Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence

Nicolas Loizou
Mila and DIRO
Université de Montréal

Sharan Vaswani[†]
University of Alberta

Issam Laradji
McGill, Element AI

Simon Lacoste-Julien
Mila and DIRO
Université de Montréal
Canada CIFAR AI Chair

Abstract

We propose a stochastic variant of the classical Polyak step-size (Polyak, 1987) commonly used in the subgradient method. Although computing the Polyak step-size requires knowledge of the optimal function values, this information is readily available for typical modern machine learning applications. Consequently, the proposed stochastic Polyak step-size (SPS) is an attractive choice for setting the learning rate for stochastic gradient descent (SGD). We provide theoretical convergence guarantees for SGD equipped with SPS in different settings, including strongly convex, convex and non-convex functions. Furthermore, our analysis results in novel convergence guarantees for SGD with a constant step-size. We show that SPS is particularly effective when training over-parameterized models capable of interpolating the training data. In this setting, we prove that SPS enables SGD to converge to the true solution at a fast rate without requiring the knowledge of any problem-dependent constants or additional computational overhead. We experimentally validate our theoretical results via extensive experiments on synthetic and real datasets. We demonstrate the strong performance of SGD with SPS compared to state-of-the-art optimization methods when training over-parameterized models.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. [†]Work done at Mila, Université de Montréal.

1 Introduction

We solve the finite-sum optimization problem:

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

This problem is prevalent in machine learning tasks where x corresponds to the model parameters, $f_i(x)$ represents the loss on the training point i and the aim is to minimize the average loss $f(x)$ across training points. We denote $\mathcal{X}^*\subset\mathbb{R}^d$ to be the set of optimal points x^* of (1) and assume that $\mathcal{X}^*$ is not empty. We use f^* to denote the minimum value of f, obtained at a point $x^*\in\mathcal{X}^*$. For each $i\in\{1,\dots,n\}$, we denote the infimum of function f_i by $f_i^* := \inf_x f_i(x)$. Depending on the model under study, the function f can either be strongly-convex, convex, or non-convex.

1.1 Background and Main Contributions

Stochastic gradient descent (SGD) (Robbins and Monro, 1951; Nemirovski and Yudin, 1978, 1983; Shalev-Shwartz et al., 2007; Nemirovski et al., 2009; Hardt et al., 2016), is the workhorse for training supervised machine learning problems that have the generic form (1).

Step-size selection for SGD. The main parameter for guaranteeing the convergence of SGD is the *step-size* or the *learning rate*. In recent years, several ways of selecting the step-size have been proposed. Moulines and Bach (2011); Needell et al. (2016); Needell and Ward (2017); Nguyen et al. (2018); Gower et al. (2019) propose a non-asymptotic analysis of SGD with *constant step-size* for convex and strongly convex functions. For non-convex functions, such an analysis can be found in Ghadimi and Lan (2013); Bottou et al. (2018). Using a constant step-size for SGD guarantees convergence to a neighbourhoood of the solution. A common technique to guarantee convergence to the exact optimum is to use a *decreasing step-size* (Robbins and Monro, 1951; Ghadimi and Lan, 2013; Gower et al., 2019; Nemirovski et al., 2009; Karimi et al., 2016). More recently, *adaptive* methods (Duchi et al., 2011; Liu et al., 2019; Kingma and Ba, 2015; Bengio, 2015; Vaswani et al., 2019b; Li and Orabona, 2019; Ward et al., 2019) that adjust the step-size on the fly have become wide-spread and are particularly beneficial when training deep neural networks.

Contributions: Inspired by the classical Polyak step-size (Polyak, 1987) commonly used with the deterministic subgradient method (Hazan and Kakade, 2019; Boyd et al., 2003), we propose a novel adaptive learning rate for SGD. The proposed step-size is a natural extension of the Polyak step-size to the stochastic setting. We name it **stochastic Polyak step-size (SPS)**. Although computing SPS requires knowledge of the f_i^*; we argue that this information is readily available for modern machine learning applications (for example, $f_i^* = 0$ for most standard surrogate losses), making SPS an attractive choice for SGD.

In Section 3, we provide theoretical guarantees for the convergence of SGD with SPS in different scenarios including strongly convex, convex and non-convex smooth functions. Although SPS is provably larger than the typically used constant step-size, we guarantee its convergence to a reasonable neighborhood around the optimum. We note that in the modern machine learning tasks that we consider, it is enough to converge to a small neighbourhood and not the exact minimizer to get good generalization performance. We also establish a connection between SPS and the optimal step-size used in sketch and project methods for solving linear systems. Furthermore, in Appendix C, we provide convergence guarantees for convex non-smooth functions. We also show that by progressively increasing the batch-size for computing the stochastic gradients, SGD with SPS converges to the optimum.

Technical assumptions and challenges for proving convergence. Besides smoothness and convexity, several papers (Shamir and Zhang, 2013; Recht et al., 2011; Hazan and Kale, 2014; Rakhlin et al., 2012) assume that the variance of the stochastic gradient is bounded; that is there exists a c such that $\mathbb{E}_i \|\nabla f_i(x)\|^2 \leq c$. However, in the unconstrained setting, this assumption contradicts the assumption of strong convexity (Nguyen et al., 2018; Gower et al., 2019). In another line of work, growth conditions on the stochastic gradients have been used to guarantee convergence. In particular, the weak growth condition has been used in Bertsekas and Tsitsiklis (1996); Bottou et al. (2018); Nguyen et al. (2018). It states that there exist constants ρ, δ such that $\mathbb{E}_i \|\nabla f_i(x)\|^2 \leq \rho \, \mathbb{E}\|\nabla f(x)\|^2 + \delta$. Its stronger variant (strong growth condition) when $\delta = 0$ has been used in several recent papers (Schmidt and Roux, 2013; Cevher and Vũ, 2019; Vaswani et al., 2019a,b). These conditions can be relaxed to the expected smoothness assumption recently used in Gower et al. (2019).

Contributions: Our analysis of SGD with SPS does not require any of these additional assumptions for guaranteeing convergence[1]. We also note that our theoretical results do not require the finite-sum assumption and can be easily adapted to the streaming setting.

In addition, unlike standard analysis for constant step-size SGD, the use of SPS requires an adaptive step-size that uses the loss and stochastic gradient estimates at an iterate, resulting in correlations. One of the main technical challenges in the proofs is to carefully analyze the SGD iterates taking these correlations into account. Furthermore, since we need to be adaptive to the Lipschitz constant, we can not use the descent lemma (implied by smoothness and SGD update). This makes the convex proof more challenging than the standard analysis.

Novel analysis for constant SGD. In the existing analyses of constant step-size SGD, the neighborhood of convergence depends on the variance of the gradients at the optimum, $z^2 := \mathbb{E}_i \left\|\nabla f_i(x^*)\right\|^2$ which is assumed to be finite.

Contributions: The proposed analysis of SGD with SPS gives a novel way to analyze constant step-size SGD. In particular, we prove convergence of *constant step-size SGD* (without SPS), to a neighbourhood that depends on $\sigma^2 := f(x^*) - \mathbb{E}[f_i^*] < \infty$ (finite optimal objective difference).

Over-parametrized models and interpolation condition. Modern machine learning models such as non-parametric regression or over-parametrized deep neural networks are highly expressive and can fit or *interpolate* the training dataset completely (Zhang et al., 2017; Ma et al., 2018). In this setting, SGD with constant step-size can been shown to converge to the exact optimum at the deterministic rate (Schmidt and Roux, 2013; Ma et al., 2018; Vaswani et al., 2019a,b; Gower et al., 2019; Berrada et al., 2020).

Contributions: As a corollary of our theoretical results, we show that SPS is particularly effective under this interpolation setting. Specifically, we prove that SPS enables SGD to converge to the true solution at a fast rate matching the deterministic case. Moreover, SPS does not require the knowledge of any problem-dependent

[1]Except for our analysis for non-convex smooth functions where the weak growth condition is used.

constants or additional computational overhead.

Experimental Evaluation. In Section 4, we experimentally validate our theoretical results via experiments on synthetic datasets. We also evaluate the performance of SGD equipped with SPS relative to the state-of-the-art optimization methods when training over-parameterized models for deep matrix factorization, binary classification using kernels and multi-class classification using deep neural networks. For each of these tasks, we demonstrate the superior convergence of the proposed method. The code to reproduce our results can be found at https://github.com/IssamLaradji/sps.

2 SGD and the Stochastic Polyak Step-size

The optimization problem (1) can be solved using SGD:

$$x^{k+1} = x^k - \gamma_k \nabla f_i(x^k),$$

where example $i \in [n]$ is chosen uniformly at random and $\gamma_k > 0$ is the step-size in iteration k.

2.1 The Polyak step-size

Before explaining the proposed stochastic Polyak step-size, we first present the deterministic variant by Polyak (Polyak, 1987). This variant is commonly used in the analysis of deterministic subgradient methods (Boyd et al., 2003; Hazan and Kakade, 2019).

The deterministic Polyak step-size. For convex functions, the deterministic Polyak step-size at iteration k is the one that minimizes an upper-bound $Q(\gamma)$ on the distance of the iterate x_{k+1} to the optimal solution: $\|x^{k+1} - x^*\|_2^2 \leq Q(\gamma)$, where $Q(\gamma) = \|x^k - x^*\|^2 - 2\gamma \left[f(x^k) - f^*\right] + \gamma^2 \|g^k\|^2$. That is,

$$\gamma_k = \operatorname{argmin}_\gamma \left[Q(\gamma)\right] = \frac{f(x^k) - f^*}{\|g^k\|^2}.$$

Here g^k denotes a subgradient of function f at point x^k and f^* the optimum function value. For more details and a convergence analysis of the deterministic subgradient method, please check Appendix A.2. Note that the above step-size can be used only when the optimal value f^* is known, however Boyd et al. (2003) demonstrate that $f^* = 0$ for several applications (for example, finding a point in the intersection of convex sets, positive semidefinite matrix completion and solving convex inequalities).

Stochastic Polyak Step-size. It is clear that using the deterministic Polyak step-size in the update rule of SGD is impractical. It requires the computation of the function value f and its full gradient in each iteration.

To avoid this, we propose the stochastic Polyak step-size (SPS) for SGD:

$$\text{SPS:} \quad \boxed{\gamma_k = \frac{f_i(x^k) - f_i^*}{c \, \|\nabla f_i(x^k)\|^2}} \tag{2}$$

Note that SPS requires the evaluation of only the stochastic gradient $\nabla f_i(x^k)$ and of the function $f_i(x^k)$ at the current iterate (quantities that can be computed in the update rule of SGD without further cost). However, it requires the knowledge of f_i^*. An important quantity in the step-size is the parameter $0 < c \in \mathbb{R}$ which can be set theoretically based on the properties of the function under study. For example, for strongly convex functions, one should select $c = 1/2$ for optimal convergence.

In addition to SPS, in some of our convergence results we require its bounded variant:

$$\text{SPS}_{\max}: \quad \boxed{\gamma_k = \min\left\{\frac{f_i(x^k) - f_i^*}{c\|\nabla f_i(x^k)\|^2}, \gamma_{\mathsf{b}}\right\}} \tag{3}$$

Here $\gamma_{\mathsf{b}} > 0$ is a bound that restricts SPS from being very large and is essential to ensure convergence to a small neighborhood around the solution. If $\gamma_{\mathsf{b}} = \infty$ then $\text{SPS}_{\max}$ is equivalent to SPS.

Though SPS and $\text{SPS}_{\max}$ require knowledge of f_i^*, this information is often readily available. For machine learning problems using standard *unregularized* surrogate loss functions (e.g. squared loss for regression, logistic loss for classification), $f_i^* = 0$ (Bartlett et al., 2006). In the presence of an additional regularization term (e.g. ℓ_2 regularization), f_i^* can be obtained in closed form for these standard losses. We emphasize that since $f_i^* = \inf_x f_i(x)$, the functions f_i are not required to achieve the minimum. This is important when using loss functions such as the logistic loss for which the infimum is achieved at infinity (Soudry et al., 2018). Furthermore, we note that the deterministic Polyak step-size requires knowledge of f^* which is a much stronger assumption than the knowledge of f_i^*.

Closely related work. We now briefly compare against the recently proposed stochastic variants of the Polyak step-size (Rolinek and Martius, 2018; Oberman and Prazeres, 2019; Berrada et al., 2020). In Section 3, we present a detailed comparison of the theoretical convergence rates.

In Rolinek and Martius (2018), the *L4 algorithm* has been proposed showing that a stochastic variant of the Polyak step for SGD achieves good empirical results for training neural networks. However it has *no theoretical* *convergence guarantees.* The step-size is very similar to SPS (2) but each update requires an online estimation of the f_i^* which does not result in robust empirical performance and *requires up to three hyper-parameters.*

Oberman and Prazeres (2019) use a different variant of the stochastic Polyak step-size: $\gamma_k = \frac{2[f(x^k)-f^*]}{\mathbb{E}_i\|\nabla f_i(x^k)\|^2}$. This step-size requires knowledge of the quantity $\mathbb{E}_i\|\nabla f_i(x^k)\|^2$ for all iterates x^k and the evaluation of $f(x^k)$ in each step, making it impractical for finite-sum problems with large n. Moreover, their theoretical results focus only on strongly convex smooth functions.

In the ALI-G algorithm proposed by Berrada et al. (2020), the step-size is set as: $\gamma_k = \min\left\{\frac{f_i(x^k)}{\|\nabla f_i(x^k)\|^2+\delta}, \eta\right\}$, where $\delta > 0$ is a positive constant. Unlike our setting, their theoretical analysis relies on an ϵ-interpolation condition. Moreover, the values of the parameter δ and η that guarantee convergence heavily depend on the smoothness parameter of the objective f, limiting the method's practical applicability. In Section 3, we show that as compared to Berrada et al. (2020), the proposed method results in both better rates and a smaller neighborhood of convergence. For the case of over-parameterized models, our step-size selection guarantees convergence to the exact solution while the step proposed in Berrada et al. (2020) finds only an approximate solution that could be δ away from the optimum. In Section 4, we also experimentally show that $\text{SPS}_{\max}$ results in better convergence than ALI-G.

2.2 Optimal Objective Difference

Unlike the typical analysis of SGD that assumes a finite gradient noise $z^2 := \mathbb{E}[\|\nabla f_i(x^*)\|^2]$, in all our results, we assume a finite optimal objective difference.

Assumption 2.1 (Finite optimal objective difference)**.**

$$\sigma^2 := \mathbb{E}_i[f_i(x^*) - f_i^*] = f(x^*) - \mathbb{E}_i[f_i^*] < \infty \quad (4)$$

This is a very weak assumption. Moreover when (1) is the training problem of an over-parametrized model such as a deep neural network or involves solving a consistent linear system or classification on linearly separable data, each individual loss function f_i attains its minimum at x^*, and thus $f_i(x^*) - f_i^* = 0$. In this *interpolation* setting, it follows that $\sigma = 0$.

3 Convergence Analysis

In this section, we present the main convergence results. For the formal definitions and properties of functions see Appendix A.1. Proofs of all key results can be found in the Appendix B.

3.1 Upper and Lower Bounds of SPS

If a function g is μ-strongly convex and L-smooth the following bounds hold: $\frac{1}{2L}\|\nabla g(x)\|^2 \leq g(x) - \inf_x g(x) \leq \frac{1}{2\mu}\|\nabla g(x)\|^2$. Using these bounds and by assuming that the functions f_i in problem (1) are μ_i-strongly convex and L_i-smooth, it is straight forward to see that SPS can be lower and upper bounded as follows:

$$\frac{1}{2cL_{\max}} \leq \frac{1}{2cL_i} \leq \gamma_k = \frac{f_i(x^k) - f_i^*}{c\|\nabla f_i(x^k)\|^2} \leq \frac{1}{2c\mu_i}, \quad (5)$$

where $L_{\max} = \max\{L_i\}_{i=1}^n$.

