha - 2\beta - \frac{1}{c} \right) \mathbb{E}_k [f(x^k) - f^*] \leq \text{dist}(x^k, \mathcal{S})^2 - \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] + 2\alpha\gamma_b\sigma_{\text{int}}^2.$$

Dividing both sides by $\gamma_{\min}(2\alpha - 2\beta - 1/c)$ and taking full expectation, we get

$$\begin{aligned} \mathbb{E} [f(x^k) - f^*] &\leq \frac{c}{\gamma_{\min}(2(\alpha-\beta)c-1)} (\mathbb{E} [\text{dist}(x^k, \mathcal{S})^2] - \mathbb{E} [\text{dist}(x^{k+1}, \mathcal{S})^2]) \\ &\quad + \frac{2\alpha c \gamma_b}{\gamma_{\min}(2(\alpha-\beta)c-1)} \sigma_{\text{int}}^2. \end{aligned}$$

Averaging for $k \in \{0, \dots, K-1\}$ we get

$$\begin{aligned} \min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{c}{\gamma_{\min}(2(\alpha-\beta)c-1)K} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] \\ &\quad + \frac{2\alpha c \gamma_b}{\gamma_{\min}(2(\alpha-\beta)c-1)} \sigma_{\text{int}}^2, \end{aligned}$$

that finalizes the proof. \square

Decreasing stepsize parameter. Now we switch to the analysis of SPS_{\max} with decreasing stepsize parameters. We consider the stepsize DecSPS introduced in [64]

$$\gamma_k = \frac{1}{c_k} \min \left\{ \frac{f_{i_k}(x^k) - f_{i_k}^*}{\|\nabla f_{i_k}(x^k)\|^2}, c_{k-1} \gamma_{k-1} \right\}$$

with $c_{-1} = c_0$ and $\gamma_{-1} = \gamma_b > 0$ to get

$$\gamma_0 = \frac{1}{c_0} \min \left\{ \frac{f_{i_0}(x^0) - f_{i_0}^*}{\|\nabla f_{i_0}(x^0)\|^2}, c_0 \gamma_b \right\}.$$

Lemma 3 (Lemma 1 from [64]). Let Assumption 1 holds. Let $\{c_k\}_{k \geq 0}$ be any non-decreasing positive sequence of real numbers. Then we have

$$\min \left\{ \frac{1}{2c_k L}, \frac{c_0 \gamma_b}{c_k} \right\} \leq \gamma_k \leq \frac{c_0 \gamma_b}{c_k}, \quad \text{and} \quad \gamma_k \leq \gamma_{k-1}. \quad (28)$$

Theorem 5. Assume that Assumptions 1-2 hold. Let $\{c_k\}_{k \geq 0}$ be any positive non-decreasing sequence such that $c_k \geq \frac{1}{\alpha-\beta}$. Let Then iterates of DecSPS satisfy

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{2\tilde{L}D^2c_{K-1}}{K(\alpha-\beta)} + \frac{2\beta}{\alpha-\beta} \sigma_{\text{int}}^2 + \frac{1}{K} \sum_{k=0}^{K-1} \frac{\sigma_{\text{int}}^2}{(\alpha-\beta)c_k}, \quad (29)$$

where $D^2 := \max_{0 \leq k \leq K} \text{dist}(x^k, \mathcal{S})^2$, $\tilde{L} := \max \left\{ L, \frac{1}{2c_0 \gamma_b} \right\}$.

Proof. From the definition of the stepsize we have

$$\gamma_k \leq \frac{1}{c_k} \frac{f_{i_k}(x^k) - f_{i_k}^*}{\|\nabla f_{i_k}(x^k)\|^2}.$$

Therefore, we have

$$\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2 \leq \frac{\gamma_k}{c_k} (f_{i_k}(x^k) - f_{i_k}^*).$$

Let $x_p^k = \text{Proj}(x^k, \mathcal{S})$ which satisfies Definition 1. Now we have

$$\begin{aligned} \text{dist}(x^{k+1}, \mathcal{S})^2 &\leq \|x^{k+1} - x_p^k\|^2 \\ &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\gamma_k \langle \nabla f_{i_k}(x^k), x^k - x_p^k \rangle + \frac{\gamma_k}{c_k} (f_{i_k}(x^k) - f_{i_k}^*). \end{aligned}$$

Using α - β -condition we get

$$\begin{aligned} \text{dist}(x^{k+1}, \mathcal{S})^2 &\leq \|x^{k+1} - x_p^k\|^2 \\ &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\gamma_k (f_{i_k}(x^k) - f_{i_k}(x_p^k)) + 2\beta\gamma_k (f_{i_k}(x^k) - f_{i_k}^*) \\ &\quad + \frac{\gamma_k}{c_k} (f_{i_k}(x^k) - f_{i_k}^*) \\ &= \text{dist}(x^k, \mathcal{S})^2 - \gamma_k (2\alpha - 2\beta - 1/c_k) (f_{i_k}(x^k) - f_{i_k}(x_p^k)) \\ &\quad + \gamma_k (2\beta + 1/c_k) (f_{i_k}(x_p^k) - f_{i_k}^*). \end{aligned}$$

Let us divide both sides by $\gamma_k > 0$

$$\begin{aligned} \frac{\text{dist}(x^{k+1}, \mathcal{S})^2}{\gamma_k} &\leq \frac{\text{dist}(x^k, \mathcal{S})^2}{\gamma_k} - (2\alpha - 2\beta - 1/c_k) (f_{i_k}(x^k) - f_{i_k}(x_p^k)) \\ &\quad + (2\beta + 1/c_k) (f_{i_k}(x_p^k) - f_{i_k}^*). \end{aligned}$$

Since by hypothesis $c_k \geq \frac{1}{\alpha - \beta}$, we get $2\alpha - 2\beta - 1/c_k \geq \alpha - \beta$. Therefore, we get

$$f_{i_k}(x^k) - f_{i_k}(x_p^k) \leq \frac{\text{dist}(x^k, \mathcal{S})^2}{(\alpha - \beta)\gamma_k} - \frac{\text{dist}(x^{k+1}, \mathcal{S})^2}{(\alpha - \beta)\gamma_k} + \frac{(2\beta + 1/c_k)}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*).$$

Summing from $k = 0$ to $K - 1$ we get

$$\begin{aligned} \sum_{k=0}^{K-1} f_{i_k}(x^k) - f_{i_k}(x_p^k) &\leq \sum_{k=0}^{K-1} \frac{\text{dist}(x^k, \mathcal{S})^2}{(\alpha - \beta)\gamma_k} - \sum_{k=0}^{K-1} \frac{\text{dist}(x^{k+1}, \mathcal{S})^2}{(\alpha - \beta)\gamma_k} \\ &\quad + \sum_{k=0}^{K-1} \frac{2\beta}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*) + \sum_{k=0}^{K-1} \frac{1}{(\alpha - \beta)c_k} (f_{i_k}(x_p^k) - f_{i_k}^*) \\ &\leq \frac{\text{dist}(x^0, \mathcal{S})^2}{(\alpha - \beta)\gamma_0} + \sum_{k=1}^{K-1} \frac{\text{dist}(x^k, \mathcal{S})^2}{(\alpha - \beta)\gamma_k} - \sum_{k=0}^{K-2} \frac{\text{dist}(x^{k+1}, \mathcal{S})^2}{(\alpha - \beta)\gamma_k} \\ &\quad - \frac{\text{dist}(x^K, \mathcal{S})^2}{(\alpha - \beta)\gamma_{K-2}} + \sum_{k=0}^{K-1} \frac{2\beta}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*) \\ &\quad + \sum_{k=0}^{K-1} \frac{1}{(\alpha - \beta)c_k} (f_{i_k}(x_p^k) - f_{i_k}^*). \end{aligned}$$

