

---

# Fast and Faster Convergence of SGD for Over-Parameterized Models (and an Accelerated Perceptron)

---

Sharan Vaswani<sup>1</sup><sup>1</sup>University of British ColumbiaFrancis Bach<sup>2</sup><sup>2</sup>INRIA, ENS, PSL Research UniversityMark Schmidt<sup>1</sup>

## Abstract

Modern machine learning focuses on highly expressive models that are able to fit or *interpolate* the data completely, resulting in zero training loss. For such models, we show that the stochastic gradients of common loss functions satisfy a *strong growth condition*. Under this condition, we prove that constant step-size stochastic gradient descent (SGD) with Nesterov acceleration matches the convergence rate of the deterministic accelerated method for both convex and strongly-convex functions. We also show that this condition implies that SGD can find a first-order stationary point as efficiently as full gradient descent in non-convex settings. Under interpolation, we further show that all smooth loss functions with a finite-sum structure satisfy a *weaker growth condition*. Given this weaker condition, we prove that SGD with a constant step-size attains the deterministic convergence rate in both the strongly-convex and convex settings. Under additional assumptions, the above results enable us to prove an  $O(1/k^2)$  mistake bound for  $k$  iterations of a stochastic perceptron algorithm using the squared-hinge loss. Finally, we validate our theoretical findings with experiments on synthetic and real datasets.

## 1 Introduction

Modern machine learning models are typically trained with iterative stochastic first-order methods [9, 41, 16, 31, 14, 8]. Stochastic gradient descent (SGD) and related methods such as Adagrad [9] or Adam [16]

compute the gradient with respect to one or a mini-batch of training examples in each iteration and take a descent step using this gradient. Since these methods use only a small part of the data in each iteration, they are the preferred way for training models on large datasets. However, in order to converge to the solution, these methods require the step-size to decay to zero in terms of the number of iterations. This implies that the gradient descent procedure takes smaller steps as the training progresses. Consequently, these methods result in slow sub-linear rates of convergence. Specifically, if  $k$  is the number of iterations, then SGD-like methods achieve a convergence rate of  $O(1/k)$  and  $O(1/\sqrt{k})$  for strongly-convex and convex functions respectively [23]. In practice, these methods are augmented with some form of momentum or acceleration [27, 25] that results in faster empirical convergence [37]. Recently, there has been some theoretical analysis for the use of such acceleration in the stochastic setting [7]. Other related work includes algorithms specifically designed to achieve an accelerated rate of convergence in the stochastic setting [1, 19, 10].

Another recent trend in the literature has been to use variance-reduction techniques [31, 14, 8] that exploit the finite-sum structure of the loss function in machine-learning applications. These methods do not require the step-size to decay to zero and are able to achieve the optimal rate of convergence. However, they require additional bookkeeping [31, 8] or need to compute the full gradient periodically [14], both of which are difficult in the context of training complex models on large datasets.

In this paper, we take further advantage of the optimization properties specific to modern machine learning models. In particular, we make use of the fact that models such as non-parametric regression or over-parameterized deep neural networks are expressive enough to fit or *interpolate* the training dataset completely [42, 22]. For an SGD-like algorithm, this implies that the gradient with respect to each training example converges to zero at the optimal solution. This property of interpolation is also true for boosting [30]

---

Proceedings of the 22<sup>nd</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

and for simple linear classifiers on separable data. For example, the perceptron algorithm [29] was first shown to converge to the optimal solution under a linear separability assumption on the data [26]. This assumption implies that the linear perceptron is able to fit the complete dataset without errors.

There has been some related work that takes advantage of the interpolation property in order to obtain faster rates of convergence for SGD [32, 22, 5]. Specifically, Schmidt and Le Roux [32] assume a *strong growth condition* on the stochastic gradients. This condition relates the  $\ell_2$  norms of the stochastic gradients to that of the full gradient. Under this assumption, they prove that constant step-size SGD can attain the same convergence rates as full gradient descent in both the strongly-convex and convex cases. Other related work has used the strong growth condition to prove convergence rates for incremental gradient methods [34, 38]. Ma et al. [22] show that under weaker conditions, SGD with constant step-size results in linear convergence for strongly-convex functions. They also investigate the effect of batch-size on the convergence and theoretically justify the *linear-scaling rule* used for training deep learning models in practice [12]. Recently, Cevher and Vũ showed the linear convergence of proximal stochastic gradient descent under a weaker growth condition for restricted strongly convex functions [5]. They also analyse the effect of an additive error term on the convergence rate.

In contrast to the above mentioned work, we first show that the strong growth condition (SGC) [32] implies that SGD with a constant step-size and Nesterov momentum [25] achieves the accelerated convergence rate of the deterministic setting for both strongly-convex and convex functions (Section 3). Our result gives some theoretical justification behind the empirical success of using Nesterov acceleration with SGD [37]. Further, in Section 4 we consider non-convex objectives and prove under the SGC that constant step-size SGD is able to find a first-order stationary point as efficiently as deterministic gradient descent. To the best of our knowledge, this is the first work to study accelerated and non-convex rates under the SGC.

After the release of the first version of this work, Liu et al. [20] also considered minimizing strongly-convex loss functions using a variant of Nesterov acceleration assuming interpolation. In this setting they show accelerated rates for the squared loss, and under additional assumptions give accelerated rates for general strongly-convex functions. However, it is not clear if these additional assumptions are satisfied by common loss functions. Indeed, these additional assumptions imply the SGC (see Section 6.1) so the result presented in Section 3 is more widely-applicable. Similarly, the

work of Jain et al. [13] uses tail-averaging to obtain accelerated rates but only for the special case of the squared loss under interpolation. Furthermore, unlike these works, we show accelerated rates for convex functions (that are not strongly-convex) under the SGC.

Another work appearing after the release of the initial version of this work is Bassily et al. [3], who considered minimizing non-convex functions satisfying the Polyak-Łojasiewicz [28] (PL) inequality (a generalization of strong-convexity) under the interpolation condition. This is a much stronger assumption than we make in Section 4 to analyze non-convex functions (since it implies all local optima are global optima), but under this condition they show that SGD can achieve a linear convergence rate. However, the step-size needed to achieve this rate is proportional to the PL constant which is typically extremely small (and is often unknown and difficult to estimate). By exploiting the stronger SGC, in this version of the paper we have added a result under the PL inequality (Section 4) that achieves a faster rate by using a step-size that depends only on the smoothness properties of the functions.