3.2 Sum of convex functions: strongly convex objective

In this section, we assume that all components f_i are convex functions and that the objective function f is μ-strongly convex.

Theorem 3.1. Let f_i be L_i-smooth convex functions and assume that the objective function f is μ-strongly convex function. Then, SGD with $\text{SPS}_{\max}$ with $c \geq 1/2$ converges as:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1-\mu\alpha)^k \|x^0 - x^*\|^2 + \frac{2\gamma_b \sigma^2}{\mu\alpha}, \quad (6)$$

where $\alpha := \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$ and $L_{\max} = \max\{L_i\}_{i=1}^n$ is the maximum smoothness constant. The best convergence rate and the tightest neighborhood are obtained for $c = 1/2$.

Note that in Theorem 3.1, we do not make any assumption on the value of the upper bound γ_b. However, it is clear that for convergence to a small neighborhood of the solution x^* (unique solution for strongly convex functions) γ_b should not be very large[2].

Another important aspect of Theorem 3.1 is that it provides convergence guarantees without requiring strong assumptions like bounded gradients or growth conditions. We do not use these conditions because SPS provides a natural bound on the norm of the gradients. In the following corollaries we make additional assumptions to better understand the convergence of SGD with $\text{SPS}_{\max}$.

In our first corollary, we assume that our model is able to interpolate the data (each individual loss function f_i attains its minimum at x^*). This condition is satisfied for unregularized least-squares regression on a realizable dataset, or when using the squared-hinge loss on a linearly-separable dataset. The interpolation assumption enables us to guarantee the convergence of SGD

[2]Note that neighborhood $\frac{2\gamma_b\sigma^2}{\mu\alpha}$ has γ_b in the numerator and for the case of large γ_b, $\alpha = \frac{1}{2cL_{\max}}$.

with SPS, without an upper-bound on the step-size ($\gamma_b = \infty$).

Corollary 3.2. Assume interpolation ($\sigma = 0$) and let all assumptions of Theorem 3.1 be satisfied. SGD with SPS with $c = 1/2$ converges as:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \left(1 - \frac{\mu}{L_{\max}}\right)^k \|x^0 - x^*\|^2.$$

We compare the convergence rate in Corollary 3.2 to that of stochastic line search (SLS) proposed in Vaswani et al. (2019b). In similar setting, SLS achieves the slower linear rate $\max\left\{1 - \frac{\bar{\mu}}{L_{\max}}, 1 - \gamma_b \bar{\mu}\right\}$, where $\bar{\mu} = \sum_{i=1}^n \mu_i/n$ is the average strong-convexity of the finite sum. In particular, according to Theorem 1 of Vaswani et al. (2019b), the convergence of SLS requires that at least one of the f_i's is μ_i-strongly convex implying that the objective function f is strongly convex. This is a stronger assumption than the one we have in Theorem 3.1. We also note that $\bar{\mu} \leq \mu$.

In Berrada et al. (2020), ALI-G is analyzed under the strong assumption that all functions f_i are μ-strongly convex and L-smooth. For detailed comparison of SPS with ALI-G, see Appendix B.1.1.

An interesting outcome of Theorem 3.1 is a novel analysis for SGD with a constant step-size. In particular, note that if the bound in $\text{SPS}_{\max}$ is selected to be $\gamma_b \leq \frac{1}{2cL_{\max}}$, then using the lower bound of (5), it can be easily shown that our method reduces to SGD with constant step-size $\gamma_k = \gamma = \gamma_b \leq \frac{1}{2cL_{\max}}$. In this case, we obtain the following convergence rate.

Corollary 3.3. Let all assumptions of Theorem 3.1 be satisfied. SGD with $\text{SPS}_{\max}$ with $c = 1/2$ and $\gamma_b \leq \frac{1}{L_{\max}}$ becomes SGD with constant step-size $\gamma \leq \frac{1}{L_{\max}}$ and converges as:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \mu\gamma)^k \|x^0 - x^*\|^2 + \frac{2\sigma^2}{\mu}.$$

If we further assume interpolation ($\sigma = 0$), the iterates of SGD with constant step-size $\gamma \leq \frac{1}{L_{\max}}$ satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \mu\gamma)^k \|x^0 - x^*\|^2.$$

To the best of our knowledge, this is the first result that shows convergence of constant step-size SGD to a neighborhood that depends on the optimal objective difference σ^2 (4) and not on the variance $z^2 = \mathbb{E}[\|\nabla f_i(x^*)\|^2$. Note that if we assume that all function f_i are μ-strongly convex and L-smooth functions then the two notions of variance satisfy the following connection: $\frac{1}{2L}z^2 \leq \sigma^2 \leq \frac{1}{2\mu}z^2$.

3.3 Sum of convex functions

Here, we derive the convergence rate when all component functions f_i are convex without any strong convexity and obtain the following theorem.

Theorem 3.4. Assume that f_i are convex, L_i-smooth functions. SGD with $\text{SPS}_{\max}$ with $c = 1$ converges as:

$$\mathbb{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{\|x^0 - x^*\|^2}{\alpha\, K} + \frac{2\sigma^2\gamma_b}{\alpha}.$$

Here $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}$ and $\bar{x}^K = \frac{1}{K}\sum_{k=0}^{K-1} x^k$.

Analogous to the strongly-convex case, the size of the neighbourhood is proportional to γ_b. When interpolation is satisfied and $\sigma = 0$, we observe that the unbounded variant of SPS with $\gamma_b = \infty$ converges to the optimum at a $O(1/K)$ rate. This rate is faster than the rates in Vaswani et al. (2019b); Berrada et al. (2020) and we refer the reader to the Appendix for a detailed comparison. As in the strongly-convex case, by setting $\gamma_b \leq \frac{1}{2cL_{max}}$, we obtain the convergence rate obtained by constant step-size SGD.

3.4 Consistent Linear Systems

In Richtárik and Takác (2020), given the consistent linear system $\mathbf{A}x = b$, the authors provide a stochastic optimization reformulation of the form (1) which is equivalent to the linear system in the sense that their solution sets are identical. That is, the set of minimizers of the stochastic optimization problem $\mathcal{X}^*$ is equal to the set of solutions of the stochastic linear system $\mathcal{L} := \{x : \mathbf{A}x = b\}$. An interesting property of this stochastic optimization problem is that[3]: $f_i(x) - f_i^* \overset{f_i^*=0}{=} f_i(x) = \frac{1}{2}\|\nabla f_i(x)\|^2 \quad \forall x \in \mathbb{R}^d$. Using the special structure of the problem, SPS (2) with $c = 1/2$ takes the following form: $\gamma_k \overset{(2)}{=} \frac{2\left[f_i(x^k) - f_i^*\right]}{\|\nabla f_i(x^k)\|^2} = 1$, which is the theoretically optimal constant step-size for SGD in this setting (Richtárik and Takác, 2020). This reduction implies that SPS results in an optimal convergence rate when solving consistent linear systems. We provide the convergence rate for SPS in this setting in Appendix B.

3.5 Sum of non-convex functions: PL Objective

We first focus on a special class of non-convex functions that satisfy the Polyak-Lojasiewicz (PL) condition (Polyak, 1987). The PL inequality is a generalization of strong-convexity and is satisfied for matrix

[3]For more details on the stochastic reformulation problem and its properties see Appendix B.3.

factorization (Sun and Luo, 2016) or when minimizing the logistic loss on a compact set (Karimi et al., 2016). In particular, we assume that function f satisfies the PL condition but do not assume convexity of the component functions f_i.

Definition 3.5 (Polyak-Lojasiewicz (PL) condition)**.** We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the PL condition if there exists $\mu > 0$ such that, $\forall x \in \mathbb{R}^n$:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \tag{7}$$

Theorem 3.6. Assume that function f satisfies the PL condition (7), and let f and f_i be smooth functions. SGD with $\text{SPS}_{\max}$ with $c > \frac{L_{\max}}{4\mu}$ and $\gamma_b \geq \frac{1}{2cL_{\max}}$ converges as:

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \nu^k \left[f(x^0) - f(x^*)\right] + \frac{L\sigma^2\gamma_b}{2(1-\nu)\,c}$$

where $\nu = \gamma_b \left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right) \in (0, 1]$ and $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}$.

Under the interpolation setting, $\sigma = 0$, and $\text{SPS}_{\max}$ converges to the optimal solution at a linear rate. If $\gamma_b \leq \min\left\{\frac{1}{2cL_{\max}}, \frac{2c}{4\mu c - L_{\max}}\right\}$ using the lower bound in (5), the analyzed method is SGD with constant step-size and we obtain the following corollary.

Corollary 3.7. Assume that f satisfies the PL condition (7), and let f and f_i be smooth functions. SGD with constant step-size $\gamma_k = \gamma \leq \frac{\mu}{L_{\max}^2}$ converges as:

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \nu^k \left[f(x^0) - f(x^*)\right] + \frac{L\sigma^2\gamma}{2(1-\nu)\,c}.$$

To the best of our knowledge this is the first result for the convergence of SGD for PL functions without assuming bounded gradient or bounded variance or interpolation (for more details see results in Karimi et al. (2016) and discussion in Gower et al. (2019)). In the interpolation case, we obtain linear convergence to the optimum with a constant step-size equal to that used in Vaswani et al. (2019a); Lei et al. (2019).

3.6 General Non-Convex Functions

In this section, we assume a common condition used to prove convergence of SGD in the non-convex setting (Bottou et al., 2018).

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq \rho\|\nabla f(x)\|^2 + \delta \tag{8}$$

where $\rho, \delta > 0$ constants.

Theorem 3.8. Let f and f_i be smooth functions and assume that there exist $\rho, \delta > 0$ such that the condition (8) is satisfied. SGD with $\text{SPS}_{\max}$ with $c > \frac{\rho L}{4L_{\max}}$ and $\gamma_b < \max\left\{\frac{2}{L\rho}, \bar{\gamma_b}\right\}$ converges as:

$$\min_{k\in[K]} \mathbb{E}\|\nabla f(x^k)\|^2 \leq \frac{2}{\zeta K}\left(f(x^0) - f(x^*)\right) + \frac{\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta}{\zeta},$$

where $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}$, $\zeta = (\gamma_b + \alpha) - \rho\left(\gamma_b - \alpha + L\gamma_b^2\right)$ and

$$\bar{\gamma_b} := \frac{-(\rho-1) + \sqrt{(\rho-1)^2 + \frac{4L\rho(\rho+1)}{2cL_{\max}}}}{2L\rho}.$$

From the above theorem, we observe that SGD with SPS results in $O(1/K)$ convergence to a neighborhoud governed by δ. For the case that $\delta = 0$, condition (8) reduces to the strong growth condition (SGC) used in several recent papers (Schmidt and Roux, 2013; Vaswani et al., 2019b,a). It can be easily shown that functions that satisfy the SGC condition necessarily satisfy the interpolation property (Vaswani et al., 2019a). In the special case of interpolation, SGD with SPS is able to find a first-order stationary point as efficiently as deterministic gradient descent. Moreover, for $c \in \left(\frac{\rho L}{4L_{\max}}, \frac{\rho L}{2L_{\max}}\right]$, the lower bound $\frac{1}{2cL_{\max}}$ of SPS lies in the range $\left[\frac{1}{\rho L}, \frac{2}{\rho L}\right)$ and thus the step-size is larger than $\frac{1}{\rho L}$, the best constant step-size analyzed in this setting (Vaswani et al., 2019a).

3.7 Additional Convergence Results

In Appendix C, we prove a $O(1/\sqrt{K})$ convergence rate for non-smooth convex functions. Furthermore, similar to Schmidt et al. (2011), we propose a method to increase the mini-batch size for evaluating the stochastic gradient and guarantee convergence to the optimal solution without interpolation.

4 Experimental Evaluation

We validate our theoretical results using synthetic experiments in Section 4.1. In Section 4.2, we evaluate the performance of SGD with SPS when training over-parametrized models. In particular, we compare against state-of-the-art optimization methods for deep matrix factorization, binary classification using kernel methods and multi-class classification using standard deep neural network models.

4.1 Synthetic experiments

We use a synthetic dataset to validate our theoretical results. Following the procedure outlined in Nutini et al. (2017), we generate a sparse dataset for binary classification with the number of examples $n = 1$k and dimension $d = 100$. We use the logistic loss with and without ℓ_2 regularization. The data is generated to ensure that the function f is strongly convex in both cases. We evaluate the performance of $\text{SPS}_{\max}$ and set set $c = 1/2$ as suggested by Theorem 3.1. We experiment with three values of $\gamma_b = \{1, 5, 100\}$. In the regularized case, f_i^* can be pre-computed in closed form for each i using the Lambert W function (Corless et al., 1996) (see Appendix D); while f_i^* is simply zero in the unregularized case. A similar observation has been used to construct a "truncated" model for improving the robustness of gradient descent in Asi and Duchi (2019). In both cases, we benchmark the performance of SPS against constant step-size SGD with $\gamma = \{0.1, 0.01\}$. From Figure 1, we observe that constant step-size SGD is not robust to the step-size; it has good convergence with step-size 0.1, slow convergence when using a step-size of 0.01 and we observe divergence for larger step-sizes. In contrast, all the variants of SPS converge to a neighbourhood of the optimum and the size of the neighbourhood increases as γ_b increases as predicted by the theory.

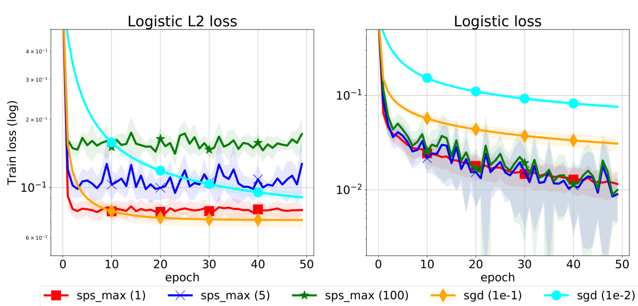

Figure 1: Synthetic experiment to benchmark SPS against constant step-size SGD for binary classification using the (left) regularized and (right) unregularized logistic loss.

4.2 Experiments for over-parametrized models

In this section, we consider training over-parameterized models that (approximately) satisfy the interpolation condition. Following the logic of the previous section, we evaluate the performance of both the SPS and $\text{SPS}_{\max}$ variants with $f_i^* = 0$. Throughout our experiments, we found that SPS without an upper-bound on the step-size is not robust to the misspecification of interpolation and results in large fluctuations when interpolation is not exactly satisfied. For $\text{SPS}_{\max}$, the value of γ_b that results in good convergence depends on the problem and requires careful parameter tuning. This is also evidenced by the highly variable performance of ALI-G (Berrada et al., 2020) that uses a constant upper-bound on the step-size. To alleviate this problem, we use a *smoothing* procedure that prevents large fluctuations in the step-size across iterations. This can be viewed as using an adaptive iteration-dependent upper-bound γ_b^k where $\gamma_b^k = \tau^{b/n} \gamma^{k-1}$. Here, τ is a tunable hyper-parameter set to 2 in all our experiments, b is the batch-size and n is the number of examples. We note that using an adaptive γ_b can be easily handled by our theoretical results. A similar smoothing procedure has been used to control the magnitude of the step-sizes when using the Barzilai-Borwein step-size selection procedure for SGD (Tan et al., 2016) and is related to the "reset" option for using larger step-sizes in Vaswani et al. (2019b). We set $c = 1/2$ for binary classification using kernels (convex case) and deep matrix factorization (non-convex PL case). For multi-class classification using deep networks, we empirically find that any value of $c \geq 0.2$ results in convergence. In this case, we observed that across models and datasets, the fastest convergence is obtained with $c = 0.2$ and use this value.

We compare our methods against Adam (Kingma and Ba, 2015), which is the most common adaptive method, and other recent methods that report better performance than Adam: (i) stochastic line-search (SLS) (Vaswani et al., 2019b) (ii) ALI-G (Berrada et al., 2020)[4] (iii) rectified Adam (RADAM) (Liu et al., 2019) (iv) Look-ahead optimizer (Zhang et al., 2019). We use the default learning rates and momentum (non-zero) parameters and the publicly available code for the competing methods. All our results are averaged across 5 independent runs.