Here we use the fact that γ_k is a non-increasing sequence with k . We continue as follows

$$\begin{aligned}
\sum_{k=0}^{K-1} f_{i_k}(x^k) - f_{i_k}(x_p^k) &\leq \frac{\text{dist}(x^0, \mathcal{S})^2}{(\alpha - \beta)\gamma_0} + \sum_{k=0}^{K-2} \left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k} \right) \frac{\text{dist}(x^{k+1}, \mathcal{S})^2}{\alpha - \beta} \\
&\quad + \sum_{k=0}^{K-1} \frac{2\beta}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*) + \sum_{k=0}^{K-1} \frac{1}{(\alpha - \beta)c_k} (f_{i_k}(x_p^k) - f_{i_k}^*) \\
&\leq \frac{D^2}{\alpha - \beta} \left(\frac{1}{\gamma_0} + \sum_{k=0}^{K-2} \left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k} \right) \right) \\
&\quad + \sum_{k=0}^{K-1} \frac{2\beta}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*) + \sum_{k=0}^{K-1} \frac{1}{(\alpha - \beta)c_k} (f_{i_k}(x_p^k) - f_{i_k}^*) \\
&\leq \frac{D^2}{(\alpha - \beta)\gamma_{K-1}} + \sum_{k=0}^{K-1} \frac{2\beta}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*) \\
&\quad + \sum_{k=0}^{K-1} \frac{1}{(\alpha - \beta)c_k} (f_{i_k}(x_p^k) - f_{i_k}^*).
\end{aligned}$$

Here we use the fact that $1/\gamma_{k+1} - 1/\gamma_k \geq 0$. From Lemma 3 we have

$$\frac{1}{\gamma_k} \leq c_k \underbrace{\max \left\{ 2L, \frac{1}{c_0 \gamma_b} \right\}}_{:= \tilde{L}}.$$

Therefore, we continue as follows

$$\begin{aligned}
\frac{1}{K} \sum_{k=0}^{K-1} f_{i_k}(x^k) - f_{i_k}(x_p^k) &\leq \frac{2\tilde{L}D^2c_{K-1}}{K(\alpha - \beta)} + \frac{1}{K} \sum_{k=0}^{K-1} \frac{2\beta}{\alpha - \beta} (f_{i_k}(x_p^k) - f_{i_k}^*) \\
&\quad + \frac{1}{K} \sum_{k=0}^{K-1} \frac{1}{(\alpha - \beta)c_k} (f_{i_k}(x_p^k) - f_{i_k}^*).
\end{aligned}$$

Taking expectation we get

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{2\tilde{L}D^2c_{K-1}}{K(\alpha - \beta)} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2 + \frac{1}{K} \sum_{k=0}^{K-1} \frac{\sigma_{\text{int}}^2}{(\alpha - \beta)c_k}.$$

□

Corollary 1. Let $c_k = \sqrt{k+1}$ with $c_{-1} = c_0$, then iterates of **DecSPS** satisfy

$$\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] \leq \frac{2\tilde{L}D^2 + 2\sigma_{\text{int}}^2}{\sqrt{K}(\alpha - \beta)} + \frac{2\beta}{\alpha - \beta} \sigma_{\text{int}}^2 \tag{30}$$

$$= \tilde{\mathcal{O}} \left(\frac{1}{\sqrt{K}} + \beta \sigma_{\text{int}}^2 \right). \tag{31}$$

Proof. The proof directly follows from Theorem 4 using the choice $c_k = \sqrt{k+1}$ and the fact that $\sum_{k=0}^{K-1} k^{-1/2} \leq 2\sqrt{K}$. □

Remark 4. It turned out that removing the bounded iterates assumption for **DecSPS** is a challenging task. Nevertheless, we believe that this is rather the technicalities of Polyak stepsize, but not of our assumption. Moreover, we highlight that [64] removed bounded iterates assumption in restrictive strongly convex setting.

Algorithm 3 NGN: Non-negative Gauss Newton

- 1: **Input:** Stepsize parameter γ
- 2: **for** $k = 0, 1, 2, \dots, K - 1$ **do**
- 3: Sample $i_k \sim \text{Unif}[n]$ and compute $\nabla f_{i_k}(x^k)$
- 4: Compute NGN stepsize

$$\gamma_k := \frac{\gamma}{1 + \frac{\gamma}{2f_{i_k}(x^k)} \|\nabla f_{i_k}(x^k)\|^2}$$

- 5: Update model

$$x^{k+1} = x^k - \gamma_k \nabla f_{i_k}(x^k)$$

- 6: **end for**

C.2.3 Convergence of NGN

Constant stepsize parameter. In this section we present the proof of convergence of NGN with constant stepsize parameter γ under α - β -condition for completeness of the presentation.

Lemma 4 (Lemma from [63]). Let f_i be L -smooth for all i , then the stepsize of NGN satisfies

$$\gamma_k \in \left[\frac{\gamma}{1 + \gamma L}, \gamma \right]. \quad (32)$$

Lemma 5 (Lemma from [63]). Let f_i be L -smooth for all i , then the iterates of NGN satisfy

$$\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2 \leq \frac{4\gamma L}{1 + 2\gamma L} \gamma_k (f_{i_k}(x^k) - f_{i_k}^*) + \frac{2\gamma^2 L}{1 + \gamma L} \max \left\{ \frac{2\gamma L - 1}{2\gamma L + 1}, 0 \right\} f_{i_k}^*. \quad (33)$$

Theorem 3. Assume that Assumptions 1 with $\alpha \geq \beta + 1$ and 1-2 hold. Assume that each function f_i is positive and $\sigma_{\text{pos}}^2 < \infty$. Then the iterates of NGN (Alg. 3) with a stepsize parameter $\gamma > 0$ satisfy

$$\begin{aligned} \min_{0 \leq k \leq K-1} \mathbb{E} [f(x^k) - f^*] &\leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{2\gamma K} \frac{(1 + 2\gamma L)^2}{c_2} + \frac{3L\gamma\alpha(1 + \gamma L)\sigma_{\text{int}}^2}{c_2} \\ &\quad + \frac{\gamma L}{a} \max \{2\gamma L - 1, 0\} \sigma_{\text{pos}}^2 + \frac{2\beta\sigma_{\text{int}}^2}{c_2}, \end{aligned} \quad (11)$$

where $c_2 := 2\gamma L(\alpha - \beta - 1) + \alpha - \beta$.

Proof. Let $x_p^k \in \text{Proj}(x^k, \mathcal{S})$ satisfying Definition 1. Then we have

$$\begin{aligned} \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \mathbb{E}_k [\|x^{k+1} - x_p^k\|^2] \\ &= \|x^k - x_p^k\|^2 - 2\mathbb{E}_k [\gamma_k \langle \nabla f_{i_k}(x^k), x^k - x_p^k \rangle] + \mathbb{E}_k [\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2] \\ &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\alpha \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}(x_p^k))] \\ &\quad + 2\beta \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] + \mathbb{E}_k [\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2]. \end{aligned} \quad (34)$$

From Lemma 5

$$\gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2 \leq \frac{4\gamma L}{1 + 2\gamma L} \gamma_k (f_{i_k}(x^k) - f_{i_k}^*) + \frac{2\gamma^2 L}{1 + \gamma L} \max \left\{ \frac{2\gamma L - 1}{2\gamma L + 1}, 0 \right\} f_{i_k}^*. \quad (35)$$

Plugging (35) in (34) we get

$$\begin{aligned} \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\alpha \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}(x_p^k))] \\ &\quad + 2\beta \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] \frac{4\gamma L}{1 + 2\gamma L} + \mathbb{E}_k [\gamma_k (f_{i_k}(x^k) - f_{i_k}^*)] \\ &\quad + \frac{2\gamma^2 L}{1 + \gamma L} \max \left\{ \frac{2\gamma L - 1}{2\gamma L + 1}, 0 \right\} \mathbb{E}_k [f_{i_k}^*]. \end{aligned}$$