In this work, we also relax the strong growth condition to a more practical *weak growth condition* (WGC). In Section 5, we prove that the weak growth condition is sufficient to obtain the optimal convergence of constant step-size SGD for smooth strongly-convex and convex functions. To demonstrate the applicability of our growth conditions in practice, we first show that for models interpolating the data, the WGC is satisfied for all smooth and convex loss functions with a finite-sum structure (Section 6.1). Furthermore, we prove that functions satisfying the WGC and the PL condition also satisfy the SGC. Under additional assumptions, we show that it is also satisfied for the squared-hinge loss. This result enables us to prove an  $O(1/k^2)$  mistake bound for  $k$  iterations of an accelerated stochastic perceptron algorithm using the squared-hinge loss (Section 7). Finally, in Section 8, we evaluate our claims with experiments on synthetic and real datasets.

## 2 Background

In this section, we give the required background and set up the necessary notation. Our aim is to minimize a differentiable function  $f(w)$ . Depending on the context, this function can be strongly-convex, convex or non-convex. We assume that we have access to noisy gradients for the function  $f$  and use stochastic gradient descent (SGD) for  $k$  iterations in order to minimize it. The SGD update rule in iteration  $k$  can be written as:  $w_{k+1} = w_k - \eta_k \nabla f(w_k, z_k)$ . Here,  $w_{k+1}$  and  $w_k$  are

the SGD iterates,  $z_k$  is the gradient noise and  $\eta_k$  is the step-size at iteration  $k$ . We assume that the gradients  $\nabla f(w, z)$  are unbiased, implying that for all  $w$  and  $z$  that  $\mathbb{E}_z [\nabla f(w, z)] = \nabla f(w)$ .

While most of our results apply for general SGD methods, a subset of our results rely on the function  $f(w)$  having a finite-sum structure meaning that  $f(w) = \frac{1}{n} \sum_{i=1}^n f_i(w)$ . In the context of supervised machine learning, given a training dataset of  $n$  points, the term  $f_i(w)$  corresponds to the loss function for the point  $(x_i, y_i)$  when the model parameters are equal to  $w$ . Here  $x_i$  and  $y_i$  refer to the feature vector and label for point  $i$  respectively. Common choices of the loss function include the squared loss where  $f_i(w) = \frac{1}{2} (w^\top x_i - y_i)^2$ , the hinge loss where  $f_i(w) = \max(0, 1 - y_i w^\top x_i)$  or the squared-hinge loss where  $f_i(w) = \max(0, 1 - y_i w^\top x_i)^2$ . The finite sum setting includes both simple models such as logistic regression or least squares and more complex models like non-parametric regression and deep neural networks.

In the finite-sum setting, SGD consists of choosing a point and its corresponding loss function (typically uniformly) at random and evaluating the gradient with respect to that function. It then performs a gradient descent step:  $w_{k+1} = w_k - \eta_k \nabla f_k(w_k)$  where  $f_k(\cdot)$  is the random loss function selected at iteration  $k$ . The unbiasedness property is automatically satisfied in this case, i.e.  $\mathbb{E}_i [\nabla f_i(w)] = \nabla f(w)$  for all  $w$ . Note that in this case, the random selection of points for computing the gradient is the source of the noise  $z_k$ . In order to converge to the optimum, SGD requires the step-size  $\eta_k$  to decrease with  $k$ ; specifically at a rate of  $\frac{1}{\sqrt{k}}$  for convex functions and at a  $\frac{1}{k}$  rate for strongly-convex functions. Decreasing the step-size with  $k$  results in sub-linear rates of convergence for SGD.

In order to derive convergence rates, we need to make additional assumptions about the function  $f$  [23]. Beyond differentiability, our results assume that the function  $f(\cdot)$  satisfies some or all of the following common assumptions. For all points  $w, v$  and for constants  $f^*$ ,  $\mu$ , and  $L$ ;

$$\begin{aligned} f(w) &\geq f^* && \text{(Bounded below)} \\ f(v) &\geq f(w) + \langle \nabla f(w), v - w \rangle && \text{(Convexity)} \\ f(v) &\geq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\mu}{2} \|v - w\|^2 && (\mu \text{ Strong-convexity}) \\ f(v) &\leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{L}{2} \|v - w\|^2 && (L \text{ Smoothness}) \end{aligned}$$

Note that some of our results in Section 6 rely on the finite-sum structure and we explicitly state when we need this additional assumption.

In this paper, we consider the case where the model is able to *interpolate* or fit the labelled training data completely. This is true for expressive models such as non-parametric regression and over-parametrized deep neural networks. For common loss functions that are lower-bounded by zero, interpolating the data results in zero training loss. Interpolation also implies that the gradient with respect to each point converges to zero at the optimum. Formally, in the finite-sum setting, if the function  $f(\cdot)$  is minimized at  $w^*$ , i.e., if  $\nabla f(w^*) = 0$ , then for all functions  $f_i(\cdot)$ ,  $\nabla f_i(w^*) = 0$ .

The *strong growth condition* (SGC) used connects the rates at which the stochastic gradients shrink relative to the full gradient. Formally, for any point  $w$  and the noise random variable  $z$ , the function  $f$  satisfies the strong growth condition with constant  $\rho$  if,

$$\mathbb{E}_z \|\nabla f(w, z)\|^2 \leq \rho \|\nabla f(w)\|^2. \quad (1)$$

Equivalently, in the finite-sum setting,

$$\mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2. \quad (2)$$

For this inequality to hold, if  $\nabla f(w) = 0$ , then  $\nabla f_i(w) = 0$  for all  $i$ . Thus, functions satisfying the SGC necessarily satisfy the above interpolation property. Schmidt and Le Roux's work [32] derives optimal convergence rates for constant step-size SGD under the above condition for both convex and strongly-convex functions. In the next section, we show that the SGC implies the accelerated rate of convergence for constant step-size SGD with Nesterov momentum.

### 3 SGD with Nesterov acceleration under the SGC

We first describe constant step-size SGD with Nesterov acceleration. The algorithm consists of three sequences  $(w_k, \zeta_k, v_k)$  updated in each iteration [24]. Specifically, it consists of the following update rules:

$$w_{k+1} = \zeta_k - \eta \nabla f(\zeta_k, z_k) \quad (3)$$

$$\zeta_k = \alpha_k v_k + (1 - \alpha_k) w_k \quad (4)$$

$$v_{k+1} = \beta_k v_k + (1 - \beta_k) \zeta_k - \gamma_k \eta \nabla f(\zeta_k, z_k). \quad (5)$$

Here,  $\eta$  is the constant step-size for the SGD step and  $\alpha_k, \beta_k, \gamma_k$  are tunable parameters to be set according to the properties of  $f$ .

In order to derive a convergence rate for the above algorithm under the SGC, we first observe that a form of the SGC is satisfied in the case of coordinate descent [39]. In this case, we choose a coordinate (typically at random) and perform a gradient descent step with respect to that coordinate. The notion of a coordinate in this case is analogous to that of an individual

loss function in the finite sum case. For coordinate descent, a zero gradient at the optimal solution implies that the partial derivative with respect to each coordinate is also equal to zero. This is analogous to the SGC in the finite-sum case, although we note the results in this section do not require the finite-sum assumption.