Deep matrix factorization. In the first experiment, we use deep matrix factorization to examine the effect of over-parametrization for the different optimizers. In particular, we solve the non-convex regression problem: $\min_{W_1, W_2} \mathbb{E}_{x \sim N(0,I)} \|W_2 W_1 x - Ax\|^2$ and use the experimental setup in Rolinek and Martius (2018); Vaswani et al. (2019b); Rahimi and Recht (2017). We choose $A \in \mathbb{R}^{10 \times 6}$ with condition number $\kappa(A) = 10^{10}$ and generate a fixed dataset of 1000 samples. We control the degree of over-parametrization via the rank k of the matrix factors $W_1 \in \mathbb{R}^{k \times 6}$ and

[4] With ALI-G we refer to the method analyzed in Berrada et al. (2020). This is SGD with step-size the one described in Section 2. We highlight that the experiments in Berrada et al. (2020) used momentum on top of the analyzed method but without any convergence quarantees. To ensure a fair comparison with SPS, we do not use such momentum.

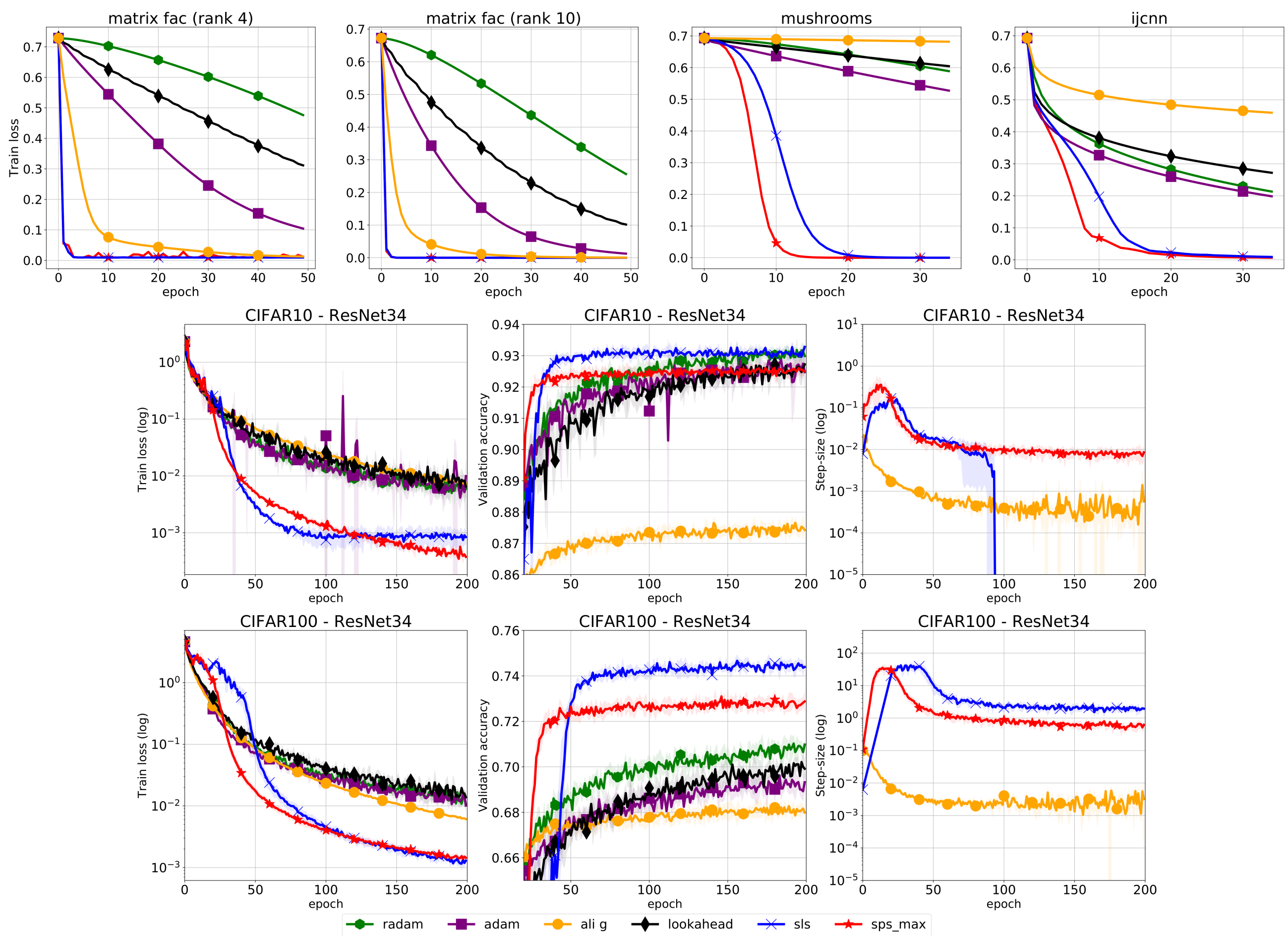

Figure 2: Comparing the performance of optimizers on deep matrix factorization (top left) and binary classification using kernels (top right) and multi-class classification on CIFAR-10 and CIFAR-100 with ResNet34.

$W_2 \in \mathbb{R}^{10\times k}$. In Figure 2, we show the training loss as we vary the rank $k \in \{4, 10\}$ (additional experiments are in Appendix E). For $k = 4$, the interpolation condition is *not* satisfied, whereas it is exactly satisfied for $k = 10$. We observe that (i) SPS is robust to the degree of over-parametrization and (ii) has performance equal to that of SLS. However, note that SPS does not require the expensive back-tracking procedure of SLS and is arguably simpler to implement.

Binary classification using kernels. Next, we compare the optimizers' performance in the convex, interpolation regime. We consider binary classification using RBF kernels, using the logistic loss without regularization. The bandwidths for the RBF kernels are set according to the validation procedure described in Vaswani et al. (2019b). We experiment with four standard datasets: mushrooms, rcv1, ijcnn, and w8a from LIBSVM (Chang and Lin, 2011). Figure 2 shows the training loss on the mushrooms and ijcnn for the different optimizers. Again, we observe the strong performance of SPS compared to the other optimizers.

Multi-class classification using deep networks. We benchmark the convergence rate and generalization performance of SPS methods on standard deep learning experiments. We consider non-convex minimization for multi-class classification using deep network models on the CIFAR10 and CIFAR100 datasets. Our experimental choices follow the setup in Luo et al. (2019). For CIFAR10 and CIFAR100, we experiment with the standard image-classification architectures: ResNet-34 (He et al., 2016) and DenseNet-121 (Huang et al., 2017). For space concerns, we report only the ResNet experiments in the main paper and relegate the DenseNet and MNIST experiments to Appendix E. From Figure 2, we observe that SPS results in the best training loss across models and datasets. For CIFAR-10, SPS results in competitive generalization performance compared to the other optimizers, whereas for CIFAR-100, its generalization performance is better than all optimizers except SLS. Note that ALI-G, the closest related optimizer results in worse generalization performance in all cases. We note that SPS is able to match the performance of SLS, but does not require an expensive back-tracking line-search or additional tricks.

For this set of experiments, we also plot how the step-size varies across iterations for SLS, SPS and ALI-G. Interestingly, for both CIFAR-10 and CIFAR-100, we find that step-size for both SPS and SLS follows a cyclic behaviour - a warm-up period where the step-size first increases and then decreases to a constant value. Such a step-size schedule has been empirically found to result in good training and generalization performance (Loshchilov and Hutter, 2017) and our results show that SPS is able to simulate this behaviour.

5 Conclusion

We proposed and theoretically analyzed a stochastic variant of the classical the Polyak step-size. We quantified the convergence rate of SPS in numerous settings and used our analysis techniques to prove new results for constant step-size SGD. Furthermore, via experiments on a variety of tasks we showed the strong performance of SGD with SPS as compared to state-of-the-art optimization methods. There are many possible interesting extensions of our work: using SPS with accelerated methods, studying the effect of mini-batching and non-uniform sampling techniques and extensions to the distributed and decentralized settings.

Acknowledgements

Nicolas Loizou and Sharan Vaswani acknowledge support by the IVADO Postdoctoral Funding Program. Issam Laradji is funded by the UBC Four-Year Doctoral Fellowships (4YF). This research was partially supported by the Canada CIFAR AI Chair Program and by a Google Focused Research award. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program.

The authors would like to thank Frederik Kunstner for help with the convex proofs, and Aaron Defazio for fruitful discussions and feedback on the manuscript.

References

Asi, H. and Duchi, J. C. (2019). The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bengio, Y. (2015). Rmsprop and equilibrated adaptive learning rates for nonconvex optimization. *corr abs/1502.04390*.

Berrada, L., Zisserman, A., and Kumar, M. P. (2020). Training neural networks for and by interpolation. In *ICML*.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, 1st edition.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Boyd, S., Xiao, L., and Mutapcic, A. (2003). Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004:2004–2005.

Cevher, V. and Vũ, B. C. (2019). On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational mathematics*, 5(1):329–359.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159.

Ghadimi, S. and Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Gower, R. and Richtárik, P. (2015). Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General analysis and improved rates. In *ICML*.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: stability of stochastic gradient descent. In *ICML*.

Harikandeh, R., Ahmed, M. O., Virani, A., Schmidt, M., Konečnỳ, J., and Sallinen, S. (2015). Stop wasting my gradients: Practical SVRG. In *NeurIPS*.

Hazan, E. and Kakade, S. (2019). Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*.

Hazan, E. and Kale, S. (2014). Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR*.

Kaczmarz, S. (1937). Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de lAcademie Polonaise des Sciences et des Lettres*, 35:355–357.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *ECML-PKDD*.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Lei, Y., Hu, T., Li, G., and Tang, K. (2019). Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10):4394–4400.

Li, X. and Orabona, F. (2019). On the convergence of stochastic gradient descent with adaptive stepsizes. In *AISTATS*.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Lohr, S. L. (2019). *Sampling: Design and Analysis: Design and Analysis*. Chapman and Hall/CRC.

Loizou, N. (2019). *Randomized iterative methods for linear systems: momentum, inexactness and gossip*. PhD thesis, University of Edinburgh.

Loizou, N. and Richtárik, P. (2019). Revisiting randomized gossip algorithms: General framework, convergence rates and novel block and accelerated protocols. *arXiv preprint arXiv:1905.08645*.

Loizou, N. and Richtárik, P. (2020a). Convergence analysis of inexact randomized iterative methods. *SIAM Journal on Scientific Computing*, 42(6):A3979–A4016.

Loizou, N. and Richtárik, P. (2020b). Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. *ICLR*.

Luo, L., Xiong, Y., Liu, Y., and Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. In *ICLR*.

Ma, S., Bassily, R., and Belkin, M. (2018). The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *ICML*.

Mező, I. and Baricz, Á. (2017). On the generalization of the Lambert W function. *Transactions of the American Mathematical Society*, 369(11):7917–7934.

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NeurIPS*.

Needell, D., Srebro, N., and Ward, R. (2016). Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming, Series A*, 155(1):549–573.

Needell, D. and Ward, R. (2017). Batched stochastic gradient descent with weighted sampling. In *Approximation Theory XV, Springer*, volume 204 of *Springer Proceedings in Mathematics & Statistics,*, pages 279 – 306.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nemirovski, A. and Yudin, D. B. (1978). On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathetmatics Doklady*, 19.

Nemirovski, A. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley Interscience.

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. (2018). SGD and hogwild! Convergence without the bounded gradients assumption. In *ICML*.

Nutini, J., Laradji, I., and Schmidt, M. (2017). Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*.

Oberman, A. M. and Prazeres, M. (2019). Stochastic gradient descent with polyak's learning rate. *arXiv preprint arXiv:1903.08688*.

Polyak, B. (1987). Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*.

Rahimi, A. and Recht, B. (2017). Reflections on random kitchen sinks - arg min blog.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*.

Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NeurIPS*.

Richtárik, P. and Takác, M. (2020). Stochastic reformulations of linear systems: algorithms and convergence theory. *SIAM Journal on Matrix Analysis and Applications*, 41(2):487–524.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Rolinek, M. and Martius, G. (2018). L4: Practical loss-based stepsize adaptation for deep learning. In *NeurIPS*.

Schmidt, M., Kim, D., and Sra, S. (2011). 11 projected Newton-type methods in machine learning. *Optimization for Machine Learning*, page 305.

Schmidt, M. and Roux, N. (2013). Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007). Pegasos: primal estimated subgradient solver for SVM. In *ICML*.

Shamir, O. and Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

Strohmer, T. and Vershynin, R. (2009). A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278.

Sun, R. and Luo, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579.

Tan, C., Ma, S., Dai, Y.-H., and Qian, Y. (2016). Barzilai-borwein step size for stochastic gradient descent. In *NeurIPS*.

Vaswani, S., Bach, F., and Schmidt, M. (2019a). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *AISTATS*.

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., and Lacoste-Julien, S. (2019b). Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *NeurIPS*.

Ward, R., Wu, X., and Bottou, L. (2019). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *ICML*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *ICLR*.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. (2019). Lookahead optimizer: k steps forward, 1 step back. In *NeurIPS*.

Supplementary Material
Stochastic Polyak Step-size for SGD:
An Adaptive Learning Rate for Fast Convergence

The Supplementary Material is organized as follows: In Section A, we provide the basic definitions mentioned in the main paper. We also present the convergence of deterministic subgradient method with the classical Polyak step-size. In Section B we present the proofs of the main theorems and in Section C we provide additional convergence results. In Section D, we provide the closed form solutions for f_i* for standard regularized binary surrogate losses. Finally, additional numerical experiments are presented in Section E.

A Technical Preliminaries

A.1 Basic Definitions

Let us present some basic definitions used throughout the paper.

Definition A.1 (Strong Convexity / Convexity)**.** The function $f : \mathbb{R}^n \to \mathbb{R}$, is μ-strongly convex, if there exists a constant $\mu > 0$ such that $\forall x, y \in \mathbb{R}^n$:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \tag{9}$$

for all $x \in \mathbb{R}^d$. If inequality (9) holds with $\mu = 0$ the function f is convex.

Definition A.2 (Polyak-Lojasiewicz Condition)**.** The function $f : \mathbb{R}^n \to \mathbb{R}$, satisfies the Polyak-Lojasiewicz (PL) condition, if there exists a constant $\mu > 0$ such that $\forall x \in \mathbb{R}^n$:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \tag{10}$$

Definition A.3 (L-smooth)**.** The function $f : \mathbb{R}^n \to \mathbb{R}$, L-smooth, if there exists a constant $L > 0$ such that $\forall x, y \in \mathbb{R}^n$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{11}$$

or equivalently:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 \tag{12}$$

A.2 The Deterministic Polyak step-size

In this section we describe the Polyak step-size for the subgradient method as presented in Polyak (1987) for solving $\min_{x \in \mathbb{R}^d} f(x)$ where f is convex, not necessarily smooth function.

Consider the subgradient method:

$$x^{k+1} = x^k - \gamma_k g^k,$$

where γ_k is the step-size (learning rate) and g^k is any subgradient of function f at point x^k.

Theorem A.4. Let f be convex function. Let $\gamma_k = \frac{f(x^k) - f(x^*)}{\|g^k\|^2}$ be the step-size in the update rule of subgradient method. Here $f(x^*)$ denotes the optimum value of function f. Let $G > 0$ such that $\|g^k\|^2 < G^2$. Then,

$$f_*^k - f(x^*) \leq \frac{G\|x^0 - x^*\|}{\sqrt{k+1}} = O\left(\frac{1}{\sqrt{k}}\right),$$

where $f_*^k = \min\{f(x^i) : i = 0, 1, \ldots, k\}$.

Proof.