We have $f_{i_k}(x^k) - f_{i_k}(x_p^k) = (f_{i_k}(x^k) - f_{i_k}^*) - (f_{i_k}(x_p^k) - f_{i_k}^*)$. Now we write $\gamma_k = \rho + \epsilon_k$ where ρ is a constant independent of i_k . Therefore, we have

$$\begin{aligned}
\mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\rho\mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\
&\quad - 2\alpha\mathbb{E}_k [\epsilon_k(f_{i_k}(x^k) - f_{i_k}^*)] + 2\alpha\mathbb{E}_k [\epsilon_k(f_{i_k}(x_p^k) - f_{i_k}^*)] \\
&\quad + 2\beta\mathbb{E}_k [\gamma_k(f_{i_k}(x^k) - f_{i_k}^*)] + \frac{4\gamma L}{1+2\gamma L}\mathbb{E}_k [\gamma_k(f_{i_k}(x^k) - f_{i_k}^*)] \\
&\quad + \frac{2\gamma^2 L}{1+\gamma L} \max\left\{\frac{2\gamma L-1}{2\gamma L+1}, 0\right\} \mathbb{E}_k [f_{i_k}^*] \\
&= \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\rho\mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\
&\quad - 2\mathbb{E}_k \left[\left(\alpha\epsilon_k - \left(\beta + \frac{2\gamma L}{1+2\gamma L} \right) \gamma_k \right) (f_{i_k}(x^k) - f_{i_k}^*) \right] \\
&\quad + 2\alpha\mathbb{E}_k [\epsilon_k(f_{i_k}(x_p^k) - f_{i_k}^*)] \\
&\quad + \frac{2\gamma^2 L}{1+\gamma L} \max\left\{\frac{2\gamma L-1}{2\gamma L+1}, 0\right\} \mathbb{E}_k [f_{i_k}^*]. \tag{36}
\end{aligned}$$

We need to find ρ such that

$$\begin{aligned}
\alpha\epsilon_k - \left(\beta + \frac{2\gamma L}{1+2\gamma L} \right) \gamma_k &\geq 0 \\
\alpha(\gamma_k - \rho) - \left(\beta + \frac{2\gamma L}{1+2\gamma L} \right) \gamma_k &\geq 0 \\
\gamma_k \left(\alpha - \beta - \frac{2\gamma L}{1+2\gamma L} \right) &\geq \alpha\rho.
\end{aligned}$$

Note that since $\gamma_k \geq \frac{\gamma}{1+\gamma L}$ from Lemma 4, then it is enough if ρ satisfies

$$\begin{aligned}
\frac{\gamma}{1+\gamma L} \left(\alpha - \beta - \frac{2\gamma L}{1+2\gamma L} \right) &\geq \alpha\rho \\
\frac{\gamma(2\gamma L(\alpha - \beta - 1) + \alpha - \beta)}{\alpha(1+\gamma L)(1+2\gamma L)} &\geq \rho.
\end{aligned}$$

Let us take this bound as a value of ρ . Note that since $\alpha \geq \beta + 1$, then $\rho \geq 0$. Therefore, the bound for ϵ_k is the following

$$\begin{aligned}
\epsilon_k &= \gamma_k - \rho \\
&\leq \gamma - \frac{\gamma(2\gamma L(\alpha - \beta - 1) + \alpha - \beta)}{\alpha(1+\gamma L)(1+2\gamma L)} \\
&= \gamma \left(\frac{\alpha(1+\gamma L)(1+2\gamma L) - 2\gamma L(\alpha - \beta - 1) - (\alpha - \beta)}{\alpha(1+\gamma L)(1+2\gamma L)} \right) \\
&= L\gamma^2 \frac{\alpha + 2\alpha\gamma L + 2(\beta + 1)}{\alpha(1+\gamma L)(1+2\gamma L)} + \gamma \frac{\beta}{\alpha(1+\gamma L)(1+2\gamma L)} \\
&\leq \underbrace{\frac{3L\gamma^2}{1+2\gamma L}}_{:= \frac{1}{2}T_2(\gamma^2)} + \underbrace{\frac{\beta\gamma}{\alpha(1+\gamma L)(1+2\gamma L)}}_{:= \frac{1}{2}\beta T_1(\gamma)}.
\end{aligned}$$

Therefore, we get the following descent inequality

$$\begin{aligned}
\mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - \underbrace{2\alpha\rho}_{:=T_0(\gamma)} \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\
&\quad + 2\alpha \mathbb{E}_k [\epsilon_k(f_{i_k}(x_p^k) - f_{i_k}^*)] + \underbrace{\frac{2\gamma^2 L}{1+\gamma L} \max\left\{\frac{2\gamma L-1}{2\gamma L+1}, 0\right\} \mathbb{E}_k [f_{i_k}^*]}_{:=T_3(\gamma^2)} \\
&\leq \text{dist}(x^k, \mathcal{S})^2 - T_0(\gamma) \mathbb{E}_k [f_{i_k}(x^k) - f_{i_k}(x_p^k)] \\
&\quad + T_2(\gamma^2) \alpha \mathbb{E}_k [f_{i_k}(x_p^k) - f_{i_k}^*] + \beta T_1(\gamma) \alpha \mathbb{E}_k [f_{i_k}(x_p^k) - f_{i_k}^*] \\
&\quad + T_3(\gamma^2) \mathbb{E}_k [f_{i_k}^*] \\
&= \text{dist}(x^k, \mathcal{S})^2 - T_0(\gamma)(f(x^k) - f^*) + T_2(\gamma^2) \alpha \mathbb{E}_k [f^* - f_{i_k}^*] \\
&\quad + \beta T_1(\gamma) \alpha \mathbb{E}_k [f^* - f_{i_k}^*] + T_3(\gamma^2) \mathbb{E}_k [f_{i_k}^*]. \tag{37}
\end{aligned}$$

Here we use the fact that x_p^k is a minimizer of f and independent of i_k . Unrolling the recursion, we get

$$\begin{aligned}
\frac{1}{K} (\mathbb{E} [\text{dist}(x^K, \mathcal{S})^2] - \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]) &\leq -\frac{T_0(\gamma)}{K} \sum_{k=0}^{K-1} \mathbb{E} [f(x_k) - f^*] \\
&\quad + E_1(\gamma) + E_2(\gamma^2), \tag{38}
\end{aligned}$$

where

$$E_1(\gamma) := \beta T_1(\gamma) \alpha \underbrace{\mathbb{E} [f^* - f_{i_k}^*]}_{\sigma_{\text{int}}^2}, \quad E_2(\gamma^2) := T_2(\gamma^2) \alpha \underbrace{\mathbb{E} [f^* - f_i^*]}_{\sigma_{\text{int}}^2} + T_3(\gamma^2) \underbrace{\mathbb{E} [f_i^*]}_{\sigma_{\text{pos}}^2}. \tag{39}$$

Therefore, we get

$$\begin{aligned}
\min_{0 \leq k \leq K-1} \mathbb{E} [f(x_k) - f^*] &\leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{KT_0(\gamma)} + \frac{E_1(\gamma)}{T_0(\gamma)} + \frac{E_2(\gamma^2)}{T_0(\gamma)} \\
&\leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{2\gamma K} \frac{(1+2\gamma L)^2}{2\gamma L(\alpha-\beta-1)+\alpha-\beta} \\
&\quad + \frac{3L\gamma\alpha}{2\gamma L(\alpha-\beta-1)+\alpha-\beta} (1+\gamma L) \sigma_{\text{int}}^2 \\
&\quad + \frac{\gamma L}{2\gamma L(\alpha-\beta-1)+\alpha-\beta} \max\{2\gamma L-1, 0\} \sigma_{\text{pos}}^2 \\
&\quad + \frac{\beta\sigma_{\text{int}}^2}{2\gamma L(\alpha-\beta-1)+\alpha-\beta}. \tag{40}
\end{aligned}$$

This finished the proof. \square

Remark 5. If all f_i are convex, i.e. we can take $\alpha = 1, \beta = 0$ in α - β -condition, then we get

$$\begin{aligned}
\min_{0 \leq k \leq K-1} \mathbb{E} [f(x_k) - f^*] &\leq \frac{\mathbb{E} [\text{dist}(x^0, \mathcal{S})^2]}{2\gamma K} (1+2\gamma L)^2 + 3L\gamma(1+\gamma L) \sigma_{\text{int}}^2 \\
&\quad + \gamma L \max\{2\gamma L-1, 0\} \sigma_{\text{pos}}^2,
\end{aligned}$$

that coincides with the results in [63].

Decreasing stepsize parameter. Now we present the results with decreasing stepsize parameter.