We use this analogy formally in order to extend the proof of Nesterov's accelerated coordinate descent [24] to derive convergence rates for the above algorithm when using the SGC. This enables us to prove the following theorems (with proofs in Appendices B.1.1 and B.1.3) in both the strongly-convex and convex settings.

**Theorem 1** (Strongly convex). *Under  $L$ -smoothness and  $\mu$  strong-convexity, if  $f$  satisfies the SGC with constant  $\rho$ , then SGD with Nesterov acceleration with the following choice of parameters,*

$$\begin{aligned}\gamma_k &= \frac{1}{\sqrt{\mu\eta\rho}} \quad ; \quad \beta_k = 1 - \sqrt{\frac{\mu\eta}{\rho}} \\ b_{k+1} &= \frac{\sqrt{\mu}}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \\ a_{k+1} &= \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \\ \alpha_k &= \frac{\gamma_k \beta_k b_{k+1}^2 \eta}{\gamma_k \beta_k b_{k+1}^2 \eta + a_k^2}; \quad \eta = \frac{1}{\rho L},\end{aligned}$$

results in the following convergence rate:

$$\begin{aligned}\mathbb{E}f(w_{k+1}) - f(w^*) \\ \leq \left(1 - \sqrt{\frac{\mu}{\rho^2 L}}\right)^k \left[f(w_0) - f(w^*) + \frac{\mu}{2} \|w_0 - w^*\|^2\right].\end{aligned}$$

**Theorem 2** (Convex). *Under  $L$ -smoothness and convexity, if  $f$  satisfies the SGC with constant  $\rho$ , then SGD with Nesterov acceleration with the following choice of parameters,*

$$\begin{aligned}\gamma_k &= \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{k-1}^2}}{2} \\ a_{k+1} &= \gamma_k \sqrt{\eta\rho} \\ \alpha_k &= \frac{\gamma_k \eta}{\gamma_k \eta + a_k^2}; \quad \eta = \frac{1}{\rho L},\end{aligned}$$

results in the following convergence rate:

$$\mathbb{E}f(w_{k+1}) - f(w^*) \leq \frac{2\rho^2 L}{k^2} \|w_0 - w^*\|^2.$$

The above theorems show that constant step-size SGD with Nesterov momentum achieves the accelerated

rate of convergence up to a  $\rho^2$  factor for both strongly-convex and convex functions.

In Appendix A, we consider the SGC with an extra additive error term, resulting in the following condition:  $\mathbb{E}_z \|\nabla f(w, z)\|^2 \leq \rho \|\nabla f(w)\|^2 + \sigma^2$ . We analyse the rate of convergence of the above algorithm under this modified condition and obtain a similar dependence on  $\sigma$  as in Cohen et al. [7].

## 4 SGD for non-convex functions satisfying the SGC

In this section, we show that the SGC results in an improvement over the  $O(1/\sqrt{k})$  rate for SGD in the non-convex setting [11]. In particular, we show that under the strong growth condition, constant step-size SGD is able to find a first-order stationary point as efficiently as deterministic gradient descent. We prove the following theorem (with the proof in Appendix B.2),

**Theorem 3** (Non-Convex). *Under  $L$ -smoothness, if  $f$  satisfies SGC with constant  $\rho$ , then SGD with a constant step-size  $\eta = \frac{1}{\rho L}$  attains the following convergence rate:*

$$\min_{i=0,1,\dots,k-1} \mathbb{E} \left[ \|\nabla f(w_i)\|^2 \right] \leq \left( \frac{2\rho L}{k} \right) [f(w_0) - f^*].$$

The above theorem shows that under the SGC, SGD with a constant step-size can attain the optimal  $O(1/k)$  rate for non-convex functions. To the best of our knowledge, this is the first result for non-convex functions under interpolation-like conditions. Under these conditions, constant step-size SGD has a better convergence rate than algorithms which have recently been proposed to improve on SGD [2, 4]. Note that the above theorem applies to neural networks with a sigmoid activation function under the assumption that the strong-growth condition is satisfied. Hence, our results also provide some theoretical justification for the effectiveness of SGD for non-convex over-parameterized models like deep neural networks.

Under the additional assumption that the function satisfies the Polyak-Lojasiewicz condition [28] (a generalization of strong-convexity), we show that SGD can obtain linear convergence. Specifically, we prove the following theorem (with the proof in Appendix B.3),

**Theorem 4** (Non-Convex + PL). *Under  $L$ -smoothness, if  $f$  satisfies SGC with constant  $\rho$  and the Polyak-Lojasiewicz inequality with constant  $\mu$ , then SGD with a constant step-size  $\eta = \frac{1}{\rho L}$  attains the fol-*

lowing convergence rate:

$$\mathbb{E}[f(w_{k+1}) - f^*] \leq \left(1 - \frac{\mu}{\rho L}\right)^k [f(w_0) - f^*].$$

Note that the PL condition or a related notion of restricted strong-convexity (RSI) [15] is satisfied by numerous non-convex optimization problems of interest. These include neural networks [18, 17, 35], matrix completion [36] and phase retrieval [6]. Under the additional SGC assumption, the above theorem implies fast rates of convergence for SGD on these problems. In contrast, Bassily et al. [3] do not assume the SGC and achieve a rate of  $\left(1 - \frac{\mu^2}{L^2}\right)^k$  using a much smaller step-size  $\eta = \frac{\mu}{L^2}$ .

## 5 Weak growth condition

In this section, we relax the strong growth condition to a more practical condition which we refer to as the *weak growth condition* (WGC). Formally, if the function  $f(\cdot)$  is  $L$ -smooth and has a minima at  $w^*$ , then it satisfies the WGC with constant  $\rho$ , if for all points  $w$  and noise random variable  $z$ ,

$$\mathbb{E}_z \|\nabla f(w, z)\|^2 \leq 2\rho L[f(w) - f(w^*)]. \quad (6)$$

Equivalently, in the finite-sum setting,

$$\mathbb{E}_i \|\nabla f_i(w)\|^2 \leq 2\rho L[f(w) - f(w^*)]. \quad (7)$$

In the above condition, notice that if  $w = w^*$ , then  $\nabla f_i(w^*) = 0$  for all points  $i$ . Thus, the WGC implies the interpolation property explained in Section 2.