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \gamma_k g^k - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, g^k \rangle + \gamma_k^2 \|g^k\|^2 \\ &\leq \|x^k - x^*\|^2 - 2\gamma_k \left[f(x^k) - f(x^*)\right] + \gamma_k^2 \|g^k\|^2 \end{aligned} \tag{13}$$

where the last line follows from the definition of subgradient:

$$f(x^*) \geq f(x^k) + \langle x^k - x^*, g^k \rangle$$

Polyak suggested to use the step-size:

$$\boxed{\gamma_k = \frac{f(x^k) - f(x^*)}{\|g^k\|^2}} \tag{14}$$

which is precisely the step-size that minimize the right hand side of (13). That is,

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\|g^k\|^2} = \text{argmin}_{\gamma_k} \left[\|x^k - x^*\|^2 - 2\gamma_k \left[f(x^k) - f(x^*)\right] + \gamma_k^2 \|g^k\|^2\right]$$

. By using this choice of step-size in (13) we obtain:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - 2\gamma_k \left[f(x^k) - f(x^*)\right] + \gamma_k^2 \|g^k\|^2 \\ &\overset{(14)}{=} \|x^k - x^*\|^2 - \frac{\left[f(x^k) - f(x^*)\right]^2}{\|g^k\|^2} \end{aligned} \tag{15}$$

From the above note that $\|x^k - x^*\|^2$ is monotonic function. Now using telescopic sum and by assuming $\|g^k\|^2 < G^2$ we obtain:

$$\|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2 - \frac{1}{G^2} \sum_{i=0}^{k} \left[f(x^i) - f(x^*)\right]^2 \tag{16}$$

Thus,

$$\frac{1}{G^2} \sum_{i=0}^{k} \left[f(x^i) - f(x^*)\right]^2 \leq \|x^0 - x^*\|^2 - \|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2$$

Let us define $f_*^k = \min\{f(x^i) : i = 0, 1, \ldots, k\}$ then: $[f_*^k - f(x^*)]^2 \leq \frac{G^2\|x^0 - x^*\|^2}{k+1}$ and

$$f_*^k - f(x^*) \leq \frac{G\|x^0 - x^*\|}{\sqrt{k+1}} = O\left(\frac{1}{\sqrt{k}}\right)$$

□

For more details and slightly different analysis check Polyak (1987) and Boyd et al. (2003). In Hazan and Kakade (2019) similar analysis to the above have been made for the deterministic gradient descent ($g^k = \nabla f(x^k)$) under several assumptions. (convex, strongly convex , smooth).

B Proofs of Main Results

In this section we present the proofs of the main theoretical results presented in the main paper. That is, the convergence analysis of SGD with $\text{SPS}_{\max}$ and SPS under different combinations of assumptions on functions f_i and f of Problem (1).

First note that the following inequality can be easily obtained by the definition of $\text{SPS}_{\max}$ (3):

$$\gamma_k^2 \|\nabla f_i(x^k)\|^2 \leq \frac{\gamma_k}{c}\left[f_i(x^k) - f_i^*\right] \tag{17}$$

We use the above inequality in several parts of our proofs. It is the reason that we are able to obtain an upper bound of $\gamma_k^2\|\nabla f_i(x^k)\|^2$ without any further assumptions. For the case of SPS (2), inequality (17) becomes equality.

B.1 Proof of Theorem 3.1

Proof.

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - \gamma_k \nabla f_i(x^k) - x^*\|^2 \\
&= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k)\rangle + \gamma_k^2 \|\nabla f_i(x^k)\|^2 \\
&\overset{(17)}{\leq} \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k)\rangle + \frac{\gamma_k}{c}\left[f_i(x^k) - f_i^*\right] \\
&\overset{c\geq 1/2}{\leq} \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k)\rangle + 2\gamma_k\left[f_i(x^k) - f_i^*\right] \\
&= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k)\rangle + 2\gamma_k\left[f_i(x^k) - f_i(x^*) + f_i(x^*) - f_i^*\right] \\
&= \|x^k - x^*\|^2 + 2\gamma_k\left[-\langle x^k - x^*, \nabla f_i(x^k)\rangle + f_i(x^k) - f_i(x^*)\right] + 2\gamma_k\left[f_i(x^*) - f_i^*\right]
\end{aligned}$$

From convexity of functions f_i it holds that $-\langle x^k - x^*, \nabla f_i(x^k)\rangle + f_i(x^k) - f_i(x^*) \leq 0$, $\forall i \in [n]$. Thus,

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 + 2\gamma_k \underbrace{\left[-\langle x^k - x^*, \nabla f_i(x^k)\rangle + f_i(x^k) - f_i(x^*)\right]}_{\leq 0} + 2\gamma_k \underbrace{\left[f_i(x^*) - f_i^*\right]}_{\geq 0} \\
&\overset{(5),(3)}{\leq} \|x^k - x^*\|^2 + 2\min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\left[-\langle x^k - x^*, \nabla f_i(x^k)\rangle + f_i(x^k) - f_i(x^*)\right] \\
&\quad + 2\gamma_b\left[f_i(x^*) - f_i^*\right]
\end{aligned}$$

By taking expectation condition on x^k

$$\begin{aligned}
\mathbb{E}_i\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 + 2\min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\left[-\langle x^k - x^*, \nabla f(x^k)\rangle + f(x^k) - f(x^*)\right] \\
&\quad + 2\gamma_b \mathbb{E}_i\left[f_i(x^*) - f_i^*\right] \\
&\overset{(4)}{=} \|x^k - x^*\|^2 + 2\min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\left[-\langle x^k - x^*, \nabla f(x^k)\rangle + f(x^k) - f(x^*)\right] \\
&\quad + 2\gamma_b\sigma^2
\end{aligned}$$

From strong convexity of the objective function f we have that $f(x^k) - f(x^*) - \langle x^k - x^*, \nabla f(x^k)\rangle \leq -\frac{\mu}{2}\|x^k - x^*\|^2$. Thus, we obtain:

$$\mathbb{E}_i\|x^{k+1} - x^*\|^2 \leq \left(1 - \mu\min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\right)\|x^k - x^*\|^2 + 2\gamma_b\sigma^2$$

Taking expectations again and using the tower property:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \left(1 - \mu\min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\right)\mathbb{E}\|x^k - x^*\|^2 + 2\gamma_b\sigma^2 \tag{18}$$

Recursively applying the above and summing up the resulting geometric series gives:

$$\begin{aligned}\mathbb{E}\|x^k - x^*\|^2 &\leq \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}\right)^k \|x^0 - x^*\|^2 \\ &\quad + 2\gamma_{\mathsf{b}}\sigma^2 \sum_{j=0}^{k-1} \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}\right)^j \\ &\leq \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}\right)^k \|x^0 - x^*\|^2 + 2\gamma_{\mathsf{b}}\sigma^2 \frac{1}{\mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}}\end{aligned}$$

Let $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}$ then,

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \mu\alpha)^k \|x^0 - x^*\|^2 + \frac{2\gamma_{\mathsf{b}}\sigma^2}{\mu\alpha} \tag{19}$$

From definition of α is clear that having small parameter c improves both the convergence rate $1 - \mu\alpha$ and the neighborhood $\frac{2\gamma_{\mathsf{b}}\sigma^2}{\mu\alpha}$. Since we have the restriction $c \geq \frac{1}{2}$ the best selection would be $c = \frac{1}{2}$. □

B.1.1 Comparison to the setting from Berrada et al. (2020)

In the next corollary, in order to compare against the results for ALI-G from Berrada et al. (2020), we make the strong assumption that all functions f_i have the same properties. We note that such an assumption in the interpolation setting is quite strong and reduces the finite-sum optimization to minimization of a single function in the finite sum.

Corollary B.1. Let all the assumptions in Theorem 3.1 be satisfied and let all f_i be μ-strongly convex[a] and L-smooth. SGD with $\text{SPS}_{\max}$ with $c = 1/2$ converges as:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \|x^0 - x^*\|^2 + \frac{2\sigma^2 L}{\mu^2}.$$

For the interpolated case ($\sigma = 0$) we obtain the same convergence as Corollary 3.2 with $L_{\max} = L$.

[a] This is a much stronger assumption than assuming that functions f_i are convex and f is strongly convex, which is the main assumption of Theorem 3.1. Nevertheless, assuming that all f_i are μ-strongly convex functions implies that the objective function f is μ-strongly convex and that functions f are convex. Thus, Theorem 3.1 still holds.

Note that, the result of Corollary B.1 is obtained by substituting $\gamma_{\mathsf{b}} \overset{(5)}{=} \frac{1}{2c\mu} \overset{c=\frac{1}{2}}{=} \frac{1}{\mu}$ into (6).

For the setting of Corollary B.1, Berrada et al. (2020) show the linear convergence to a much larger neighborhood than ours and with slower rate. In particular, their rate is $1 - \frac{\mu}{8L}$ and the neighborhood is $\frac{8L}{\mu}(\frac{\epsilon}{L} + \frac{\delta}{4L^2} + \frac{\epsilon}{2\mu})$ where $\delta > 2L\epsilon$ and ϵ is the ϵ-interpolation parameter $\epsilon > \max_i[f_i(x^*) - f_i^*]$ which by definition is bigger than σ^2. Under interpolation where $\sigma = 0$, our method converges linearly to the x^* while the algorithm proposed by Berrada et al. (2020) still converges to a neighborhood that is proportional to the parameter δ.

B.2 Proof of Theorem 3.4

Proof.

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 &= \|x^k-\gamma_k\nabla f_i(x^k)-x^*\|^2 \\
&= \|x^k-x^*\|^2-2\gamma_k\langle x^k-x^*,\nabla f_i(x^k)\rangle+\gamma_k^2\|\nabla f_i(x^k)\|^2 \\
&\overset{convexity}{\leq} \|x^k-x^*\|^2-2\gamma_k\left[f_i(x^k)-f_i(x^*)\right]+\gamma_k^2\|\nabla f_i(x^k)\|^2 \\
&\overset{(17)}{\leq} \|x^k-x^*\|^2-2\gamma_k\left[f_i(x^k)-f_i(x^*)\right]+\frac{\gamma_k}{c}\left[f_i(x^k)-f_i^*\right] \\
&= \|x^k-x^*\|^2-2\gamma_k\left[f_i(x^k)-f_i^*+f_i^*-f_i(x^*)\right]+\frac{\gamma_k}{c}\left[f_i(x^k)-f_i^*\right] \\
&= \|x^k-x^*\|^2-\gamma_k\left(2-\frac{1}{c}\right)\underbrace{\left[f_i(x^k)-f_i^*\right]}_{>0}+2\gamma_k\underbrace{\left[f_i(x^*)-f_i^*\right]}_{>0} \qquad (20)
\end{aligned}$$

Let $\alpha=\min\left\{\frac{1}{2cL_{\max}},\gamma_b\right\}$ and recall that from the definition of $\text{SPS}_{\max}$ (3) we obtain:

$$\alpha \overset{(5),(3)}{\leq} \gamma_k \leq \gamma_b \qquad (21)$$

From the above if $\alpha=\frac{1}{2cL_{\max}}$ then the step-size is in the regime of the stochastic Polyak step (5). In the case that $\alpha=\gamma_b$ then the analyzed method becomes the constant step-size SGD with stepsize $\gamma_k=\gamma_b$.

Since $c>\frac{1}{2}$ it holds that $\left(2-\frac{1}{c}\right)>0$. Using (21) into (20) we obtain:

$$\begin{aligned}
\|x^{k+1}-x^*\|^2 &\leq \|x^k-x^*\|^2-\gamma_k\left(2-\frac{1}{c}\right)\left[f_i(x^k)-f_i^*\right]+2\gamma_k\left[f_i(x^*)-f_i^*\right] \\
&\overset{(21)}{\leq} \|x^k-x^*\|^2-\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^k)-f_i^*\right]+2\gamma_b\left[f_i(x^*)-f_i^*\right] \\
&= \|x^k-x^*\|^2-\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^k)-f_i(x^*)+f_i(x^*)-f_i^*\right] \\
&\quad +2\gamma_b\left[f_i(x^*)-f_i^*\right] \\
&= \|x^k-x^*\|^2-\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^k)-f_i(x^*)\right]-\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^*)-f_i^*\right] \\
&\quad +2\gamma_b\left[f_i(x^*)-f_i^*\right] \\
&\leq \|x^k-x^*\|^2-\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^k)-f_i(x^*)\right]+2\gamma_b\left[f_i(x^*)-f_i^*\right] \qquad (22)
\end{aligned}$$

where in the last inequality we use that $\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^*)-f_i^*\right]>0$.

By rearranging:

$$\alpha\left(2-\frac{1}{c}\right)\left[f_i(x^k)-f_i(x^*)\right] \leq \|x^k-x^*\|^2-\|x^{k+1}-x^*\|^2+2\gamma_b\left[f_i(x^*)-f_i^*\right] \qquad (23)$$

By taking expectation condition on x^k and dividing by $\alpha\left(2-\frac{1}{c}\right)$:

$$f(x^k)-f(x^*) \leq \frac{c}{\alpha(2c-1)}\left(\|x^k-x^*\|^2-\mathbb{E}_i\|x^{k+1}-x^*\|^2\right)+2\gamma_b\frac{c}{\alpha(2c-1)}\sigma^2$$

Taking expectation again and using the tower property:

$$\mathbb{E}[f(x^k)-f(x^*)] \leq \frac{c}{\alpha(2c-1)}\left(\mathbb{E}\|x^k-x^*\|^2-\mathbb{E}\|x^{k+1}-x^*\|^2\right)+2\gamma_b\frac{c}{\alpha(2c-1)}\sigma^2$$

Summing from $k = 0$ to $K - 1$ and dividing by K:

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k) - f(x^*)\right] &= \frac{c}{\alpha(2c-1)}\frac{1}{K}\sum_{k=0}^{K-1}\left(\mathbb{E}\|x^k - x^*\|^2 - \mathbb{E}\|x^{k+1} - x^*\|^2\right) \\
&\quad + \frac{1}{K}\sum_{k=0}^{K-1}\frac{2c\gamma_b\sigma^2}{\alpha(2c-1)} \\
&= \frac{2c}{\alpha(2c-1)}\frac{1}{K}\left[\|x^0 - x^*\|^2 - \mathbb{E}\|x^K - x^*\|^2\right] + \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)} \\
&\leq \frac{c}{\alpha(2c-1)}\frac{1}{K}\|x^0 - x^*\|^2 + \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)} \qquad (24)
\end{aligned}
$$

Let $\bar{x}^K = \frac{1}{K}\sum_{k=0}^{K-1} x^k$, then:

$$
\mathbb{E}\left[f(\bar{x}^K) - f(x^*)\right] \overset{Jensen}{\leq} \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq \frac{c}{\alpha(2c-1)}\frac{1}{K}\|x^0 - x^*\|^2 + \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)}
$$

For $c = 1$:

$$
\mathbb{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{\|x^0 - x^*\|^2}{\alpha K} + \frac{2\gamma_b\sigma^2}{\alpha} \qquad (25)
$$

and this completes the proof. $\square$

At this point we highlight that $c = 1$ is selected to simplify the expression of the upper bound in (25). This is not the optimum choice (the one that makes the rate and the neighborhood of the upper bound smaller). In order to compute the optimum value of c one needs to follow similar procedure to Gower et al. (2019) and Needell et al. (2016). In this case c will depend on parameter σ and the desired accuracy ϵ of convergence.

However as we show bellow having $c = 1$ allows SGD with SPS to convergence faster than the ALI-G algorithm (Berrada et al., 2020) and the SLS algorithm (Vaswani et al., 2019b) for the case of smooth convex functions.

Comparison with other methods Similar to the strongly convex case let us compare the above convergence for smooth convex functions with the convergence rates proposed in Vaswani et al. (2019b) and Berrada et al. (2020).

For the smooth convex functions, Berrada et al. (2020) show the linear convergence to a much larger neighborhood than ours and with slower rate. In particular, their rate is $\frac{1}{K}\left(\frac{2L}{1-\frac{2L\epsilon}{\delta}}\right)$ and the neighborhood is $\frac{\delta}{L(1-\frac{2L\epsilon}{\delta})}$ where $\delta > 2L\epsilon$ and ϵ is the ϵ-interpolation parameter $\epsilon > \max_i[f_i(x^*) - f_i^*]$ which by definition is bigger than σ^2. Under interpolation where $\sigma = 0$, our method converges with a $O(1/K)$ rate to the x^* while the algorithm proposed by Berrada et al. (2020) still converges to a neighborhood that is proportional to the parameter δ.