Theorem 6. Assume that Assumptions 1 with $\alpha \geq \beta + 1$ and 1-2 hold. Assume that each function f_i is positive and $\sigma_{\text{pos}}^2 < \infty$. Then the iterates of NGN with decreasing stepsize of the form

$$\gamma_k = \frac{\tilde{\gamma}_k}{1 + \frac{\tilde{\gamma}_k}{2f_{i_k}(x^k)} \|\nabla f_{i_k}(x^k)\|^2}, \quad \tilde{\gamma}_k := \frac{\gamma}{\sqrt{k+1}}$$

satisfy

$$\begin{aligned} \min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{C_1}{\sqrt{K}} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] + \frac{C_2 \log(K+1)}{\sqrt{K}} \alpha \sigma_{\text{int}}^2 + \frac{C_3}{\alpha - \beta} \beta \sigma_{\text{int}}^2 \\ &\quad + \frac{C_4 \log(K+1)}{\sqrt{K}} \sigma_{\text{pos}}^2 \end{aligned} \quad (41)$$

$$= \tilde{\mathcal{O}} \left(\frac{1}{\sqrt{K}} + \beta \sigma_{\text{int}}^2 \right). \quad (42)$$

where C_1, C_2, C_3 , and C_4 are defined in (43).

Proof. Similarly to the proof with constant stepsize parameter we can obtain (similar to (37))

$$\begin{aligned} \mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \text{dist}(x^k, \mathcal{S})^2 - T_0(\tilde{\gamma}_k)(f(x^k) - f^*) + T_2(\tilde{\gamma}_k^2) \alpha \sigma_{\text{int}}^2 \\ &\quad + \beta T_1(\tilde{\gamma}_k) \alpha \sigma_{\text{int}}^2 + T_3(\tilde{\gamma}_k^2) \sigma_{\text{pos}}^2. \end{aligned}$$

Note that we have

$$\begin{aligned} T_0(\tilde{\gamma}_k) &= \frac{2\tilde{\gamma}_k(2\tilde{\gamma}_k L(\alpha - \beta - 1) + \alpha - \beta)}{(1 + \tilde{\gamma}_k L)(1 + 2\tilde{\gamma}_k L)} \\ &\geq \frac{2\tilde{\gamma}_k(\alpha - \beta)}{(1 + \tilde{\gamma}_0 L)(1 + 2\tilde{\gamma}_0 L)} =: \tilde{T}_0(\tilde{\gamma}_k), \\ T_1(\tilde{\gamma}_k) &= \frac{2\tilde{\gamma}_k}{\alpha(1 + \tilde{\gamma}_k L)(1 + 2\tilde{\gamma}_k L)} \\ &\leq \frac{2\tilde{\gamma}_k}{\alpha} =: \tilde{T}_1(\tilde{\gamma}_k), \\ T_2(\tilde{\gamma}_k^2) &= \frac{6L\tilde{\gamma}_k^2}{1 + 2\tilde{\gamma}_k L} \\ &\leq 6L\tilde{\gamma}_k^2 =: \tilde{T}_2(\tilde{\gamma}_k^2), \\ T_3(\tilde{\gamma}_k^2) &= \frac{2\tilde{\gamma}_k^2 L}{1 + \tilde{\gamma}_k L} \max \left\{ \frac{2\tilde{\gamma}_k L - 1}{2\tilde{\gamma}_k L + 1}, 0 \right\} \\ &\leq 2\tilde{\gamma}_k^2 L \max\{2\gamma L - 1, 0\} =: \tilde{T}_3(\tilde{\gamma}_k^2). \end{aligned}$$

Therefore, we can continue as follows

$$\begin{aligned} \mathbb{E} [\text{dist}(x^{k+1}, \mathcal{S})^2] &\leq \mathbb{E} [\text{dist}(x^k, \mathcal{S})^2] - \tilde{T}_0(\tilde{\gamma}_k) \mathbb{E} [f(x^k) - f^*] + \tilde{T}_2(\tilde{\gamma}_k^2) \alpha \sigma_{\text{int}}^2 \\ &\quad + \beta \tilde{T}_1(\tilde{\gamma}_k) \alpha \sigma_{\text{int}}^2 + \tilde{T}_3(\tilde{\gamma}_k^2) \sigma_{\text{pos}}^2. \end{aligned}$$

This leads to

$$\begin{aligned} \sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k) \mathbb{E} [f(x^k) - f^*] &\leq \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] + \sum_{k=0}^{K-1} \tilde{T}_2(\tilde{\gamma}_k^2) \alpha \sigma_{\text{int}}^2 + \sum_{k=0}^{K-1} \beta \tilde{T}_1(\tilde{\gamma}_k) \alpha \sigma_{\text{int}}^2 \\ &\quad + \sum_{k=0}^{K-1} \tilde{T}_3(\tilde{\gamma}_k^2) \sigma_{\text{pos}}^2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{1}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] + \frac{\sum_{k=0}^{K-1} \tilde{T}_2(\tilde{\gamma}_k^2) \alpha \sigma_{\text{int}}^2}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)} \\ &\quad + \frac{\beta \sum_{k=0}^{K-1} \tilde{T}_1(\tilde{\gamma}_k) \alpha \sigma_{\text{int}}^2}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)} + \frac{\sum_{k=0}^{K-1} \tilde{T}_3(\tilde{\gamma}_k^2) \sigma_{\text{pos}}^2}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)}. \end{aligned}$$

Following the results of [64] and [24] we get

$$\begin{aligned}
\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k) &= \sum_{k=0}^{K-1} \frac{2\tilde{\gamma}_k}{(1+\gamma L)(1+2\gamma L)} \\
&\geq \frac{8\gamma\sqrt{K}}{5(1+\gamma L)(1+2\gamma L)}, \\
\sum_{k=0}^{K-1} \tilde{T}_2(\tilde{\gamma}_k^2) &= \sum_{k=0}^{K-1} 6L\tilde{\gamma}_k^2 \\
&\leq 12L\gamma^2 \log(K+1), \\
\sum_{k=0}^{K-1} \tilde{T}_3(\tilde{\gamma}_k^2) &= \sum_{k=0}^{K-1} 2\tilde{\gamma}_k^2 L \max\{2\gamma L-1, 0\} \\
&\leq 4\gamma^2 L \max\{2\gamma L-1, 0\} \log(K+1), \\
\frac{\sum_{k=0}^{K-1} \tilde{T}_1(\tilde{\gamma}_k)}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)} &= \frac{\sum_{k=0}^{K-1} \frac{2\tilde{\gamma}_k}{\alpha}}{\sum_{k=0}^{K-1} \frac{2\tilde{\gamma}_k(\alpha-\beta)}{(1+\gamma L)(1+2\gamma L)}} \\
&= \frac{(1+\gamma L)(1+2\gamma L)}{\alpha(\alpha-\beta)}.
\end{aligned}$$

Thus, the final result can be written as follows

$$\begin{aligned}
\min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{5(1+\gamma L)(1+2\gamma L)}{8\gamma\sqrt{K}} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] + \frac{12L\gamma^2 \log(K+1)}{\frac{8\gamma\sqrt{K}}{5(1+\gamma L)(1+2\gamma L)}} \alpha \sigma_{\text{int}}^2 \\
&\quad + \frac{\sum_{k=0}^{K-1} \tilde{T}_1(\tilde{\gamma}_k)}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)} \alpha \sigma_{\text{int}}^2 + \frac{\sum_{k=0}^{K-1} \tilde{T}_3(\tilde{\gamma}_k^2)}{\sum_{k=0}^{K-1} \tilde{T}_0(\tilde{\gamma}_k)} \sigma_{\text{pos}}^2 \\
&= \frac{5(1+\gamma L)(1+2\gamma L)}{8\gamma\sqrt{K}} \mathbb{E} [\text{dist}(x^0, \mathcal{S})^2] \\
&\quad + \frac{15L\gamma(1+\gamma L)(1+2\gamma L) \log(K+1)}{2\sqrt{K}} \alpha \sigma_{\text{int}}^2 \\
&\quad + \frac{(1+\gamma L)(1+2\gamma L)}{(\alpha-\beta)} \beta \sigma_{\text{int}}^2 \\
&\quad + \frac{5(1+\gamma L)(1+2\gamma L)\gamma L \max\{2\gamma L-1, 0\} \log(K+1)}{2\sqrt{K}} \sigma_{\text{pos}}^2.
\end{aligned}$$