### 5.1 Relation between WGC and SGC

In this section, we relate the two growth conditions. We first prove that SGC implies WGC with the same  $\rho$  without any additional assumptions, formally showing that the WGC is indeed weaker than the corresponding SGC. For the converse, a function satisfying the WGC satisfies the SGC with a worse constant if it also satisfies the Polyak-Lojasiewicz (PL) inequality [28]. The above relations are captured by the following proposition, proved in Appendix B.6

**Proposition 1.** *If  $f(\cdot)$  is  $L$ -smooth, satisfies the WGC with constant  $\rho$  and the PL inequality with constant  $\mu$ , then it satisfies the SGC with constant  $\frac{\rho L}{\mu}$ .*

*Conversely, if  $f(\cdot)$  is  $L$ -smooth, convex and satisfies the SGC with constant  $\rho$ , then it also satisfies the WGC with the same constant  $\rho$ .*

### 5.2 SGD under the weak growth condition

Using the WGC, we obtain the following convergence rates for SGD with a constant step-size.

**Theorem 5** (Strongly-convex). *Under  $L$ -smoothness and  $\mu$  strong-convexity, if  $f$  satisfies the WGC with constant  $\rho$ , then SGD with a constant step-size  $\eta = \frac{1}{\rho L}$  achieves the following rate:*

$$\mathbb{E} \|w_{k+1} - w^*\|^2 \leq \left(1 - \frac{\mu}{\rho L}\right)^k \|w_0 - w^*\|^2.$$

**Theorem 6** (Convex). *Under  $L$ -smoothness and convexity, if  $f$  satisfies the WGC with constant  $\rho$ , then SGD with a constant step-size  $\eta = \frac{1}{4\rho L}$  and iterate averaging achieves the following rate:*

$$\mathbb{E}[f(\bar{w}_k)] - f(w^*) \leq \frac{4L(1+\rho)\|w_0 - w^*\|^2}{k}.$$

Here,  $\bar{w}_k = \frac{[\sum_{i=1}^k w_i]}{k}$  is the averaged iterate after  $k$  iterations.

The proofs for Theorems 5 and 6 are deferred to Appendices B.4 and B.5 respectively. In these cases, the WGC is sufficient to show that constant step-size SGD can attain the deterministic rates up to a factor of  $\rho$ . Since this condition is weaker than the corresponding strong growth condition, our results subsume the SGC results [32]. Note that an alternative way to obtain the result in Theorem 5 would be to observe that the WGC and strong convexity imply the SGC (with a constant  $\frac{\rho L}{\mu}$ ) (Proposition 1) and then use the result by Schmidt et al. [32]. This would result in an additional dependence on  $\frac{\mu}{\rho L}$  which is worse than the rate in Theorem 5.

In the next section, we characterize the functions satisfying the growth conditions in practice.

## 6 Growth conditions in practice

In this section, we give examples of functions that satisfy the weak and strong growth conditions. In Section 6.1, we first show that for models interpolating the data, the WGC is satisfied by all smooth functions with a finite-sum structure. In section 6.2, we show that the SGC is satisfied by the squared-hinge loss under additional assumptions.

### 6.1 Functions satisfying WGC

To characterize the functions satisfying the WGC, we first prove the following proposition (with the proof in Appendix B.7):

**Proposition 2.** *If the function  $f(\cdot)$  is convex and has a finite-sum structure for a model that interpolates the data and  $L_{\max}$  is the maximum smoothness constant amongst the functions  $f_i(\cdot)$ , then for all  $w$ ,*

$$\mathbb{E}_i \|\nabla f_i(w)\|^2 \leq 2L_{\max} [f(w) - f(w^*)]. \quad (8)$$

Comparing the above equation to Equation 7, we see that any smooth finite-sum problem under interpolation satisfies the WGC with  $\rho = \frac{L_{\max}}{L}$ . The WGC is thus satisfied by common loss functions such as the squared and squared-hinge loss. For these loss functions, if  $L_i = L$  for all  $i$ , then Theorem 5 implies that SGD with  $\eta = \frac{1}{L}$  results in linear convergence for strongly-convex functions. This matches the recently proved result of Ma et al. [22], whereas Theorem 6 allows us to generalize their result beyond strongly-convex functions.

## 6.2 Functions satisfying SGC

We now show that under additional assumptions on the data, the squared-hinge loss also satisfies the SGC. We first assume that the data is linearly separable with a margin equal to  $\tau$ , implying that for all  $x$ ,  $\tau = \max_{|w|=1} \inf_{x \in \mathcal{S}} w^\top x$ . Here,  $\mathcal{S}$  is the support of the distribution of the features  $x$ . Note that the above assumption implies the existence of a classifier  $w^*$  such that  $\|w^*\| = \frac{1}{\tau}$ . In addition to this, we assume that the features have a finite support, meaning that the set  $\mathcal{S}$  is finite and has a cardinality equal to  $c$ . Under these assumptions, we prove the following lemma in Appendix B.8,

**Lemma 1.** *For linearly separable data with margin  $\tau$  and a finite support of size  $c$ , the squared-hinge loss satisfies the SGC with the constant  $\rho = \frac{c}{\tau^2}$ .*

In the next section, we use the above lemma to prove a mistake bound for the perceptron algorithm using the squared-hinge loss.

## 7 Implication for Faster Perceptron

In this section, we use the strong growth property of the squared-hinge function in order to prove a bound on the number of mistakes made by the perceptron algorithm [29] using a squared-hinge loss. The perceptron algorithm is used for training a linear classifier for binary classification and is guaranteed to converge for linearly separable data [26]. It can be considered as stochastic gradient descent on the loss  $f_i(w) = \max\{0, y_i x_i^\top w\}$ .

The common way to characterize the performance of a perceptron is by bounding the number of mistakes

(in the binary classification setting) after  $k$  iterations of the algorithm. In other words, we care about the quantity  $\mathbb{P}(yx^\top w_k \geq 0)$ . Assuming linear separability of the data and that  $\|x\| = 1$  for all points  $(x, y)$ , the perceptron achieves a mistake bound of  $O(\frac{1}{\tau^2})$  [26].

In this paper, we consider a modified perceptron algorithm using the squared-hinge function as the loss. Note that since we assume the data to be linearly separable, a linear classifier is able to fit all the training data. Since the squared-hinge loss function is smooth, the conditions of Proposition 2 are satisfied, which implies that it satisfies the WGC with  $\rho = \frac{L_{\max}}{L}$ . Also observe that since we assume that  $\|x\| = 1$ ,  $L_{\max} = L = 1$ . Using these facts with Theorem 6 and assuming that we start the optimization with  $w_0 = \mathbf{0}$ , we obtain the following convergence rate using SGD with  $\eta = 1/4$ ,

$$\mathbb{E}[f(w_{k+1})] \leq \frac{8}{\tau^2 k}.$$

To see this, recall that  $\|w^*\| = \frac{1}{\tau}$  and the loss is equal to zero at the optima, implying that  $f(w^*) = 0$ .

The above result gives us a bound on the training loss. We use the following lemma (proved using the Markov inequality in Appendix B.9) to relate the mistake bound to the training loss.