In the interpolation setting our rate is similar to the one obtain for the stochastic line search (SLS) proposed in Vaswani et al. (2019b). In particular in the interpolation setting, SLS achieves the following $O(1/K)$ rate $\mathbb{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{\max\{3L_{\max}, 2/\gamma_b\}}{K}\|x^0 - x^*\|^2$ which has slightly worse constants than SGD with SPS.

B.3 SPS on Methods for Solving Consistent Linear Systems

Recently several new randomized iterative methods (sketch and project methods) for solving large-scale linear systems have been proposed (Richtárik and Takác, 2020; Loizou and Richtárik, 2020b,a; Gower and Richtárik, 2015). The main algorithm in this literature is the celebrated randomized Kaczmarz (RK) method (Kaczmarz, 1937; Strohmer and Vershynin, 2009) which can be seen as special case of SGD for solving least square problems (Needell et al., 2016). In this area of research, it is well known that the theoretical best constant step-size for RK method is $\gamma = 1$.

As we have already mentioned in Section 3.4, given the consistent linear system

$$
\mathbf{A}x = b, \qquad (26)
$$

Richtárik and Takác (2020) provide a stochastic optimization reformulation of the form (1) which is equivalent to the linear system in the sense that their solution sets are identical. That is, the set of minimizers of the stochastic optimization problem $\mathcal{X}^*$ is equal to the set of solutions of the stochastic linear system $\mathcal{L} := \{x : \mathbf{A}x = b\}$.

In particular, the stochastic convex quadratic optimization problem proposed in Richtárik and Takác (2020), can be expressed as follows:

$$\min_{x\in\mathbb{R}^n} f(x) := \mathbb{E}_{\mathbf{S}\sim\mathcal{D}} f_{\mathbf{S}}(x). \tag{27}$$

Here the expectation is over random matrices $\mathbf{S}$ drawn from an arbitrary, user defined, distribution $\mathcal{D}$ and $f_{\mathbf{S}}$ is a stochastic convex quadratic function of a least-squares type, defined as

$$f_{\mathbf{S}}(x) := \frac{1}{2}\|\mathbf{A}x - b\|^2_{\mathbf{H}} = \frac{1}{2}(\mathbf{A}x - b)^\top \mathbf{H}(\mathbf{A}x - b). \tag{28}$$

Function $f_{\mathbf{S}}$ depends on the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^m$ of the linear system (26) and on a random symmetric positive semidefinite matrix $\mathbf{H} := \mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S})^\dagger\mathbf{S}^\top$. By $\dagger$ we denote the Moore-Penrose pseudoinverse.

For solving problem (27), Richtárik and Takác (2020) analyze SGD with constant step-size:

$$x^{k+1} = x^k - \gamma\nabla f_{\mathbf{S}_k}(x^k), \tag{29}$$

where $\nabla f_{\mathbf{S}_k}(x^k)$ denotes the gradient of function $f_{\mathbf{S}_k}$. In each step the matrix $\mathbf{S}_k$ is drawn from the given distribution $\mathcal{D}$.

The above update of SGD is quite general and as explained by Richtárik and Takác (2020) the flexibility of selecting distribution $\mathcal{D}$ allow us to obtain different stochastic reformulations of the linear system (26) and different special cases of the SGD update. For example the celebrated randomized Kaczmarz (RK) method can be seen as special cases of the above update as follows:

Randomized Kaczmarz Method: Let pick in each iteration the random matrix $\mathbf{S} = e_i$ (random coordinate vector) with probability $p_i = \|\mathbf{A}_{i:}\|^2/\|\mathbf{A}\|^2_F$. In this setup the update rule of SGD (29) simplifies to

$$x^{k+1} = x^k - \omega\frac{\mathbf{A}_{i:}x^k - b_i}{\|\mathbf{A}_{i:}\|^2}\mathbf{A}_{i:}^\top$$

Many other methods like Gaussian Kacmarz, Randomized Coordinate Descent, Gaussian Decsent and their block variants can be cast as special cases of the above framework. For more details on the general framework and connections with other research areas we also suggest (Loizou and Richtárik, 2019; Loizou, 2019).

Lemma B.2 (Properties of stochastic reformulation (Richtárik and Takác, 2020))**.** For all $x \in \mathbb{R}^n$ and any $\mathbf{S} \sim \mathcal{D}$ it holds that:

$$f_{\mathbf{S}}(x) - f_{\mathbf{S}}(x^*) \overset{f^*_{\mathbf{S}}=0}{=} f_{\mathbf{S}}(x) = \frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}} = \frac{1}{2}\langle\nabla f_{\mathbf{S}}(x), x - x^*\rangle_{\mathbf{B}}. \tag{30}$$

Let x^* is the projection of vector x onto the solution set $\mathcal{X}^*$ of the optimization problem $\min_{x\in\mathbb{R}^n} f(x)$ (Recall that by the construction of the stochastic optimization problems we have that $\mathcal{X}^* = \mathcal{L}$). Then:

$$\frac{\lambda^+_{\min}(\mathbf{W})}{2}\|x - x^*\|^2_{\mathbf{B}} \leq f(x). \tag{31}$$

where $\lambda^+_{\min}$ denotes the minimum non-zero eigenvalue of matrix $\mathbf{W} = \mathbb{E}[\mathbf{A}^\top\mathbf{H}\mathbf{A}]$.

As we will see in the next Theorem, using the special structure of the stochastic reformulation (27), SPS (2) with $c = 1/2$ takes the following form:

$$\gamma_k \overset{(2)}{=} \frac{2\left[f_{\mathbf{S}}(x^k) - f^*_{\mathbf{S}}\right]}{\|\nabla f_{\mathbf{S}}(x^k)\|^2} \overset{(30)}{=} 1,$$

which is the theoretically optimal constant step-size for SGD in this setting (Richtárik and Takác, 2020). This reduction implies that SPS results in an optimal convergence rate when solving consistent linear systems. We provide the convergence rate for SPS in the next Theorem.

Though a straight forward verification of the optimality of SPS, we believe that this is the first time that SGD with adaptive step-size is reduced to constant step-size when is used for solving linear systems. SPS does that by obtaining the best convergence rate in this setting.

Theorem B.3. Let $\mathbf{A}x = b$ be a consistent linear system and let x^* is the projection of vector x onto the solution set $\mathcal{X}^* = \mathcal{L}$. Then the SGD with SPS (2) with $c = 1/2$ for solving the stochastic optimization reformulation (27) satisfies:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \left(1 - \lambda_{\min}^+(\mathbf{W})\right)^k \|x^0 - x^*\|^2 \tag{32}$$

where $\lambda_{\min}^+$ denotes the minimum non-zero eigenvalue of matrix $\mathbf{W} = \mathbb{E}[\mathbf{A}^\top \mathbf{H} \mathbf{A}]$.

Proof.

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\overset{(29)}{=} \|x^k - \gamma_k \nabla f_{\mathbf{S}_k}(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle + \gamma_k^2 \|\nabla f_{\mathbf{S}_k}(x^k)\|^2 \end{aligned} \tag{33}$$

Let us select γ_k such that the RHS of inequality (33) is minimized. That is, let us select:

$$\gamma_k = \frac{\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2} \overset{(30)}{=} \frac{2\left[f_{\mathbf{S}}(x) - f_{\mathbf{S}}(x^*)\right]}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2}$$

Substitute this step-size to (33) we obtain:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - 2\frac{\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2}\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle \\ &\quad + \left[\frac{\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2}\right]^2 \|\nabla f_{\mathbf{S}_k}(x^k)\|^2 \\ &= \|x^k - x^*\|^2 - 2\frac{[\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle]^2}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2} + \frac{\left[\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle\right]^2}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2} \\ &= \|x^k - x^*\|^2 - \frac{[\langle x^k - x^*, \nabla f_{\mathbf{S}_k}(x^k)\rangle]^2}{\|\nabla f_{\mathbf{S}_k}(x^k)\|^2} \\ &\overset{(30)}{=} \|x^k - x^*\|^2 - 2f_{\mathbf{S}}(x^k) \end{aligned} \tag{34}$$

By taking expectation with respect to $\mathbf{S}_k$ and using quadratic growth inequality (31):

$$\begin{aligned}\mathbb{E}_{\mathbf{S}_k}[\|x^{k+1} - x^*\|^2] &= \|x^k - x^*\|^2 - 2f(x^k) \\ &\overset{(31)}{\leq} \|x^k - x^*\|^2 - \lambda_{\min}^+(\mathbf{W})\|x^k - x^*\|^2 \\ &= \left[1 - \lambda_{\min}^+(\mathbf{W})\right]\|x^k - x^*\|^2. \end{aligned} \tag{35}$$

Taking expectation again and by unrolling the recurrence we obtain (32). □

We highlight that the above proof provides a different viewpoint on the analysis of the optimal constant step-size for the sketch and project methods for solving consistent liner systems. The expression of Theorem B.3 is the same with the one proposed in Richtárik and Takác (2020).

B.4 Proof of Theorem 3.6

Proof. By the smoothness of function f we have that

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2.$$

Combining this with the update rule of SGD we obtain:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \gamma_k \left\langle \nabla f(x^k), \nabla f_i(x^k) \right\rangle + \frac{L\gamma_k^2}{2}\|\nabla f_i(x^k)\|^2 \end{aligned} \tag{36}$$

By rearranging:

$$\begin{aligned} \frac{f(x^{k+1}) - f(x^k)}{\gamma_k} &\leq -\left\langle \nabla f(x^k), \nabla f_i(x^k) \right\rangle + \frac{L\gamma_k}{2}\|\nabla f_i(x^k)\|^2 \\ &\overset{(17)}{\leq} -\left\langle \nabla f(x^k), \nabla f_i(x^k) \right\rangle + \frac{L}{2c}\left[f_i(x^k) - f_i^*\right] \\ &= -\left\langle \nabla f(x^k), \nabla f_i(x^k) \right\rangle + \frac{L}{2c}\left[f_i(x^k) - f_i(x^*)\right] + \frac{L}{2c}\left[f_i(x^*) - f_i^*\right] \end{aligned}$$

and by taking expectation condition on x^k:

$$\begin{aligned} \mathbb{E}_i\left[\frac{f(x^{k+1}) - f(x^k)}{\gamma_k}\right] &\leq -\left\langle \nabla f(x^k), \nabla f(x^k) \right\rangle + \frac{L}{2c}\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\mathbb{E}_i\left[f_i(x^*) - f_i^*\right] \\ &\overset{(4)}{\leq} -\|\nabla f(x^k)\|^2 + \frac{L}{2c}\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\sigma^2 \\ &\overset{(10)}{\leq} -2\mu\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\sigma^2 \end{aligned}$$

Let $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}$. Then,

$$\begin{aligned} \mathbb{E}_i\left[\frac{f(x^{k+1}) - f(x^*)}{\gamma_k}\right] &\leq \mathbb{E}_i\left[\frac{f(x^k) - f(x^*)}{\gamma_k}\right] - 2\mu\left[f(x^k) - f(x^*)\right] \\ &\quad + \frac{L}{2c}\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\sigma^2 \\ &\overset{(5),(3)}{\leq} \frac{1}{\alpha}\left[f(x^k) - f(x^*)\right] - 2\mu\left[f(x^k) - f(x^*)\right] \\ &\quad + \frac{L}{2c}\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\sigma^2 \\ &= \left(\frac{1}{\alpha} - 2\mu + \frac{L}{2c}\right)\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\sigma^2 \\ &\overset{L \leq L_{\max}}{=} \left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right)\left[f(x^k) - f(x^*)\right] + \frac{L}{2c}\sigma^2 \end{aligned} \tag{37}$$

Using $\gamma_k \leq \gamma_b$ and by taking expectations again:

$$\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq \underbrace{\gamma_b\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right)}_{\nu}\mathbb{E}\left[f(x^k) - f(x^*)\right] + \frac{L\sigma^2\gamma_b}{2c} \tag{38}$$

By having $\nu \in (0, 1]$ and by recursively applying the above and summing the resulting geometric series we obtain:

$$\begin{aligned} \mathbb{E}\left[f(x^k) - f(x^*)\right] &\leq \nu^k\left[f(x^0) - f(x^*)\right] + \frac{L\sigma^2\gamma_b}{2c}\sum_{j=0}^{k-1}\nu^j \\ &\leq \nu^k\left[f(x^0) - f(x^*)\right] + \frac{L\sigma^2\gamma_b}{2(1-\nu)c} \end{aligned} \tag{39}$$

In the above result we require that $0 < \nu = \gamma_b\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right) \leq 1$. In order for this to hold we need to make extra assumptions on the values of γ_b and parameter c. This is what we do next.

Let us divide the analysis into two cases based on the value of parameter α. That is:

- (i) If $\frac{1}{2cL_{\max}} \leq \gamma_{\mathsf{b}}$ then,

$$\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\} = \frac{1}{2cL_{\max}} \quad \text{and} \quad \nu = \gamma_{\mathsf{b}}\left((2c + \frac{1}{2c})L_{\max} - 2\mu\right).$$

By preliminary computations, it can be easily shown that $\nu > 0$ for every $c \geq 0$. However for $\nu \leq 1$ we need to require that $\gamma_{\mathsf{b}} \leq \frac{1}{\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right)}$ and since we are already assume that $\frac{1}{2cL_{\max}} \leq \gamma_{\mathsf{b}}$ we need to force

$$\frac{1}{2cL_{\max}} \leq \frac{1}{\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right)}$$

to avoid contradiction. This is true only if $c > \frac{L_{\max}}{4\mu}$ which is the assumption of Theorem 3.6.

- (ii) If $\gamma_{\mathsf{b}} \leq \frac{1}{2cL_{\max}}$ then,

$$\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\} = \gamma_{\mathsf{b}} \quad \text{and} \quad \nu = \gamma_{\mathsf{b}}\left(\frac{1}{\gamma_{\mathsf{b}}} - 2\mu + \frac{L_{\max}}{2c}\right) = 1 - 2\mu\gamma_{\mathsf{b}} + \frac{L_{\max}}{2c}\gamma_{\mathsf{b}}.$$

Note that if we have $c > \frac{L_{\max}}{4\mu}$ (an assumption of Theorem 3.6) it holds that $\nu \leq 1$. In addition, by preliminary computations, it can be shown that $\nu > 0$ if $\gamma_{\mathsf{b}} < \frac{2c}{4\mu c - L_{\max}}$. Finally, for $c > \frac{L_{\max}}{4\mu}$ it holds that $\frac{1}{2cL_{\max}} \leq \frac{2c}{4\mu c - L_{\max}}$, and as a result $\nu > 0$ for all $\gamma_{\mathsf{b}} \leq \frac{1}{2cL_{\max}}$.

By presenting the above cases on bound of ν we complete the proof. □

Remark B.4. The expression of Corollary 3.7 is obtained by simply use $c = \frac{L_{\max}}{2\mu}$ in the case (ii) of the above proof. In this case we have $\gamma \leq \frac{\mu}{L_{\max}^2}$ and $\nu = 1 - \mu\gamma$.