Now it is left to use the definitions of constants C_1, C_2 , and C_3

$$\begin{aligned}
C_1 &:= \frac{5(1+\gamma L)(1+2\gamma L)}{8\gamma}, \\
C_2 &:= \frac{15L\gamma(1+\gamma L)(1+2\gamma L)}{2}, \\
C_3 &:= (1+\gamma L)(1+2\gamma L), \\
C_4 &:= \frac{5(1+\gamma L)(1+2\gamma L)\gamma L \max\{2\gamma L-1, 0\}}{2}.
\end{aligned} \tag{43}$$

□

C.2.4 Convergence of AdaGrad-norm-max

Theorem 7. Assume that Assumptions 1-2 hold. Assume that for all $k \geq 0$ stochastic gradients satisfy $\|\nabla f_{i_k}(x^k)\|^2 \leq G^2$ for some $G > 0$ and $b_{-1} \geq \frac{2L\gamma}{\alpha-\beta}$. Let $D^2 = \max_i \|\nabla f_i(x^0)\|^2$. Then the

Algorithm 4 AdaGrad-norm-max

1: **Input:** Stepsize parameter $\gamma > 0$, $c_{-1} = 0$, $b_{-1} \geq \frac{2L\gamma}{\alpha-\beta}$
 2: **for** $k = 0, 1, 2, \dots, K-1$ **do**
 3: Sample $i_k \sim \text{Unif}[n]$ and compute $\nabla f_{i_k}(x^k)$
 4: Compute AdaGrad-norm-max stepsize

$$c_k^2 := \max \{c_{k-1}^2, \|\nabla f_{i_k}(x^k)\|^2\}$$

$$b_k^2 := b_{k-1}^2 + c_k^2, \quad \gamma_k = \frac{\gamma}{b_k}$$

 5: Update model

$$x^{k+1} = x^k - \gamma_k \nabla f_{i_k}(x^k)$$

 6: **end for**

iterates of AdaGrad-norm-max (Alg. 4) satisfy

$$\begin{aligned}
 \min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{\text{dist}(x^k, \mathcal{S})^2}{\gamma K(\alpha - \beta)} \sqrt{b_{-1}^2 + G^2 K} \\
 &\quad + \frac{2\alpha}{(\alpha - \beta) K D^2} \sigma_{\text{int}}^2 \sqrt{b_{-1}^2 + G^2 K} \sqrt{b_{-1}^2 + D^2(K+1)} \\
 &= \mathcal{O} \left(\frac{1}{\sqrt{K}} + \alpha \sigma_{\text{int}}^2 \right).
 \end{aligned} \tag{44}$$

We see that if K is large enough, then we recover the standard convex rate of Adagrad of order $\tilde{\mathcal{O}}(K^{-1/2})$ [46].

Proof. Let $x_p^k = \text{Proj}(x^k, \mathcal{S})$ satysfying Definition 1. Then we have

$$\begin{aligned}
 \text{dist}(x^{k+1}, \mathcal{S})^2 &\leq \|x^{k+1} - x_p^k\|^2 \\
 &= \|x^k - x_p^k\|^2 - 2\gamma_k \langle \nabla f_{i_k}(x^k), x^k - x_p^k \rangle + \gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2 \\
 &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\gamma_k (f_{i_k}(x^k) - f_{i_k}(x_p^k)) + 2\beta\gamma_k (f_{i_k}(x^k) - f_{i_k}^*) \\
 &\quad + \gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2 \\
 &\leq \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\gamma_k (f_{i_k}(x^k) - f_{i_k}(x_p^k)) + 2\beta\gamma_k (f_{i_k}(x^k) - f_{i_k}^*) \\
 &\quad + 2\gamma_k^2 L (f_{i_k}(x^k) - f_{i_k}^*) \\
 &= \text{dist}(x^k, \mathcal{S})^2 - 2\alpha\gamma_k (f_{i_k}(x^k) - f_{i_k}^* + f_{i_k}^* - f_{i_k}(x_p^k)) + 2\beta\gamma_k (f_{i_k}(x^k) - f_{i_k}^*) \\
 &\quad + 2\gamma_k^2 L (f_{i_k}(x^k) - f_{i_k}^*) \\
 &= \text{dist}(x^k, \mathcal{S})^2 - 2\gamma_k (\alpha - \beta - L\gamma_k) (f_{i_k}(x^k) - f_{i_k}^*) \\
 &\quad + 2\alpha\gamma_k (f_{i_k}(x_p^k) - f_{i_k}^*).
 \end{aligned} \tag{45}$$

Note that we choose $b_{-1} \geq \frac{2L\gamma}{\alpha-\beta}$. Since b_k is inreasing sequence, then we have for any k that $b_k \geq \frac{2L\gamma}{\alpha-\beta}$. Therefore,

$$L\gamma_k = L \frac{\gamma}{b_k} \leq L \frac{\gamma}{b_{-1}} \leq L \frac{\gamma}{2L\gamma/\alpha-\beta} = \frac{\alpha-\beta}{2}.$$

This means that

$$-2\gamma_k (\alpha - \beta - L\gamma_k) \leq -2\gamma_k (\alpha - \beta - \alpha-\beta/2) = -\gamma_k (\alpha - \beta).$$

Thus, we can continue (45) as follows

$$\text{dist}(x^{k+1}, \mathcal{S})^2 \leq \text{dist}(x^k, \mathcal{S})^2 - \frac{\gamma}{b_k} (\alpha - \beta) (f_{i_k}(x^k) - f_{i_k}^*) + 2\alpha \frac{\gamma}{b_k} (f_{i_k}(x_p^k) - f_{i_k}^*).$$

We know that b_k is increasing sequence that satisfy

$$\sqrt{b_{-1}^2 + D^2(k+1)} \leq b_k \leq b_{K-1} \leq \sqrt{b_{-1}^2 + G^2K}.$$

This leads together with the fact that both $f_{i_k}(x^k) - f_{i_k}^*$ and $f_{i_k}(x_p^k) - f_{i_k}^*$ are non-negative to

$$\begin{aligned} \text{dist}(x^{k+1}, \mathcal{S})^2 &\leq \text{dist}(x^k, \mathcal{S})^2 - \frac{\gamma}{\sqrt{b_{-1}^2 + G^2K}}(\alpha - \beta)(f_{i_k}(x^k) - f_{i_k}^*) \\ &\quad + 2\alpha \frac{\gamma}{\sqrt{b_{-1}^2 + D^2(k+1)}}(f_{i_k}(x_p^k) - f_{i_k}^*) \\ &= \text{dist}(x^k, \mathcal{S})^2 - \frac{\gamma(\alpha - \beta)}{\sqrt{b_{-1}^2 + G^2K}}(f_{i_k}(x^k) - f_{i_k}(x_p^k)) \\ &\quad + \frac{2\alpha\gamma}{\sqrt{b_{-1}^2 + D^2(k+1)}}(f_{i_k}(x_p^k) - f_{i_k}^*). \end{aligned}$$

Taking the conditional expectation we get

$$\frac{\mathbb{E}_k [f(x^k) - f^*]}{\sqrt{b_{-1}^2 + G^2K}} \leq \frac{\text{dist}(x^k, \mathcal{S})^2}{\gamma(\alpha - \beta)} - \frac{\mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2]}{\gamma(\alpha - \beta)} + \frac{2\alpha}{\sqrt{b_{-1}^2 + D^2(k+1)}(\alpha - \beta)}\sigma_{\text{int}}^2,$$

which leads to the following bound

$$\begin{aligned} \mathbb{E}_k [f(x^k) - f^*] &\leq \left(\frac{\text{dist}(x^k, \mathcal{S})^2}{\gamma(\alpha - \beta)} - \frac{\mathbb{E}_k [\text{dist}(x^{k+1}, \mathcal{S})^2]}{\gamma(\alpha - \beta)} \right) \sqrt{b_{-1}^2 + G^2K} \\ &\quad + \frac{2\alpha}{\sqrt{b_{-1}^2 + D^2(k+1)}(\alpha - \beta)}\sigma_{\text{int}}^2\sqrt{b_{-1}^2 + G^2K}. \end{aligned}$$