**Lemma 2.** *If  $f(w, x, y)$  represents the loss on the point  $(x, y)$ , then*

$$\mathbb{P}(yx^\top w \leq 0) \leq \mathbb{E}_{x,y} f(w, x, y).$$

Combining the above results, we obtain a mistake bound of  $O(\frac{1}{\tau^2 k})$  when using the squared-hinge loss on linearly separable data. We thus recover the standard results for the stochastic perceptron.

Note that for a finite amount of data (when the expectation is with respect to a discrete distribution), if we use batch accelerated gradient descent (which is not one of the stochastic gradient algorithms studied in this paper, and for which no growth condition is needed), we obtain a mistake bound that decreases as  $1/k^2$ . This improves on existing mistake bounds that scale as  $1/k$  [33, 40]. Note that both sets of algorithms have the same dependence on the margin  $\tau$ , but this deterministic accelerated method would require evaluating  $n$  gradients on each iteration.

From Lemma B.9, we know that the squared-hinge loss satisfies the SGC with  $\rho = \frac{c}{\tau^2}$ . Under the same conditions as above, this lemma along with the result of Theorem 2 gives us the following bound:

$$\mathbb{E} f(w_{k+1}) \leq \frac{2c^2}{\tau^6 k^2}.$$

Using the result from Lemma 2, this results in a mistake bound of the order  $O\left(\frac{1}{\tau^6 k^2}\right)$  while only requiring one gradient per iteration. Hence, the use of acceleration leads to an improved novel dependence of  $O(1/k^2)$ , but requires the additional assumptions of Lemma B.9 and has a worse dependence on the margin  $\tau$ .

## 8 Experiments

In this section, we empirically validate our theoretical results. For the first set of experiments (Figures 1(a)-1(d)), we generate a synthetic binary classification dataset with  $n = 8000$  and the dimension  $d = 100$ . We ensure that the data is linearly separable with a margin  $\tau$ , thus satisfying the interpolation property for training a linear classifier. We seek to minimize the finite-sum squared-hinge loss,  $f(w) = \sum_{i=1}^n \max(0, 1 - y_i x_i^\top w)^2$ . In Figure 1, we vary the margin  $\tau$  and plot the logarithm of the loss with the number of effective passes (one pass is equal to  $n$  iterations of SGD) over the data. In all of our experiments, we estimate the value of the smoothness parameter  $L$  as the maximum eigenvalue of the Gram matrix  $X^T X$ .

We evaluate the performance of constant step-size SGD with and without acceleration. Since the squared-hinge loss satisfies the WGC with  $\rho = \frac{L_{\max}}{L}$  (Proposition 2), we use SGD with a constant step-size  $\eta = 1/L_{\max}$ <sup>1</sup> (denoted as SGD in the plots). For using Nesterov acceleration, we experimented with the dependence of the margin  $\tau$  on the constant  $\rho$  in the SGC. We found that setting  $\rho = 1/\tau$  results in consistently stable but fast convergence across different choices of  $\tau$ . We thus use a step-size  $\eta = \tau/L$  and set the tunable parameters in the update Equations 3-5 as specified by Theorem 2. We denote this variant of accelerated SGD as Acc-SGD in the subsequent plots. In Appendix C, we propose a line-search heuristic to dynamically estimate the value of  $\rho$ .

In each of the Figures 1(a)-1(d), we make the following observations: (i) SGD results in reasonably slow convergence. This observation is in line with other SGD methods using  $1/L$  as the step-size [31]. (ii) Acc-SGD with  $\eta = \tau/L$  is consistently stable and as suggested by the theory, it results in faster convergence as compared to using SGD. (iii) For larger values of  $\tau$  (Figures 1(a)- 1(b)), the training loss becomes equal to zero, verifying the interpolation property.

The next set of experiments (Figure 2) considers bi-

nary classification on the CovType<sup>2</sup> and Protein<sup>3</sup> datasets. For this, we train a linear classifier using the radial basis (non-parametric) features. Non-parametric regression models of this form are capable of interpolating the data [22] and thus satisfy our assumptions. We subsample  $n = 8000$  random points from the datasets and use the squared-hinge loss as above. In this case, we perform a grid-search to obtain a good estimate of  $\rho$ . We choose  $\rho = 1$  for the CovType dataset and equal to 0.1 for the Protein dataset.

From Figures 2(a) and 2(b), we make the following observations: (i) In Figure 2(a), both variants have similar performance. (ii) In Figure 2(b), the Acc-SGD leads to considerably faster convergence as compared to SGD. These experiments show that in cases where the interpolation property is satisfied, both SGD and accelerated SGD with a constant step-size can result in good empirical performance.

## 9 Conclusion

In this paper, we showed that under interpolation, the stochastic gradients of common loss functions satisfy specific growth conditions. Under these conditions, we proved that it is possible for constant step-size SGD (with and without Nesterov acceleration) to achieve the convergence rates of the corresponding deterministic settings. These are the first results achieving optimal rates in the accelerated and non-convex settings under interpolation-like conditions. We used these results to demonstrate the fast convergence of the stochastic perceptron algorithm employing the squared-hinge loss. We showed that both SGD and accelerated SGD with a constant step-size can lead to good empirical performance when the interpolation property is satisfied. As opposed to determining the step-size and the schedule for annealing it for current SGD-like methods, our results imply that under interpolation, we only need to automatically determine the constant step-size for SGD. In the future, we hope to develop line-search techniques for automatically determining this step-size for both the accelerated and non-accelerated variants.

<sup>1</sup>Note that using  $\eta = 1/L_{\max}$  lead to consistently better results as compared to using  $\eta = 1/4L_{\max}$  as suggested by Theorem 6.