B.5 Proof of Theorem 3.8

Proof. First note that:

$$\begin{aligned}
-\gamma_k \left\langle \nabla f(x^k), \nabla f_i(x^k)\right\rangle &= \frac{\gamma_k}{2}\|\nabla f_i(x^k) - \nabla f(x^k)\|^2 - \frac{\gamma_k}{2}\|\nabla f_i(x^k)\|^2 - \frac{\gamma_k}{2}\|\nabla f(x^k)\|^2 \\
&\overset{(21)}{\leq} \frac{\gamma_{\mathsf{b}}}{2}\|\nabla f_i(x^k) - \nabla f(x^k)\|^2 - \frac{\alpha}{2}\|\nabla f_i(x^k)\|^2 - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= \frac{\gamma_{\mathsf{b}}}{2}\|\nabla f_i(x^k)\|^2 + \frac{\gamma_{\mathsf{b}}}{2}\|\nabla f(x^k)\|^2 - \gamma_{\mathsf{b}}\left\langle \nabla f(x^k), \nabla f_i(x^k)\right\rangle \\
&\quad -\frac{\alpha}{2}\|\nabla f_i(x^k)\|^2 - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= \left(\frac{\gamma_{\mathsf{b}}}{2} - \frac{\alpha}{2}\right)\|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_{\mathsf{b}}}{2} - \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \\
&\quad -\gamma_{\mathsf{b}}\left\langle \nabla f(x^k), \nabla f_i(x^k)\right\rangle
\end{aligned} \tag{40}$$

By the smoothness of function f we have that

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2.$$

Combining this with the update rule of SGD we obtain:

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k\right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&\leq f(x^k) - \gamma_k \left\langle \nabla f(x^k), \nabla f_i(x^k)\right\rangle + \frac{L\gamma_k^2}{2}\|\nabla f_i(x^k)\|^2 \\
&\overset{(3)}{\leq} f(x^k) - \gamma_k \left\langle \nabla f(x^k), \nabla f_i(x^k)\right\rangle + \frac{L\gamma_{\mathsf{b}}^2}{2}\|\nabla f_i(x^k)\|^2 \\
&\overset{(40)}{\leq} f(x^k) + \left(\frac{\gamma_{\mathsf{b}}}{2} - \frac{\alpha}{2} + \frac{L\gamma_{\mathsf{b}}^2}{2}\right)\|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_{\mathsf{b}}}{2} - \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \\
&\quad -\gamma_{\mathsf{b}}\left\langle \nabla f(x^k), \nabla f_i(x^k)\right\rangle
\end{aligned} \tag{41}$$

By taking expectation condition on x^k:

$$\begin{aligned}
\mathbb{E}_i f(x^{k+1}) &\overset{(41)}{\leq} f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2}\right)\mathbb{E}_i\|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \\
&\quad -\gamma_b \left\langle \nabla f(x^k), \mathbb{E}_i \nabla f_i(x^k)\right\rangle \\
&= f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2}\right)\mathbb{E}_i\|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \\
&\quad -\gamma_b\|\nabla f(x^k)\|^2 \\
&= f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2}\right)\mathbb{E}_i\|\nabla f_i(x^k)\|^2 - \left(\frac{\gamma_b}{2} + \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \qquad (42)
\end{aligned}$$

Since $0 < \alpha \leq \gamma_b$ we have that $\left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2}\right) > 0$. Thus, we are able to use (8):

$$\begin{aligned}
\mathbb{E}_i f(x^{k+1}) &\leq f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2}\right)\mathbb{E}_i\|\nabla f_i(x^k)\|^2 - \left(\frac{\gamma_b}{2} + \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \\
&\overset{(8)}{\leq} f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2}\right)\left[\rho\|\nabla f(x)\|^2 + \delta\right] - \left(\frac{\gamma_b}{2} + \frac{\alpha}{2}\right)\|\nabla f(x^k)\|^2 \\
&= f(x^k) + \frac{1}{2}\left[\left(\gamma_b - \alpha + L\gamma_b^2\right)\rho - (\gamma_b + \alpha)\right]\|\nabla f(x)\|^2 + \frac{1}{2}\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta
\end{aligned}$$

By rearranging and taking expectation again:

$$\underbrace{\left[(\gamma_b + \alpha) - \left(\gamma_b - \alpha + L\gamma_b^2\right)\rho\right]}_{\zeta}\mathbb{E}\left[\|\nabla f(x)\|^2\right] \leq 2\left(\mathbb{E}[f(x^k)] - \mathbb{E}[f(x^{k+1})]\right) + \left(\gamma_b - \alpha + L\gamma_b^2\right)\delta$$

If $\zeta > 0$ then:

$$\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\zeta}\left(\mathbb{E}[f(x^k)] - \mathbb{E}[f(x^{k+1})]\right) + \frac{\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta}{\zeta} \qquad (43)$$

By summing from $k = 0$ to $K - 1$ and dividing by K:

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|^2] &\leq \frac{2}{\zeta}\frac{1}{K}\sum_{k=0}^{K-1}\left(\mathbb{E}[f(x^k)] - \mathbb{E}[f(x^{k+1})]\right) + \frac{1}{K}\sum_{k=0}^{K-1}\frac{\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta}{\zeta} \\
&\leq \frac{2}{\zeta}\frac{1}{K}\left(f(x^0) - \mathbb{E}[f(x^K)]\right) + \frac{\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta}{\zeta} \\
&\leq \frac{2}{\zeta K}\left(f(x^0) - f(x^*)\right) + \frac{\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta}{\zeta} \qquad (44)
\end{aligned}$$

In the above result we require that $\zeta = (\gamma_b + \alpha) - \left(\gamma_b - \alpha + L\gamma_b^2\right)\rho > 0$. In order for this to hold we need to make extra assumptions on the values of γ_b and parameter c. This is what we do next.

Let us divide the analysis into two cases. That is:

- (i) If $\frac{1}{2cL_{\max}} \leq \gamma_b$ then $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\} = \frac{1}{2cL_{\max}}$ and

$$\zeta = (\gamma_b + \alpha) - \left(\gamma_b - \alpha + L\gamma_b^2\right)\rho = \left(\gamma_b + \frac{1}{2cL_{\max}}\right) - \left(\gamma_b - \frac{1}{2cL_{\max}} + L\gamma_b^2\right)\rho.$$

By solving the quadratic expression of ζ with respect to γ_b, it can be easily shown that $\zeta > 0$ if

$$0 < \gamma_b < \bar{\gamma}_b := \frac{-(\rho - 1) + \sqrt{(\rho-1)^2 + \frac{4L\rho(\rho+1)}{2cL_{\max}}}}{2L\rho}.$$

To avoid contradiction the inequality $\frac{1}{2cL_{\max}} < \bar{\gamma}_b$ needs to be true, where $\bar{\gamma}_b$ is the above upper bound of γ_b. This is the case of $c > \frac{L\rho}{4L_{\max}}$ which is the assumption of Theorem 3.8.

- (ii) If $\gamma_b \le \frac{1}{2cL_{\max}}$ then $\alpha = \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\} = \gamma_b$ and

$$\zeta = (\gamma_b + \alpha) - \left(\gamma_b - \alpha + L\gamma_b^2\right)\rho = (\gamma_b + \gamma_b) - \left(\gamma_b - \gamma_b + L\gamma_b^2\right)\rho = 2\gamma_b - L\gamma_b^2\rho$$

In this case, by preliminary computations, it can be shown that $\zeta > 0$ if $\gamma_b < \frac{2}{L\rho}$. For $c > \frac{L\rho}{4L_{\max}}$ it also holds that $\frac{1}{2cL_{\max}} < \frac{2}{L\rho}$.

□

B.5.1 Additional convergence result for nonconvex smooth functions: Assuming independence of step-size and stochastic gradient

Let us now present an extra theoretical result in which we assume that the step-size γ_k and the stochastic gradient $\nabla f_i(x^k)$ in each step are not correlated. Such assumption has been recently used to prove convergence of SGD with the AdaGrad step-size (Ward et al., 2019) and for the analysis of stochastic line search in Vaswani et al. (2019b). From a technical viewpoint, we highlight that for the proofs in the non-convex setting, we use the lower and upper bound of SPS rather than its exact form. This is what allows us to use this independence.

Let us state the main Theorem with the extra condition of independence and present its proof.

Theorem B.5. Let f and f_i be smooth functions and assume that there exist $\rho, \delta > 0$ such that the condition (8) is satisfied. Assuming independence of the step-size γ_k and the stochastic gradient $\nabla f_{i_k}(x^k)$ at every iteration k, SGD with SPS$_{\max}$ with $c > \frac{\rho L}{4L_{\max}}$ and $\gamma_b < \max\left\{\frac{2}{L\rho}, \frac{1}{\sqrt{\rho c\, L L_{\max}}}\right\}$ converges as:

$$\min_{k\in[K]} \mathbb{E}\|\nabla f(x^k)\|^2 \le \frac{f(x^0) - f(x^*)}{\alpha\, K} + \frac{L\delta\gamma_b^2}{2\alpha}$$

where $\beta_1 = 1 - \frac{\rho c\, L L_{max}\, \gamma_b^2}{2}$ and $\beta_2 = 1 - \frac{\rho L\, \gamma_b}{2}$, $\alpha = \min\left\{\frac{\beta_1}{2cL_{\max}}, \gamma_b\, \beta_2\right\}$.

Proof. By the smoothness of function f we have that

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2.$$

Combining this with the update rule of SGD we obtain:

$$\begin{aligned} f(x^{k+1}) &\le f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k\right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &\le f(x^k) - \gamma_k\left\langle \nabla f(x^k), \nabla f_{i_k}(x^k)\right\rangle + \frac{L\gamma_k^2}{2}\|\nabla f_i(x^k)\|^2 \end{aligned} \tag{45}$$

Taking expectations with respect to i_k and noting that γ_k is independent of $\nabla f_i(x^k)$ yields:

$$\begin{aligned} \mathbb{E}_{i_k} f(x^{k+1}) &\le f(x^k) - \gamma_k\left\langle \nabla f(x^k), \mathbb{E}_i \nabla f_{i_k}(x^k)\right\rangle + \frac{L\gamma_k^2}{2}\mathbb{E}_{i_k}\|\nabla f_i(x^k)\|^2 \\ &= f(x^k) - \gamma_k\|\nabla f(x^k)\|^2 + \frac{L\gamma_k^2}{2}\mathbb{E}_{i_k}\|\nabla f_i(x^k)\|^2 \\ &\overset{(8)}{\le} f(x^k) - \gamma_k\|\nabla f(x^k)\|^2 + \frac{L\gamma_k^2}{2}\left[\rho\|\nabla f(x)\|^2 + \delta\right] \\ &\le f(x^k) - \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\|\nabla f(x^k)\|^2 + \frac{L\gamma_b^2}{2}\left[\rho\|\nabla f(x)\|^2 + \delta\right] \\ &= f(x^k) - \left(\min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\} - \frac{L\gamma_b^2\rho}{2}\right)\|\nabla f(x^k)\|^2 + \frac{L\gamma_b^2}{2}\delta \end{aligned} \tag{46}$$

By rearranging and taking expectations again:

$$\underbrace{\left(\min\left\{\frac{1}{2cL_{\max}},\gamma_{\mathsf{b}}\right\}-\frac{L\gamma_{\mathsf{b}}^2}{2}\rho\right)}_{\alpha}\mathbb{E}[\|\nabla f(x^k)\|^2] \quad\leq\quad \mathbb{E}[f(x^k)]-\mathbb{E}[f(x^{k+1})]+\frac{L\gamma_{\mathsf{b}}^2}{2}\delta \tag{47}$$

Let $\alpha>0$ then:

$$\mathbb{E}[\|\nabla f(x^k)\|^2] \quad\leq\quad \frac{1}{\alpha}\left(\mathbb{E}[f(x^k)]-\mathbb{E}[f(x^{k+1})]\right)+\frac{L\gamma_{\mathsf{b}}^2\delta}{2\alpha} \tag{48}$$

By summing from $k=0$ to $K-1$ and dividing by K:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|^2] \quad&\leq\quad \frac{1}{\alpha}\frac{1}{K}\sum_{k=0}^{K-1}\left(\mathbb{E}[f(x^k)]-\mathbb{E}[f(x^{k+1})]\right)+\frac{1}{K}\sum_{k=0}^{K-1}\frac{L\gamma_{\mathsf{b}}^2\delta}{2\alpha}\\ &\leq\quad \frac{1}{\alpha}\frac{1}{K}\left(f(x^0)-\mathbb{E}[f(x^K)]\right)+\frac{L\gamma_{\mathsf{b}}^2\delta}{2\alpha}\\ &\leq\quad \frac{1}{\alpha K}\left(f(x^0)-f(x^*)\right)+\frac{L\gamma_{\mathsf{b}}^2\delta}{2\alpha}\end{aligned} \tag{49}$$

In the above result we require that $\alpha=\left(\min\left\{\frac{1}{2cL_{\max}},\gamma_{\mathsf{b}}\right\}-\frac{L\gamma_{\mathsf{b}}^2}{2}\rho\right)>0$. In order for this to hold we need to make extra assumptions on the values of γ_{b} and parameter c. This is what we do next.

Let us divide the analysis into two cases. That is:

- (i) If $\frac{1}{2cL_{\max}}\leq\gamma_{\mathsf{b}}$ then,

$$\alpha=\left(\min\left\{\frac{1}{2cL_{\max}},\gamma_{\mathsf{b}}\right\}-\frac{L\gamma_{\mathsf{b}}^2}{2}\rho\right)=\left(\frac{1}{2cL_{\max}}-\frac{L\gamma_{\mathsf{b}}^2}{2}\rho\right).$$

  By preliminary computations, it can be easily shown that $\alpha>0$ if $\gamma_{\mathsf{b}}<\frac{1}{\sqrt{L\rho cL_{\max}}}$. To avoid contraction the inequality $\frac{1}{2cL_{\max}}<\frac{1}{\sqrt{L\rho cL_{\max}}}$ needs to be true. This is the case of $c>\frac{L\rho}{4L_{\max}}$ which is the assumptions of Theorem B.5.

- (ii) If $\gamma_{\mathsf{b}}\leq\frac{1}{2cL_{\max}}$ then,

$$\alpha=\left(\min\left\{\frac{1}{2cL_{\max}},\gamma_{\mathsf{b}}\right\}-\frac{L\gamma_{\mathsf{b}}^2}{2}\rho\right)=\gamma_{\mathsf{b}}-\frac{L\gamma_{\mathsf{b}}^2}{2}\rho=\gamma_{\mathsf{b}}\left(1-\frac{L\gamma_{\mathsf{b}}}{2}\rho\right).$$

  In this case, by preliminary computations, it can be shown that $\alpha>0$ if $\gamma_{\mathsf{b}}<\frac{2}{L\rho}$. For $c>\frac{L\rho}{4L_{\max}}$ it also holds that $\frac{1}{2cL_{\max}}<\frac{2}{L\rho}$.

$\square$

C Additional Convergence Results

In this section we present some additional convergence results. We first prove a $O(1/\sqrt{K})$ convergence rate of stochastic subgradient method with SPS for non-smooth convex functions in the interpolated setting. Furthermore, similar to Schmidt et al. (2011), we propose a way to increase the mini-batch size for evaluating the stochastic gradient and guarantee convergence to the optimal solution without interpolation.

C.1 Non-smooth Convex Functions

In all of our previous results we assume that functions f_i are smooth. As a result, in the proofs of our theorems we were able to use the lower bound (5) of SPS. In the case that functions f_i are not smooth using this lower is clearly not possible. Below we present a Theorem that handles the case of non-smooth function for the convergence of stochastic subgradient method[5]. For this result we require that a constant G exists such that $\|g_i(x)\|^2 < G^2$ for each subgradient of function f_i. This is equivalent with assuming that functions f_i are G-Lipschitz. To keep the presentation simple we only present the interpolated case. Using the proof techniques from the rest of the paper one can easily obtain convergence for the more general setting.

Theorem C.1. Assume interpolation and that f and f_i are convex non-smooth functions. Let G be a constant such that $\|g_i(x)\|^2 < G^2, \forall i \in [n] \quad \text{and} \quad x \in \mathbb{R}^n$. Let γ^k be the subgradient counterpart of SPS (2) with $c = 1$. Then the iterates of the stochastic subgradient method satisfy:

$$\mathbb{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{G\|x^0 - x^*\|}{\sqrt{K}} = O\left(\frac{1}{\sqrt{K}}\right)$$

where $\bar{x}^K = \frac{1}{K}\sum_{k=0}^{K-1} x^k$.