Averaging over K iterations we get

$$\begin{aligned} \min_{0 \leq k < K} \mathbb{E} [f(x^k) - f^*] &\leq \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} [f(x^k) - f^*] \\ &\leq \frac{\text{dist}(x^k, \mathcal{S})^2}{\gamma K(\alpha - \beta)} \sqrt{b_{-1}^2 + G^2K} \\ &\quad + \frac{2\alpha}{(\alpha - \beta)K}\sigma_{\text{int}}^2\sqrt{b_{-1}^2 + G^2K} \sum_{k=0}^{K-1} \frac{1}{\sqrt{b_{-1}^2 + D^2(k+1)}} \\ &\leq \frac{\text{dist}(x^k, \mathcal{S})^2}{\gamma K(\alpha - \beta)} \sqrt{b_{-1}^2 + G^2K} \\ &\quad + \frac{2\alpha}{(\alpha - \beta)KD^2}\sigma_{\text{int}}^2\sqrt{b_{-1}^2 + G^2K}\sqrt{b_{-1}^2 + D^2(K+1)} \\ &= \mathcal{O}\left(\frac{1}{\sqrt{K}} + \alpha\sigma_{\text{int}}^2\right). \end{aligned}$$

□

D Additional experiments

For all the experiments, we make use of PyTorch [65] package. LSTM, MLP, CNN and Resnet experiments are performed using one NVIDIA GeForce RTX 3090 GPU with a memory of 24 GB. For training Algoperf and Pythia language models, we resort instead to 4xA100-SXM4 GPUs, with a memory of 40 GB each, and employ data parallelism for efficient distributed training. When necessary, we disable CUDA non-deterministic operations, to allow consistency between runs with the same random seed.

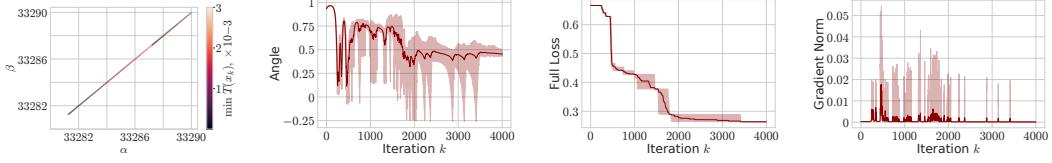


Figure 9: Training for half-space learning problem with SGD. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* \approx 0.000523853$; angle denotes $\angle(\nabla f(x^k), x^k - x^K)$.

D.1 Half space learning

Half Space Learning problem corresponds to the following optimization problem

$$f(x) = \frac{1}{n} \sum_{i=1}^n \sigma(-b_i x^\top a_i) + \frac{\lambda}{2} \|x\|^2,$$

where $\{a_i, b_i\}_{i=1}^n, a_i \in \mathbb{R}^d, b_i \in \{0, 1\}$ is a given dataset, $\lambda = 10^{-5}$, and σ is a sigmoid function. For the test, we create a synthetic dataset that contains 20 samples for both classes. We sample data points from normal distribution where each class has its own mean and variance value of 2. We use SGD with learning rate $\frac{1}{4}$ and batch size 1 for minimization task.⁵

We observe that the gradient norm becomes zero quite which means that SGD trajectory goes through saddle point. This leads to possible negative angle between full gradient and direction to minimizer. Nevertheless, we demonstrate that even for such highly non-convex landscape with many saddle points our α - β -condition holds for large enough values of α and β .

D.2 Experiment setup from section 2.2

We use 3 layer LSTM based model from Hübler et al. [33]⁶. The model for PTB dataset has 35441600 parameters while for Wikitext-2 dataset the model has 44798534 parameters. To train the model, we choose NSGD with momentum [15] with decaying stepsize and momentum parameters according to the experiment section from [33]. We train the model for 1000 epochs with initial stepsize values 158 and 900 for PTB and Wikitext-2 datasets respectively. We switch off dropout during measuring stochastic gradients and losses for α - β -condition. We run the experiments for 7 different random seeds and plot the mean with maximum and minimum fluctuations.

In Figure 10, we plot empirical aiming coefficient $\frac{\langle \nabla f(x^k), x^k - x^K \rangle}{f(x^k)}$ for full loss f ; mean of stochastic losses across 7 runs along with maximum and minimum fluctuations from the mean; pairs of (α, β) that satisfy α - β -condition across all 7 runs. We observe that for both datasets, there is a plateau at the beginning of the training. This part of the training corresponds to possible negative values of the empirical coefficient Aiming condition. After this, it becomes stable and positive.

Besides, we demonstrate that possible values of pairs of (α, β) that satisfy α - β -condition are large. We believe, this happens because the full loss f has a minimum value of around 3.5 while individual losses have f_i^* , i.e. the model is far from the interpolation regime (when $f^* = f_i^*$).

D.3 MLP architecture

We use MLP model with 3 fully connected layers. We fix the dimensions of the second layer to be equal, i.e. the parameter matrix of this layer is square. After the first and second fully connected layers we use ReLU activation function. We train the model in all cases with fixed learning rate 0.09 for 1500 epochs and batch size 64 on Fashion-MNIST [83] dataset. In Figure 12, we plot the

⁵We use the implementation from [17] that can be found <https://github.com/archana95in/Escaping-saddles-using-Stochastic-Gradient-Descent>.

⁶The implementation can be found following the link <https://github.com/fhueb/parameter-agnostic-lzlo>

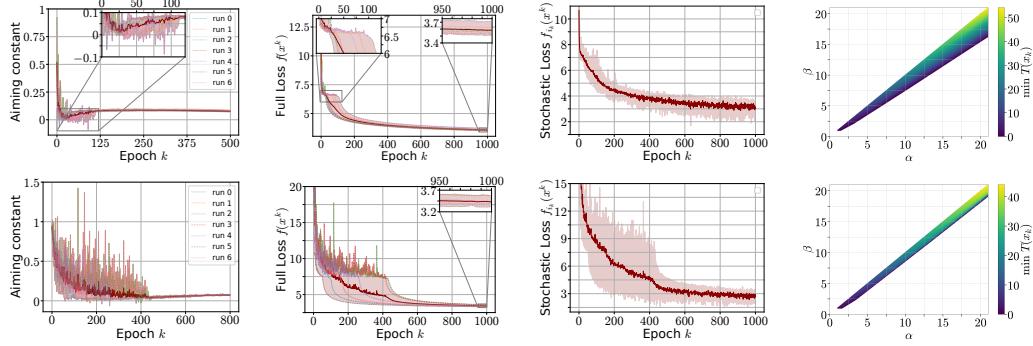


Figure 10: Training of 3 layer LSTM model on PTB (first row) and Wikitext-2 (second row) datasets. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$.

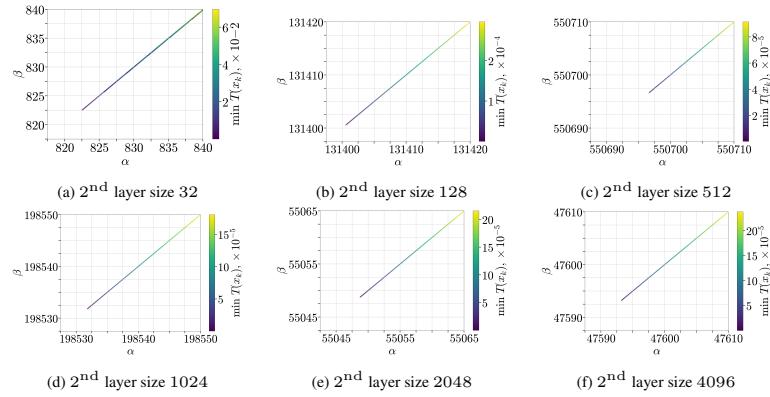


Figure 11: Values of α and β during the training of 3 layer MLP model on Fashion-MNIST dataset varying the dimension of the second layer. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$.

mean stochastic loss across 4 runs along with maximum and minimum fluctuations, and in Figure 11 possible values of α and β that work over all random seeds and iterations satisfying $\alpha \geq \beta + 0.1$.