<sup>2</sup><http://osmot.cs.cornell.edu/kddcup>

<sup>3</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>

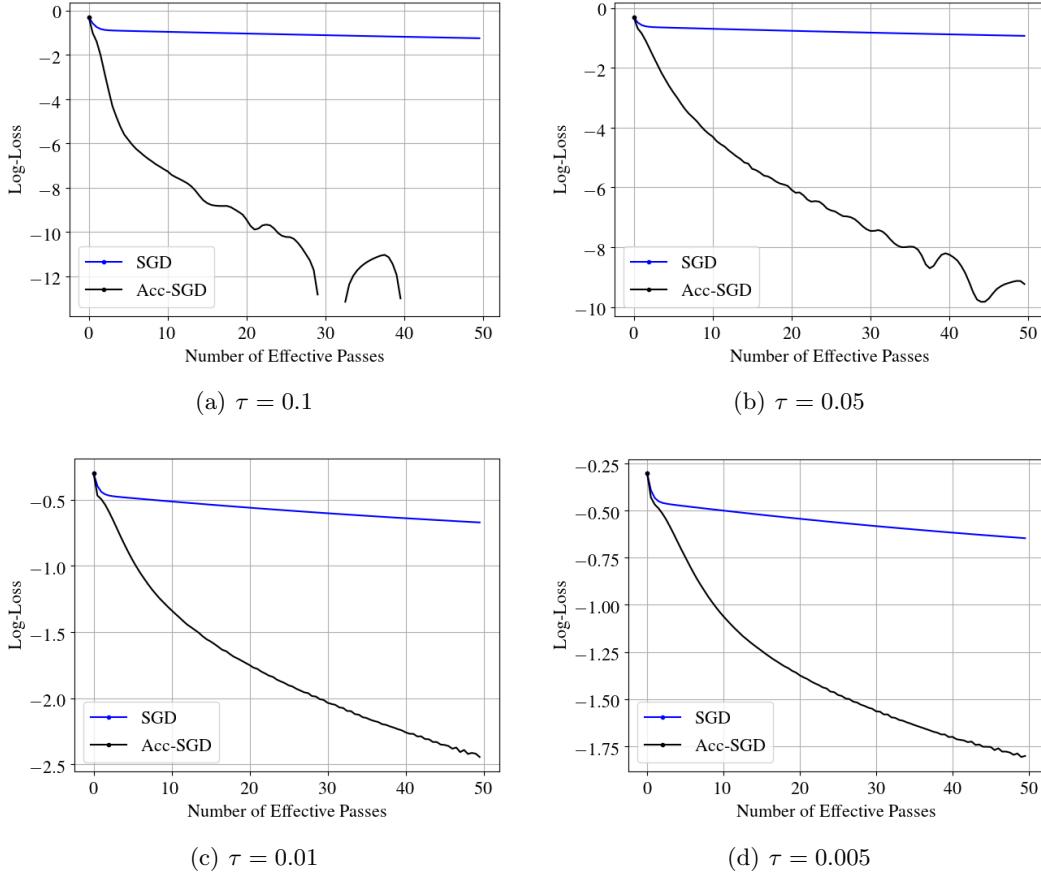


Figure 1: Comparison of SGD and variants of accelerated SGD on a synthetic linearly separable dataset with margin  $\tau$ . Accelerated SGD with  $\eta = \tau/L$  leads to faster convergence as compared to SGD with  $\eta = 1/L$ .

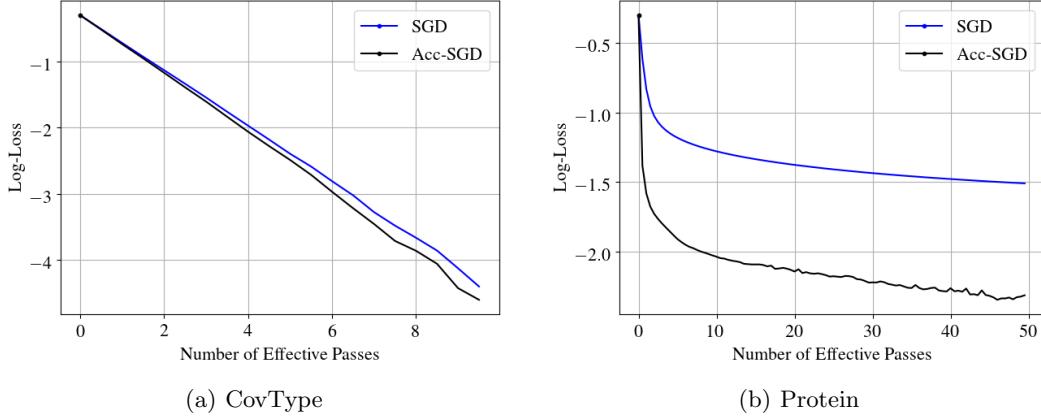


Figure 2: Comparison of SGD and accelerated SGD for learning a linear classifier with RBF features on the (a) CovType and (b) Protein datasets. Accelerated SGD leads to better performance as compared to SGD with  $\eta = 1/L$ .

## 10 Acknowledgements

We acknowledge support from the European Research Council (grant SEQUOIA 724063) and the CIFAR

program on Learning with Machines and Brains. We also thank Nicolas Flammarion, Reza Babanezhad and Adrien Taylor for discussions related to this work. We also thank Kevin Scaman for discussions and insights

on using acceleration with multiplicative noise.

## References

- [1] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.
- [2] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In *Advances in Neural Information Processing Systems*, pages 2680–2691, 2018.
- [3] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.
- [4] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 654–663. JMLR.org, 2017.
- [5] Volkan Cevher and Bang Cong Vu. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, pages 1–11, 2018.
- [6] Yuxin Chen and Emmanuel Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.
- [7] Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsm  ssan, Stockholm, Sweden, July 10-15, 2018*, pages 1018–1027, 2018.
- [8] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.
- [9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- [10] Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *International Conference on Machine Learning*, pages 2540–2548, 2015.
- [11] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [12] Priya Goyal, Piotr Doll  r, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [13] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 545–604, 2018.
- [14] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.
- [15] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.
- [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [17] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pages 2703–2712, 2018.
- [18] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.
- [19] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.
- [20] Chaoyue Liu and Mikhail Belkin. Mass: an accelerated stochastic method for over-parametrized learning. *arXiv preprint arXiv:1810.13395*, 2018.

- [21] Jun Liu, Jianhui Chen, and Jieping Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.
- [22] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3331–3340, 2018.
- [23] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
- [24] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- [25] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
- [26] Albert B Novikoff. On convergence proofs for perceptrons. Technical report, 1963.
- [27] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- [28] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.
- [29] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [30] Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.
- [31] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.
- [32] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.
- [33] Negar Soheili and Javier Pena. A primal-dual smooth perceptron-von neumann algorithm. In *Discrete Geometry and Optimization*, pages 303–320. Springer, 2013.
- [34] Mikhail V Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.
- [35] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.
- [36] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.
- [37] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- [38] Paul Tseng. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.
- [39] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.
- [40] Adams Wei Yu, Fatma Kilinc-Karzan, and Jaime Carbonell. Saddle points and accelerated perceptron algorithms. In *International Conference on Machine Learning*, pages 1827–1835, 2014.
- [41] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
- [42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

## A Incorporating additive error for Nesterov acceleration

For this section, we assume an additive error in the the strong growth condition implying that the following equation is satisfied for all  $w, z$ .

$$\mathbb{E}_z \|\nabla f(w, z)\|^2 \leq \rho \|\nabla f(w)\|^2 + \sigma^2$$

In this case, we have the counterparts of Theorems 1 and 2 as follows:

**Theorem 7** (Strongly convex). *Under  $L$ -smoothness and  $\mu$  strongly-convexity, if  $f$  satisfies SGC with constant  $\rho$  and an additive error  $\sigma$ , then SGD with Nesterov acceleration with the following choice of parameters,*

$$\begin{aligned} \gamma_k &= \frac{1}{\sqrt{\mu\eta\rho}} ; \quad \beta_k = 1 - \sqrt{\frac{\mu\eta}{\rho}} \\ b_{k+1} &= \frac{\sqrt{\mu}}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \\ a_{k+1} &= \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \\ \alpha_k &= \frac{\gamma_k \beta_k b_{k+1}^2 \eta}{\gamma_k \beta_k b_{k+1}^2 \eta + a_k^2}; \quad \eta = \frac{1}{\rho L} \end{aligned}$$

results in the following convergence rate:

$$[\mathbb{E}[f(w_{k+1})] - f(w^*)] \leq \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k \left[ f(x_0) - f(w^*) + \frac{\mu}{2} \|x_0 - w^*\|^2 \right] + \frac{\sigma^2 \sqrt{\eta}}{\sqrt{\rho\mu}}$$

**Theorem 8** (Convex). *Under  $L$ -smoothness and convexity, if  $f$  satisfies SGC with constant  $\rho$  and an additive error  $\sigma$ , then SGD with Nesterov acceleration with the following choice of parameters,*

$$\begin{aligned} \gamma_k &= \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{k-1}^2}}{2} \\ a_{k+1} &= \gamma_k \sqrt{\eta\rho} \\ \alpha_k &= \frac{\gamma_k \eta}{\gamma_k \eta + a_k^2}; \quad \eta = \frac{1}{\rho L} \end{aligned}$$

results in the following convergence rate:

$$[\mathbb{E}f(w_{k+1}) - f(w^*)] \leq \frac{2\rho}{k^2\eta} \|x_0 - w^*\|^2 + \frac{k\sigma^2\eta}{\rho}$$

The above theorems are proved in appendices B.1.1 and B.1.3

## B Proofs

### B.1 Proofs for SGD with Nesterov Acceleration

Recall the update equations for SGD with Nesterov acceleration as follows:

$$\begin{aligned} w_{k+1} &= \zeta_k - \eta \nabla f(\zeta_k, z_k) \\ \zeta_k &= \alpha_k v_k + (1 - \alpha_k) w_k \\ v_{k+1} &= \beta_k v_k + (1 - \beta_k) \zeta_k - \gamma_k \eta \nabla f(\zeta_k, z_k) \end{aligned}$$

Since the stochastic gradients are unbiased, we obtain the following equation,

$$\mathbb{E}_z [\nabla f(y, z)] = \nabla f(y) \quad (9)$$

For the proof, we consider the more general strong-growth condition with an additive error  $\sigma^2$ .

$$\mathbb{E}_z \|\nabla f(w, z)\|^2 \leq \rho \|\nabla f(w)\|^2 + \sigma^2 \quad (10)$$

We choose the parameters  $\gamma_k, \alpha_k, \beta_k, a_k, b_k$  such that the following equations are satisfied:

$$\gamma_k = \frac{1}{\rho} \cdot \left[ 1 + \frac{\beta_k(1 - \alpha_k)}{\alpha_k} \right] \quad (11)$$

$$\alpha_k = \frac{\gamma_k \beta_k b_{k+1}^2 \eta}{\gamma_k \beta_k b_{k+1}^2 \eta + a_k^2} \quad (12)$$

$$\beta_k \geq 1 - \gamma_k \mu \eta \quad (13)$$

$$a_{k+1} = \gamma_k \sqrt{\eta \rho} b_{k+1} \quad (14)$$

$$b_{k+1} \leq \frac{b_k}{\sqrt{\beta_k}} \quad (15)$$

We now prove the following lemma assuming that the function  $f(\cdot)$  is  $L$ -smooth and  $\mu$  strongly-convex.

**Lemma 3.** *Assume that the function is  $L$ -smooth and  $\mu$  strongly-convex and satisfies the strong-growth condition in Equation 10. Then, using the updates in Equation 3-5 and setting the parameters according to Equations 11-15, if  $\eta \leq \frac{1}{\rho L}$ , then the following relation holds:*

$$b_{k+1}^2 \gamma_k^2 [\mathbb{E}f(w_{k+1}) - f^*] \leq \frac{a_0^2}{\rho \eta} [f(x_0) - f^*] + \frac{b_0^2}{2\rho \eta} \|x_0 - w^*\|^2 + \frac{\sigma^2 \eta}{\rho} \sum_{i=0}^k [\gamma_i^2 b_{i+1}^2]$$

*Proof.*

Let  $r_{k+1} = \|v_{k+1} - w^*\|$ , then using equation 5

$$\begin{aligned} r_{k+1}^2 &= \|\beta_k v_k + (1 - \beta_k) \zeta_k - w^* - \gamma_k \eta \nabla f(\zeta_k, z_k)\|^2 \\ r_{k+1}^2 &= \|\beta_k v_k + (1 - \beta_k) \zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \|\nabla f(\zeta_k, z_k)\|^2 + 2\gamma_k \eta \langle w^* - \beta_k v_k - (1 - \beta_k) \zeta_k, \nabla f(\zeta_k, z_k) \rangle \end{aligned}$$

Taking expectation wrt to  $z_k$ ,

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &= \mathbb{E}[\|\beta_k v_k + (1 - \beta_k) \zeta_k - w^*\|^2] + \gamma_k^2 \eta^2 \mathbb{E} \|\nabla f(\zeta_k, z_k)\|^2 + 2\gamma_k \eta [\mathbb{E} \langle w^* - \beta_k v_k - (1 - \beta_k) \zeta_k, \nabla f(\zeta_k, z_k) \rangle] \\ &\leq \|\beta_k v_k + (1 - \beta_k) \zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta [\langle w^* - \beta_k v_k - (1 - \beta_k) \zeta_k, \nabla f(\zeta_k) \rangle] + \gamma_k^2 \eta^2 \sigma^2 \\ &= \|\beta_k(v_k - w^*) + (1 - \beta_k)(\zeta_k - w^*)\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta [\langle w^* - \beta_k v_k - (1 - \beta_k) \zeta_k, \nabla f(\zeta_k) \rangle] + \gamma_k^2 \eta^2 \sigma^2 \\ &\leq \beta_k \|v_k - w^*\|^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta [\langle w^* - \beta_k v_k - (1 - \beta_k) \zeta_k, \nabla f(\zeta_k) \rangle] + \gamma_k^2 \eta^2 \sigma^2 \quad (\text{By convexity of } \|\cdot\|^2) \\ &= \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta [\langle w^* - \beta_k v_k - (1 - \beta_k) \zeta_k, \nabla f(\zeta_k) \rangle] + \gamma_k^2 \eta^2 \sigma^2 \\ &= \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta [\langle \beta_k(\zeta_k - v_k) + w^* - \zeta_k, \nabla f(\zeta_k) \rangle] + \gamma_k^2 \eta^2 \sigma^2 \\ &= \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta \left[ \langle \frac{\beta_k(1 - \alpha_k)}{\alpha_k} (w_k - \zeta_k) + w^* - \zeta_k, \nabla f(\zeta_k) \rangle \right] + \gamma_k^2 \eta^2 \sigma^2 \quad (\text{From equation 4}) \\ &= \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta \left[ \frac{\beta_k(1 - \alpha_k)}{\alpha_k} \langle \nabla f(\zeta_k), (w_k - \zeta_k) \rangle + \langle \nabla f(\zeta_k), w^* - \zeta_k \rangle \right] + \gamma_k^2 \eta^2 \sigma^2 \\ &\leq \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 + 2\gamma_k \eta \left[ \frac{\beta_k(1 - \alpha_k)}{\alpha_k} (f(w_k) - f(\zeta_k)) + \langle \nabla f(\zeta_k), w^* - \zeta_k \rangle \right] + \gamma_k^2 \eta^2 \sigma^2 \quad (\text{By convexity}) \end{aligned}$$