Proof. The proof is similar to the deterministic case (see Theorem A.4). That is, we select the γ_k that minimize the right hand side of the inequality after the use of convexity.

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - \gamma_k g_i^k - x^*\|^2 \\
&= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, g_i^k\rangle + \gamma_k^2 \|g_i^k\|^2 \\
&\overset{convexity}{\leq} \|x^k - x^*\|^2 - 2\gamma_k \left[f_i(x^k) - f_i(x^*)\right] + \gamma_k^2 \|g_i^k\|^2
\end{aligned} \tag{50}$$

Using the subgradient counterpart of SPS (2) with $c = 1$, that is, $\gamma_k = \frac{f_i(x^k) - f_i(x^*)}{\|g_i^k\|^2}$[6] we obtain:

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - 2\frac{f_i(x^k) - f_i(x^*)}{\|g_i^k\|^2}\left[f_i(x^k) - f_i(x^*)\right] \\
&\quad + \left[\frac{f_i(x^k) - f_i(x^*)}{\|g_i^k\|^2}\right]^2 \|g^k\|^2 \\
&= \|x^k - x^*\|^2 - \frac{\left[f_i(x^k) - f_i(x^*)\right]^2}{\|g_i^k\|^2} \\
&\overset{\|g_i(x)\|^2 < G^2}{\leq} \|x^k - x^*\|^2 - \frac{\left[f_i(x^k) - f_i(x^*)\right]^2}{G^2}
\end{aligned} \tag{51}$$

taking expectation condition on x^k:

$$\begin{aligned}
\mathbb{E}_i\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \frac{\mathbb{E}_i\left[f_i(x^k) - f_i(x^*)\right]^2}{G^2} \\
&\overset{Jensen}{\leq} \|x^k - x^*\|^2 - \frac{\left[\mathbb{E}[f_i(x^k) - f_i(x^*)]\right]^2}{G^2} \\
&= \|x^k - x^*\|^2 - \frac{\left[f(x^k) - f(x^*)\right]^2}{G^2}
\end{aligned} \tag{52}$$

Taking expectation again and using the tower property:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \mathbb{E}\|x^k - x^*\|^2 - \frac{\mathbb{E}\left[f(x^k) - f(x^*)\right]^2}{G^2} \tag{53}$$

[5]Note that for non-smooth functions, it is required to have stochastic subgradient method instead of SGD. That is, in each iteration we replace the evaluation of $\nabla f_i(x)$ with its subgradient counterpart $g_i(x)$

[6]Recall that in the interpolation setting it holds that $f_i^* = f^* = f(x^*)$.

By rearranging, summing from $k=0$ to $K-1$ and dividing by K:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\frac{\mathbb{E}\left[f(x^k)-f(x^*)\right]^2}{G^2} &\leq \frac{1}{K}\sum_{k=0}^{K-1}\left[\mathbb{E}\|x^k-x^*\|^2-\mathbb{E}\|x^{k+1}-x^*\|^2\right]\\ &= \frac{1}{K}\left[\|x^0-x^*\|^2-\mathbb{E}\|x^K-x^*\|^2\right]\\ &= \frac{1}{K}\left[\|x^0-x^*\|^2\right] \end{aligned}\tag{54}$$

Taking square roots and using Jensen's inequality:

$$\frac{1}{GK}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f(x^*)\right] \overset{Jensen}{\leq} \frac{1}{G}\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f(x^*)\right]^2} \leq \frac{1}{\sqrt{K}}\left[\|x^0-x^*\|\right]\tag{55}$$

Thus,

$$\mathbb{E}\left[f(\bar{x}^K)-f(x^*)\right] \overset{Jensen}{\leq} \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f(x^*)\right] \leq \frac{G\|x^0-x^*\|}{\sqrt{K}},\tag{56}$$

where $\bar{x}^K=\frac{1}{K}\sum_{k=0}^{K-1}x^k$. □

C.2 Increasing Mini-batch Size

We propose a way to increase the mini-batch size for evaluating the stochastic gradient and guarantee convergence to the optimal solution without interpolation. We present two main Theorems. In the first Theorem we assume that functions f_i of problem (1) are μ_i-strongly convex functions and in the second that each function f_i satisfies the PL condition (10) with μ_i parameter.

Theorem C.2. Let us have the same assumptions as in Theorem 3.1 and let all f_i be μ_i-strongly convex functions then SGD with SPS and increasing the batch-size progressively such that the batch-size b_k at iteration k satisfies:

$$b_k \geq \left[\frac{1}{n}+\frac{1}{4\gamma_{max}\,z^2}\frac{\mu_{\min}\mu}{L_{\max}}\left(\frac{\|\nabla f(x^k)\|}{L}\right)^2\right]^{-1}$$

where $z^2=\sup_x\mathbb{E}\left[|]|\nabla f_i(x)-\nabla f(x)||^2\right.$. Then SGD with SPS converges as:

$$\mathbb{E}\|x^k-x^*\|^2\leq\left(1-\frac{\mu}{4c\,L_{\max}}\right)^k\|x^0-x^*\|^2.$$

Proof. Following the proof of Theorem 3.1 for the batch b. From Equation 18,

$$\mathbb{E}_i\|x^{k+1}-x^*\|^2 \leq \left(1-\mu\min\left\{\frac{1}{2cL_{\max}},\gamma_{\mathrm{b}}\right\}\right)\|x^k-x^*\|^2+2\gamma_{\mathrm{b}}\sigma^2$$

By strong-convexity of all f_i, $\forall i\in[n]$ the minibatch function f_b is μ_b-strongly convex and it holds that:

$$\begin{aligned}\sigma^2=\mathbb{E}\left[f_b(x^*)-f_b^*\right] &\leq \mathbb{E}\left[\frac{1}{2\mu_b}\|\nabla f_b(x^*)\|^2\right]\\ &\leq \frac{1}{2\mu_{\min}}\mathbb{E}\|\nabla f_b(x^*)\|^2,\end{aligned}\tag{57}$$

where in the last inequality we use that $\mu_{\min}=\min\{\mu_i\}_{1=1}^n\leq\mu_i\leq\mu_b$. By the assumption that the gradients at the optimum have bounded variance, from Harikandeh et al. (2015); Lohr (2019):

$$\mathbb{E}\|\nabla f_b(x^*)\|^2\leq\frac{n-b}{n\,b}z^2,\tag{58}$$

Thus,

$$\sigma^2 = \mathbb{E}\left[f_b(x^*) - f_b^*\right] \overset{(57),(58)}{\leq} \frac{1}{2\mu_{\min}} \frac{n-b}{n\,b} z^2$$

and as a result,

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \quad \leq \quad \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}\right) \mathbb{E}\|x^k - x^*\|^2 + \frac{\gamma_{\mathsf{b}}}{\mu_{\min}} \frac{n-b}{n\,b} z^2$$

If we set the batch-size in iteration k such that,

$$\begin{aligned} \frac{\gamma_{\mathsf{b}}}{\mu_{min}} \frac{n-b}{n\,b} z^2 \quad &\leq \quad \frac{\mu}{4cL_{\max}} \left(\frac{\|\nabla f(x^k)\|}{L}\right)^2 \\ \implies b \quad &\geq \quad \left[\frac{1}{n} + \frac{1}{\gamma_{max}\, z^2} \frac{\mu_{\min}\mu}{4cL_{\max}} \left(\frac{\|\nabla f(x^k)\|}{L}\right)^2\right]^{-1} \end{aligned} \tag{59}$$

$$\begin{aligned} \mathbb{E}_i\|x^{k+1} - x^*\|^2 \quad &\leq \quad \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}\right) \|x^k - x^*\|^2 + \frac{\mu}{4cL_{\max}} \left(\frac{\|\nabla f(x^k)\|}{L}\right)^2 \\ &\leq \quad \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_{\mathsf{b}}\right\}\right) \|x^k - x^*\|^2 + \frac{\mu}{4cL_{\max}} \|x^k - x^*\|^2 \\ \mathbb{E}_i\|x^{k+1} - x^*\|^2 \quad &\leq \quad \left(1 - \mu \min\left\{\frac{1}{4cL_{\max}}, \gamma_{\mathsf{b}} - \frac{1}{4cL_{\max}}\right\}\right) \|x^k - x^*\|^2 \end{aligned}$$

Following the remaining proof of Theorem 3.1,

$$\mathbb{E}\|x^k - x^*\|^2 \quad \leq \quad \left(1 - \mu \min\left\{\frac{1}{4cL_{\max}}, \gamma_{\mathsf{b}} - \frac{1}{4cL_{\max}}\right\}\right)^k \|x^0 - x^*\|^2$$

For SPS it holds that $\gamma_{\mathsf{b}} = \infty$. Thus,

$$\mathbb{E}\|x^k - x^*\|^2 \quad \leq \quad \left(1 - \frac{\mu}{4cL_{\max}}\right)^k \|x^0 - x^*\|^2 \tag{60}$$

□

Theorem C.3. Assume that all functions f_i satisfy the PL inequality (10) and let f and f_i be smooth functions. SGD with $\text{SPS}_{\max}$ and increasing the batch-size progressively such that the batch-size b_k at iteration k satisfies:

$$b \quad \geq \quad \left[\frac{1}{n} + \frac{2}{\gamma_{max}\, z^2} \frac{\mu_{\min} v}{cL} [f(x^k) - f(x^*)]\right]^{-1} \tag{61}$$

where $z^2 = \sup_x \mathbb{E}\left[|]|\nabla f_i(x) - \nabla f(x)||^2\right.$. Then SGD with SPS converges as:

$$\mathbb{E}[f(x^k) - f(x^*)] \leq (1 - v/2)^k \left(f(x^0 - f(x^*))\right) \tag{62}$$

where $\nu = 1 - \gamma_{\mathsf{b}}\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right) \in (0, 1)$.

Proof. Following the proof of Theorem 3.6, from Equation 38,

$$\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \quad \leq \quad \underbrace{\gamma_{\mathsf{b}}\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c}\right)}_{1-\nu} \mathbb{E}\left[f(x^k) - f(x^*)\right] + \frac{L\gamma_{\mathsf{b}}}{2c} \mathbb{E}\left[f_b(x^*) - f_b^*\right]$$

Similar to the proof of Theorem C.2, since each function f_i is PL,

$$\begin{aligned}
\mathbb{E}\left[f_b(x^*) - f_b^*\right] &\leq \mathbb{E}\left[\frac{1}{2\mu_i}\|\nabla f_b(x^*)\|^2\right] \\
&\leq \frac{1}{2\mu_{\min}}\mathbb{E}\|\nabla f_b(x^*)\|^2 \\
\mathbb{E}\left[f_b(x^*) - f_b^*\right] &\leq \frac{1}{2\mu_{\min}}\frac{n-b}{n\,b}z^2 \qquad (63)
\end{aligned}$$

From the above relations,

$$\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq (1-\nu)\,\mathbb{E}\left[f(x^k) - f(x^*)\right] + \frac{L\gamma_b}{2c}\frac{1}{2\mu_{\min}}\frac{n-b}{n\,b}z^2 \tag{64}$$

If we set the batch-size b s.t.

$$\begin{aligned}
\frac{L\gamma_b}{2c}\frac{1}{2\mu_{\min}}\frac{n-b}{n\,b}z^2 &\leq \frac{v}{2}\left[f(x^k) - f(x^*)\right] \\
\implies b &\geq \left[\frac{1}{n} + \frac{2}{\gamma_{max}\,z^2}\frac{\mu_{\min}v}{cL}[f(x^k) - f(x^*)]\right]^{-1}
\end{aligned}$$

$$\implies \mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq (1-\nu/2)\,\mathbb{E}\left[f(x^k) - f(x^*)\right] \tag{65}$$

Following the remaining proof of Theorem 3.6,

$$\mathbb{E}\left[f(x^{k+1}) - f(x^*)\right] \leq (1-\nu/2)^k\left[f(x^0) - f(x^*)\right] \tag{66}$$

$\square$

D Computing f_i^* for ℓ_2-regularized standard surrogate losses

In this section, we explain how the values of f_i^* can be computed in closed form expressions for some standard binary surrogate losses from Bartlett et al. (2006) with ℓ_2-regularization. These closed form expressions are using the Lambert W function (Corless et al., 1996) or the more general r-Lambert function (Mező and Baricz, 2017). While these functions have efficient numerical routines to compute them (see e.g. Corless et al. (1996)), we note that we can also compute easily f_i^* for the cases in this section by solving simple strongly convex minimization problems in 1 dimension (see (72) and (76) below). This can be done efficiently to machine precision using Newton's method for example, and could be used to pre-compute f_i^* for each i in our synthetic experiments of Section 4.1.

Following Mező and Baricz (2017), we first start by presenting the definition of the Lambert W function and its recent generalization, the r-Lambert function.

Definition D.1 (Lambert W function)**.** Consider the transcendental equation

$$xe^x = a. \tag{67}$$

The inverse of the function on the left-hand side of the above equation (xe^x) is called the Lambert W function and is denoted by W. In general, W is a multivalued function on complex numbers. But for $a \geq 0$, there is a unique real solution (called the principal branch) and this is the one that we will consider in this paper, i.e. the unique real solution of (67) for $a \geq 0$ is given by $x = W(a)$. We note that $W(a)$ is a strictly increasing function for $a \geq 0$ with $W(0) = 0$, and that there are efficient numerical routines to compute it (Corless et al., 1996).

Definition D.2 (r-Lambert function)**.** The r-Lambert function is a direct generalization of the Lambert W function first proposed by Mező and Baricz (2017). It is used to express the solution of the transcendental equation

$$xe^x + rx = a, \tag{68}$$

where r is a fixed real number. The inverse of the function $xe^x + rx$ is called the r-Lambert function and is denoted by W_r. Again, there are multiple possible inverses in general; we will consider the principal branch in this paper, which is a strictly increasing function on its domain of definition (which always includes $\mathbb{R}_+$ – see Theorem 4 in Mező and Baricz (2017)). Thus $x = W_r(a)$ solution of (68), at least valid for $a \geq 0$

From the above definitions, it is clear that the classical Lambert W function is a special case of the r-Lambert function when $r = 0$. In this case we write W_0.

Manipulating the transcendental equation (68), we can easily solve the slightly more general equation as given in the following Theorem.

Theorem D.3 (Variant of Theorem 3 from Mező and Baricz (2017))**.** Let $c \in \mathbb{R}$. The equation $xe^{cx} + rx = a$ can be resolved by the r-Lambert function and the solution can be expressed as:

$$x = \frac{1}{c} W_r(ca) \tag{69}$$

In the rest of this section, we provide the analytical derivations of the closed form expression of f_i^* when the given losses are ℓ_2-regularized version of:

1. the binary log-loss. We show that in this case that f_i^* can be computed in closed form expression using the r-Lambert function.

2. the binary exponential loss. As mentioned by Bartlett et al. (2006) this loss appears in the Adaboost algorithm, amongst others. In this case, the Lambert W function is used.

D.1 Binary Log-loss

The ℓ_2-regularized logistic regression problem is given by:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-b_i\langle A_i, x\rangle}\right) + \frac{\lambda}{2}\|x\|^2, \tag{70}$$

where A_i is the input feature vector for the i^{th} datapoint while $b_i \in \{-1, 1\}$ is its label, and λ is the regularization parameter.