We observe that the magnitude of the smallest possible values of α and β increase till up to the second layer size 512, but then it starts decreasing as the model becomes more over-parameterized. This leads to smaller values of neighborhood $\mathcal{O}(\beta \sigma_{\text{int}}^2)$ as we expect in this setting.

D.4 CNN architecture

We use CNN model with 2 convolution layers followed by a fully connected one. After each convolution layer, we use max-pooling and ReLU activation functions. We train the model with a cosine annealing learning rate scheduler with a maximum value 0.01 and batch size 64. We train the model on CIFAR10 dataset [41] for 1500 epochs. We run the experiments for 4 different random seeds. In Figure 13 we plot possible values of α and β satisfying $\alpha \geq \beta + 0.1$ that work across all runs and iterations.

We observe that increasing the dimension of the second layer of the model makes the model closer to over-parameterization: values of stochastic losses decrease. We observe the same phenomenon as in Appendix D.3: minimum possible values of α and β increase up to 128 number of convolutions, but then it decreases for a larger number of convolutions. This happens because the model becomes more over-parameterized. Moreover, the possible difference between α and β tends to increase with number of convolutions starting from 128 convolutions.

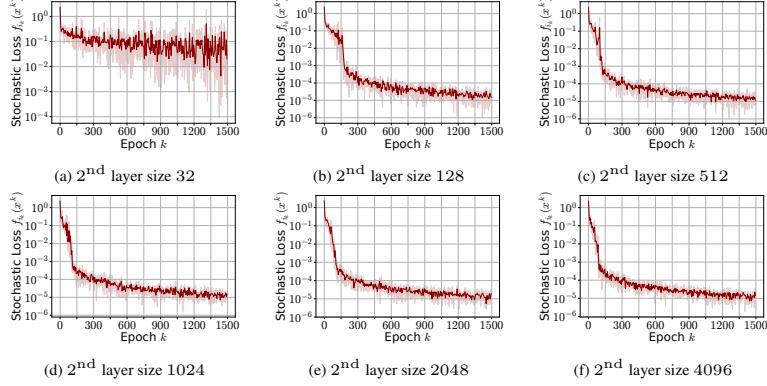


Figure 12: Values of stochastic loss during the training of 3 layer MLP model on Fashion-MNIST dataset varying the dimension of the second layer.

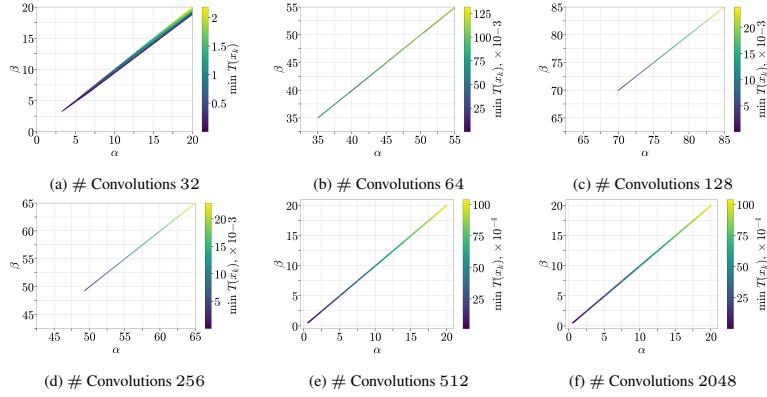


Figure 13: Values of α and β during the training of CNN model on CIFAR10 dataset varying the number of convolutions in the second layer. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$.

D.5 Resnet architecture

We use the implementation from Kumar [42]. We train the model on CIFAR100 dataset [41] for 1000 epochs. We use one-cycle scheduler with a maximum learning rate 0.01. To compute dense stochastic gradients, we switch off dropout during evaluations of stochastic gradients and losses for α - β -condition. We run the experiments for 4 different random seeds and plot the mean along with maximum and minimum fluctuations.

In Figure 15 we observe that the minimum value of stochastic loss increases with batch size which means that the model becomes further from over-parameterization.

D.6 Verification of α - β -condition by different optimizers

Now we turn to another interesting question: how does the choice of an optimizer affect the practical verification of the α - β -condition? To explore this question, we train Resnet9 model with SGD, SGDM, and Adam. We report the results in Figure 16 varying the batch size used in the training. Comparing the values of α and β for SGD (from Figure 6), SGDM, and Adam, we observe that the loss landscape explored by the Adam optimizer achieves smaller values of α and β . Moreover, the values of α and β found by SGDM are typically smaller than those found by SGD. This result may shed light on why momentum (from a comparison of SGDM against SGD) and adaptive stepsize (from a comparison of SGDM against Adam) are typically beneficial in practice: these more advanced algorithms explore better part of a loss landscape from the α - β -condition point of view.

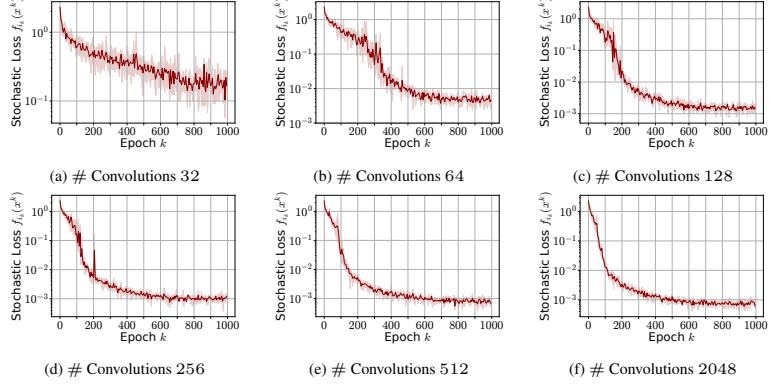


Figure 14: Values of stochastic loss during the training of CNN model on CIFAR10 dataset varying the number of convolutions in the second layer.

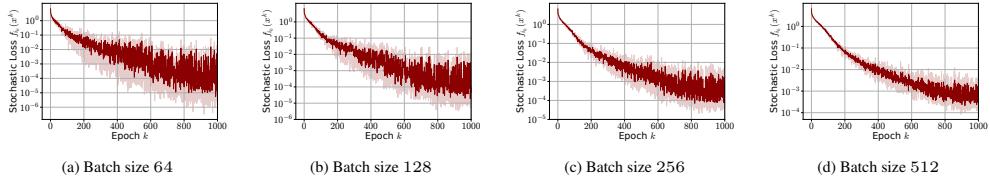


Figure 15: Training of Resnet9 model on CIFAR100 dataset varying the batch size.

Table 3: Training details of large models from Appendix D.8 and Appendix D.9

Dataset	Model	Batch Size	LR	β_1	β_2	Weight Decay	Warmup
Criteo 1TB	DLRMsmall	262144	0.0017	0.93	0.995	0.08	0.02
Fastmri	U-Net	32	0.001	0.9	0.998	0.15	0.1
OBGB	GNN	512	0.0017	0.93	0.995	0.08	0.02
WMT	Transformer	128	0.001	0.97	0.999	0.15	0.1
Slim-Pajama-627B	Pythia-70M	256	0.01	0.9	0.95	0.1	0.1
Slim-Pajama-627B	Pythia-160M	256	0.006	0.9	0.95	0.1	0.1

D.7 Increasing the depth of Resnet architecture

In our next experiment, we aim to investigate how the α - β -condition behaves increasing the depth of a model. To do so, in addition to training Resnet9 model, we test Resnet18 and Resnet34 models on CIFAR100 dataset. The results in Figure 17 suggest that the values of α and β tend to decrease with the depth of a model. These observations align with those from Section 5.1 and Section 5.2 as the constants α and β decrease as the depth of the model increases.

D.8 AlgoPerf experiments

For each of all aforementioned tasks, we repeat the training with 3 random seeds to create more stable results. The detailed model architectures are given in [16]. In Table 3 we provide the parameters of optimizers we use for each task. The loss curves are presented in Figure 18. For each workload, we run the experiments for 3 different random seeds to obtain more stable results. The hyperparameters of optimizer NadamW are chosen such that we can reach the validation threshold set by the organizers⁷. We employ a cosine annealing learning rate schedule that reduces the learning to $1e-10$, with an initial linear warm-up. For each workload, we run the experiments sufficiently enough so that we reach the validation target threshold and the stochastic loss becomes sufficiently stable.

⁷The quality of performance is measured differently from one task to another; we defer to the [16] for a more detailed description of the competition.

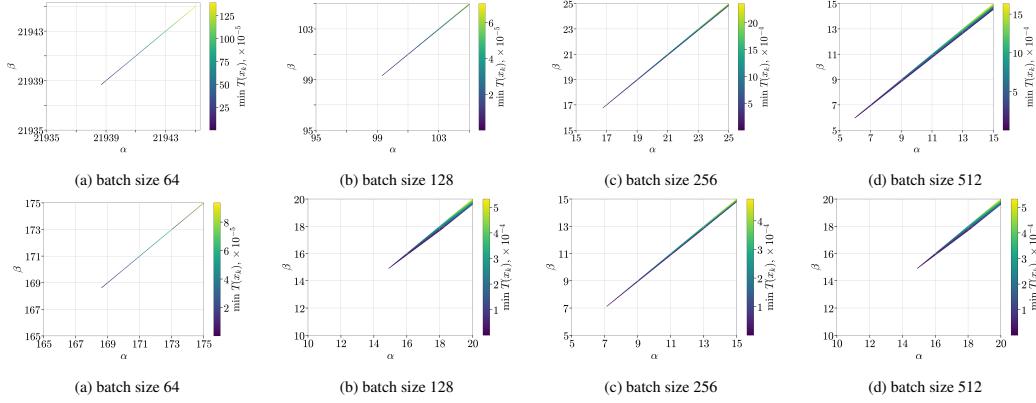


Figure 16: Training of ResNet9 model on CIFAR100 dataset varying the batch size with SGDM (stepsize 0.01 and momentum 0.9 with OneCycle learning rate scheduler) and Adam (stepsize 0.0001 with default momentum parameters with OneCycle learning rate scheduler) optimizers. Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for a given pair of (α, β) . We plot values of α, β in α - β -condition for SGDM (**first row**) and Adam (**second row**).

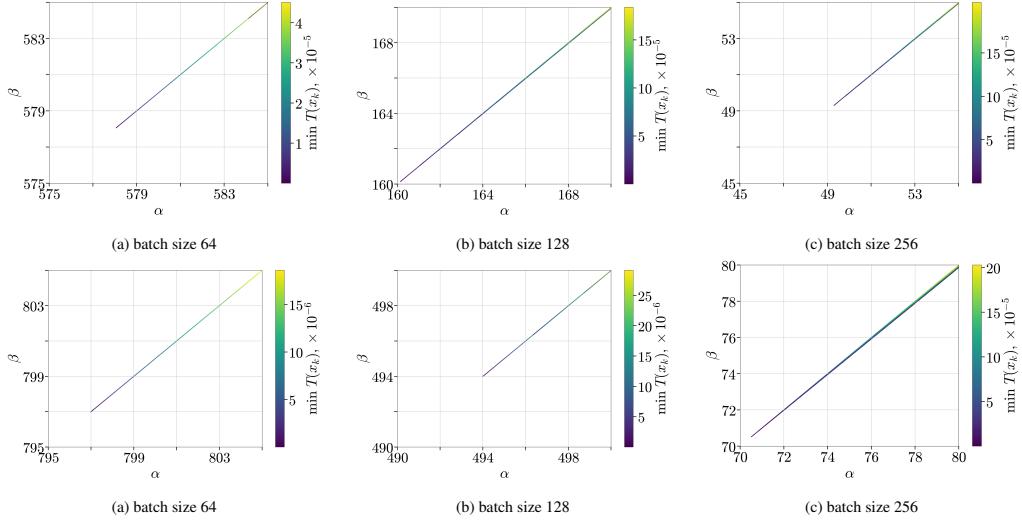


Figure 17: Training of ResNet18 and Resnet34 model on CIFAR100 dataset varying the batch size SGD (stepsize 0.01 and momentum 0.9 with OneCycle learning rate scheduler). Here $T(x_k) = \langle \nabla f_{i_k}(x^k), x^k - x^K \rangle - \alpha(f_{i_k}(x^k) - f_{i_k}(x^K)) - \beta f_{i_k}(x^k)$ assuming that $f_i^* = 0$. Minimum is taken across all runs and iterations for a given pair of (α, β) .

D.9 Pythia experiments

For each of all aforementioned tasks, we repeat the training with 3 random seeds to create more stable results. We train Pythia 70M and Pythia 160M [8] on publicly available Slim-Pajama-627B dataset [72]. Both models are trained on sequences of length 2048, and makes use of a batch size of 0.5M tokens, which amounts to a batch size of 256 samples. We use AdamW optimizer and a cosine annealing with linear warmup, with hyperparameters specified in Table 3. The stochastic loss and training perplexity are reported in Figure 19.

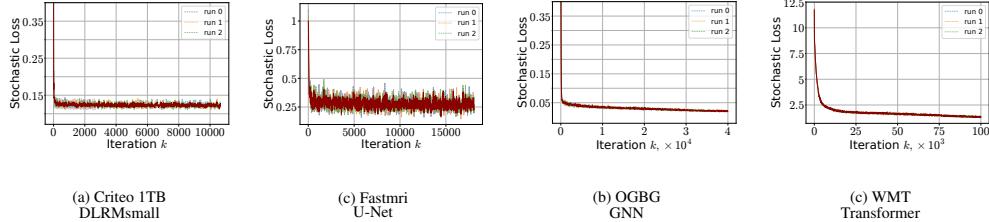


Figure 18: Training of large models from AlgoPerf benchmark.

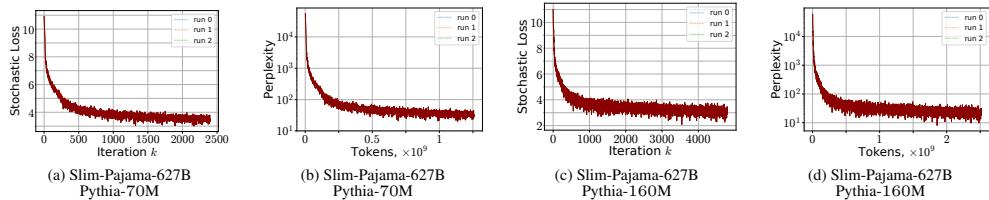


Figure 19: Training statistics for Pythia language models.

E Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: In the abstract, we shortly list all the main results of the paper. In the introduction and related works, we demonstrate the limitations of previous works and explicitly list the contributions of our work. We support theoretical claims by extensive experimental results on various deep learning benchmarks.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We provide a section in the end of the main body where we discuss several limitations of our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We explicitly state a new class of functions we propose in the paper as well as other assumptions on the smoothness and noise of ERM problem. We explicitly state all convergence guarantees for 4 different optimization algorithms. The proofs are deferred to the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: For most of the experiments we use existing open source Github repositories making small changes in the code to track necessary quantities during the training. We provide the links to all used repositories. Besides, we list all parameters of optimizers to reproduce the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[Yes]**

Justification: For each open source data and code we attach the link. All codes can be used almost immediately with small changes that are needed to track necessary quantities for validation of the proposed condition.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all hyperparameters and training details in the corresponding section of the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For each instance of the experiment section, we run tests for several random seeds (exact numbers are indicated in the experiment section of the appendix) to get more stable results. For each figure with training curves, we plot the mean and maximum and minimum fluctuations around it.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide information about GPU that we use for our experiments in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We follow all the rules listed following the link above. In particular, we anonymized the paper and attached the code.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: we provide an impact statement in ??.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[NA\]](#)

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use open-source datasets and code. For all used codes we explicitly cite the corresponding paper or Github repository.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.