Note that by following the notation of the rest of the paper, in (70) we have

$$f_i(x) = \log\left(1 + e^{-b_i\langle A_i, x\rangle}\right) + \frac{\lambda}{2}\|x\|^2.$$

To simplify the notation, let us define $z_i := b_i A_i$. Let us also decompose the vector x in its direction $\hat{x}$ (element of the unit sphere, i.e. $\|\hat{x}\| = 1$) and its norm $\alpha = \|x\|$, and thus $x = \alpha\hat{x}$ for $\alpha \in \mathbb{R}$. Then we obtain the following:

$$\begin{aligned} f_i^* := \inf_x f_i(x) &= \inf_x \log\left(1 + e^{-\langle x, z_i\rangle}\right) + \frac{\lambda}{2}\|x\|^2 \\ &= \inf_\alpha \inf_{\hat{x}} \log\left(1 + e^{-\alpha\langle \hat{x}, z_i\rangle}\right) + \frac{\lambda}{2}\alpha^2 \end{aligned} \tag{71}$$

Note that $\log\left(1 + e^{-\alpha\langle \hat{x}, z_i\rangle}\right)$ is decreasing as $\langle \hat{x}, z_i\rangle$ increases. Thus, by Cauchy-Schwartz inequality, $\inf_{\hat{x}}$ is reached when $\hat{x} = \frac{z_i}{\|z_i\|}$ and equation (71) takes the following form:

$$f_i^* = \inf_\alpha \underbrace{\log\left(1 + e^{-\alpha\|z_i\|}\right) + \frac{\lambda}{2}\alpha^2}_{:=g(\alpha)} \tag{72}$$

$g(\alpha)$ is a strongly convex function of α, and we can find its global minimum by setting its gradient to zero:

$$\nabla g(\alpha) = \frac{-\|z_i\| e^{-\alpha\|z_i\|}}{1 + e^{-\alpha\|z_i\|}} + \lambda\alpha = 0.$$

Let $c = \|z_i\|$, then by rearranging the last equation, we obtain:

$$\alpha + \alpha e^{\alpha c} = \frac{c}{\lambda}.$$

Using Theorem D.3, the solution of the above equation can be expressed as follows (using $r = 1$):

$$\alpha^* = \frac{1}{c} W_1\left(\frac{c^2}{\lambda}\right). \tag{73}$$

Thus to get a closed form expression for f_i^*, one can plug α^* in (72), i.e. $f_i^* = g(\alpha^*)$.

D.2 Binary Exponential Loss

The ℓ_2-regularized binary exponential loss problem is given by:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} e^{-b_i\langle A_i, x\rangle} + \frac{\lambda}{2}\|x\|^2, \tag{74}$$

with the same notation as for the logistic regression problem (70).

From (74) it is clear that,

$$f_i(x) = e^{-b_i\langle A_i, x\rangle} + \frac{\lambda}{2}\|x\|^2.$$

As for the logistic regression derivation, defining $z_i := b_i A_i$, and letting $x = \alpha \hat{x}$ with $\|\hat{x}\| = 1$, then we obtain the following:

$$\begin{aligned} f_i^* := \inf_x f_i(x) &= \inf_x e^{-\langle x, z_i \rangle} + \frac{\lambda}{2}\|x\|^2 \\ &= \inf_\alpha \inf_{\hat{x}} e^{-\alpha\langle \hat{x}, z_i \rangle} + \frac{\lambda}{2}\alpha^2 \end{aligned} \tag{75}$$

As for the logistic regression, note that by Cauchy-Schwartz inequality, $\inf_{\hat{x}}$ is reached when $\hat{x} = \frac{z_i}{\|z_i\|}$. Thus,

$$f_i^* = \inf_\alpha \underbrace{e^{-\alpha\|z_i\|} + \frac{\lambda}{2}\alpha^2}_{:=g(\alpha)} \tag{76}$$

Again, $g(\alpha)$ is a strongly convex function of α, and we can find its global minimum by setting its gradient to zero:

$$\nabla g(\alpha) = -\|z_i\| e^{-\alpha\|z_i\|} + \lambda\alpha = 0.$$

Let $c = \|z_i\|$, then by rearranging the last equation, we obtain:

$$\alpha e^{\alpha c} = \frac{c}{\lambda}$$

Using Theorem D.3, the solution of the above equation can be expressed as follows (using $r = 0$):

$$\alpha^* = \frac{1}{c} W_0\left(\frac{c^2}{\lambda}\right), \tag{77}$$

where W_0 is the classical Lambert W function.

Thus to get a closed form expression for f_i^*, one can plug α^* in (76), i.e. $f_i^* = g(\alpha^*)$.

E Additional Experiments

Following the experiments presented in the main paper, we further evaluate the performance of SGD with SPS when training over-parametrized models.

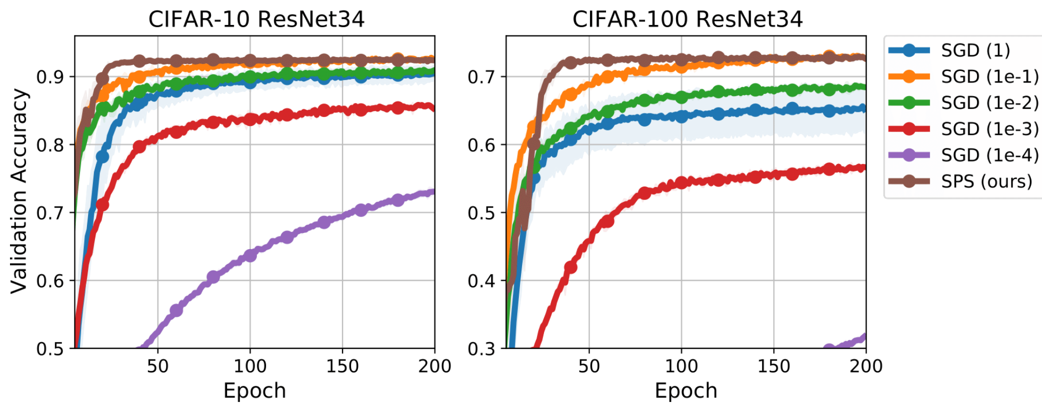

Figure 3: Comparison between SGD with different learning rates and SPS.

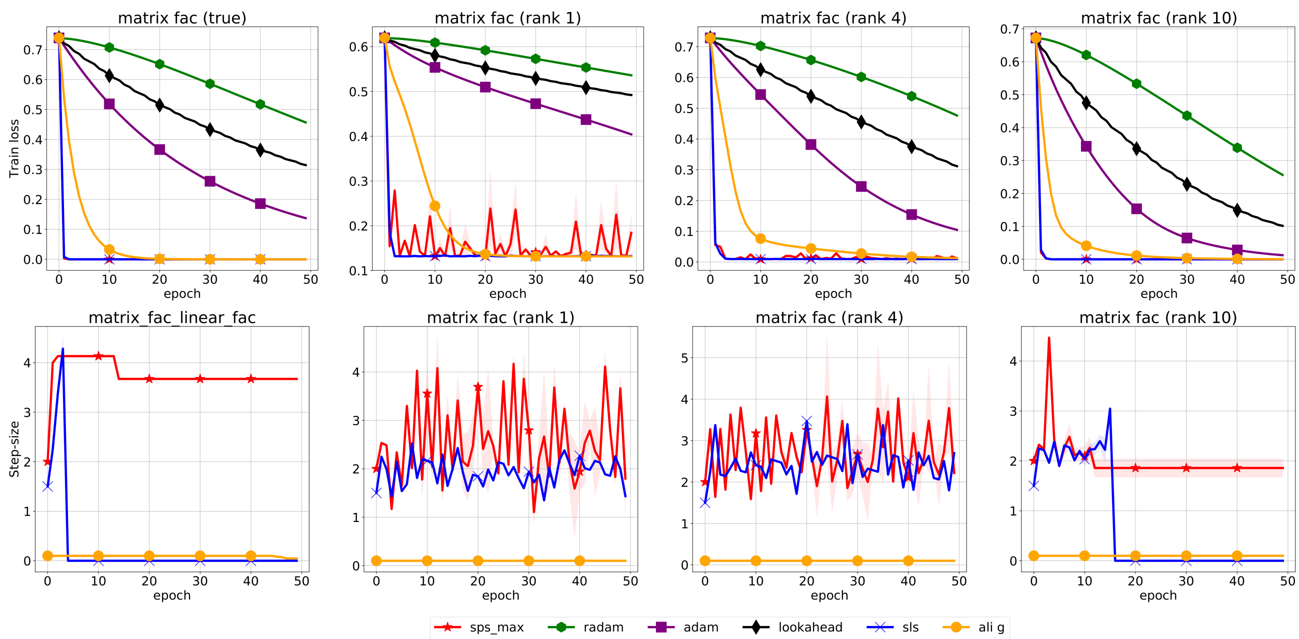

Figure 4: Deep matrix factorization.

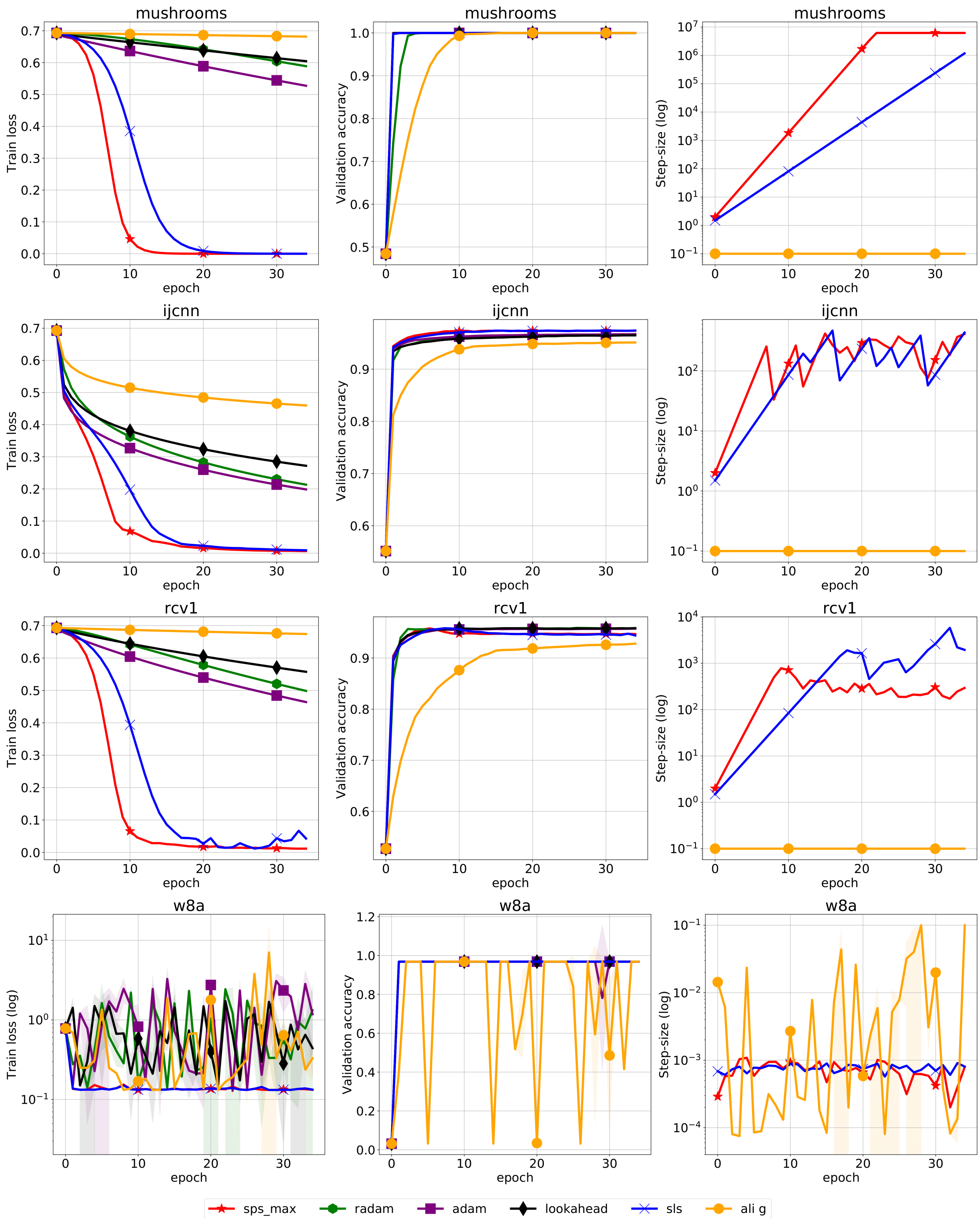

Figure 5: Binary classification using kernels. Data: mushrooms, ijcnn, rcv1, w8a.

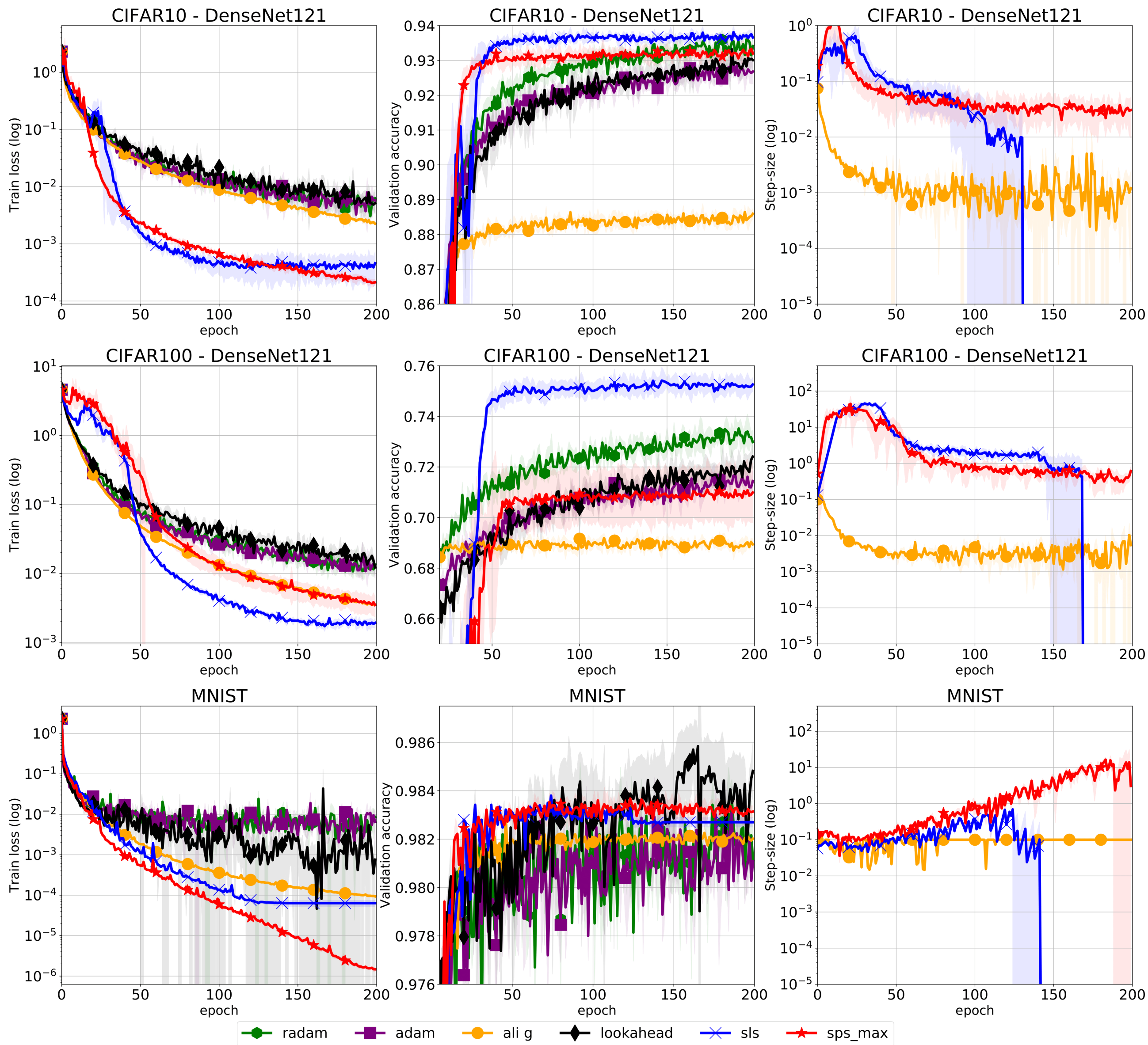

Figure 6: Further experiments on Multi-class classification using deep networks. Setting: CIFAR10-DenseNet121, CIFAR100-DenseNet121 and MNIST.