By strong convexity,

$$\begin{aligned}\mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + \gamma_k^2 \eta^2 \rho \|\nabla f(\zeta_k)\|^2 \\ &+ 2\gamma_k \eta \left[ \frac{\beta_k(1 - \alpha_k)}{\alpha_k} (f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2} \|\zeta_k - w^*\|^2 \right] + \gamma_k^2 \eta^2 \sigma^2\end{aligned}\quad (16)$$

By Lipschitz continuity of the gradient,

$$\begin{aligned}f(w_{k+1}) - f(\zeta_k) &\leq \langle \nabla f(\zeta_k), w_{k+1} - \zeta_k \rangle + \frac{L}{2} \|w_{k+1} - \zeta_k\|^2 \\ &\leq -\eta \langle \nabla f(\zeta_k), \nabla f(\zeta_k, z_k) \rangle + \frac{L\eta^2}{2} \|\nabla f(\zeta_k, z_k)\|^2\end{aligned}$$

Taking expectation wrt  $z_k$  and using equations 9, 10

$$\begin{aligned}\mathbb{E}[f(w_{k+1}) - f(\zeta_k)] &\leq -\eta \|\nabla f(\zeta_k)\|^2 + \frac{L\rho\eta^2}{2} \|\nabla f(\zeta_k)\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \mathbb{E}[f(w_{k+1}) - f(\zeta_k)] &\leq \left[ -\eta + \frac{L\rho\eta^2}{2} \right] \|\nabla f(\zeta_k)\|^2 + \frac{L\eta^2\sigma^2}{2}\end{aligned}$$

If  $\eta \leq \frac{1}{\rho L}$ ,

$$\begin{aligned}\mathbb{E}[f(w_{k+1}) - f(\zeta_k)] &\leq \left( \frac{-\eta}{2} \right) \|\nabla f(\zeta_k)\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \implies \|\nabla f(\zeta_k)\|^2 &\leq \left( \frac{2}{\eta} \right) \mathbb{E}[f(\zeta_k) - f(w_{k+1})] + L\eta\sigma^2\end{aligned}\quad (17)$$

From equations 16 and 17,

$$\begin{aligned}\mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + 2\gamma_k^2 \rho \eta \mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\ &+ 2\gamma_k \eta \left[ \frac{\beta_k(1 - \alpha_k)}{\alpha_k} (f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2} \|\zeta_k - w^*\|^2 \right] + \gamma_k^2 \eta^2 \sigma^2 + L\gamma_k^2 \eta^3 \rho \sigma^2 \\ &\leq \beta_k r_k^2 + (1 - \beta_k) \|\zeta_k - w^*\|^2 + 2\gamma_k^2 \eta \rho \mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\ &+ 2\gamma_k \eta \left[ \frac{\beta_k(1 - \alpha_k)}{\alpha_k} (f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2} \|\zeta_k - w^*\|^2 \right] + 2\gamma_k^2 \eta^2 \sigma^2 \quad (\text{Since } \eta \leq \frac{1}{\rho L}) \\ &= \beta_k r_k^2 + \|\zeta_k - w^*\|^2 [(1 - \beta_k) - \gamma_k \mu \eta] + f(\zeta_k) \left[ 2\gamma_k^2 \eta \rho - 2\gamma_k \eta \cdot \frac{\beta_k(1 - \alpha_k)}{\alpha_k} - 2\gamma_k \eta \right] \\ &- 2\gamma_k^2 \eta \rho \mathbb{E}f(w_{k+1}) + 2\gamma_k \eta f^* + \left[ 2\gamma_k \eta \cdot \frac{\beta_k(1 - \alpha_k)}{\alpha_k} \right] f(w_k) + 2\gamma_k^2 \eta^2 \sigma^2\end{aligned}$$

Since  $\beta_k \geq 1 - \gamma_k \mu \eta$  and  $\gamma_k = \frac{1}{\rho} \cdot \left( 1 + \frac{\beta_k(1 - \alpha_k)}{\alpha_k} \right)$ ,

$$\mathbb{E}[r_{k+1}^2] \leq \beta_k r_k^2 - 2\gamma_k^2 \eta \rho \mathbb{E}f(w_{k+1}) + 2\gamma_k \eta f^* + \left[ 2\gamma_k \eta \cdot \frac{\beta_k(1 - \alpha_k)}{\alpha_k} \right] f(w_k) + 2\gamma_k^2 \eta^2 \sigma^2$$

Multiplying by  $b_{k+1}^2$ ,

$$b_{k+1}^2 \mathbb{E}[r_{k+1}^2] \leq b_{k+1}^2 \beta_k r_k^2 - 2b_{k+1}^2 \gamma_k^2 \eta \rho \mathbb{E}f(w_{k+1}) + 2b_{k+1}^2 \gamma_k \eta f^* + \left[ 2b_{k+1}^2 \gamma_k \eta \cdot \frac{\beta_k(1 - \alpha_k)}{\alpha_k} \right] f(w_k) + 2b_{k+1}^2 \gamma_k^2 \eta^2 \sigma^2$$

Since  $b_{k+1}^2 \beta_k \leq b_k^2$ ,  $b_{k+1}^2 \gamma_k^2 \eta \rho = a_{k+1}^2$ ,  $\frac{\gamma_k \eta \beta_k (1-\alpha_k)}{\alpha_k} = \frac{a_k^2}{b_{k+1}^2}$

$$\begin{aligned} b_{k+1}^2 \mathbb{E}[r_{k+1}^2] &\leq b_k^2 r_k^2 - 2a_{k+1}^2 \mathbb{E}f(w_{k+1}) + 2b_{k+1}^2 \gamma_k \eta f^* + 2a_k^2 f(w_k) + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho} \\ &= b_k^2 r_k^2 - 2a_{k+1}^2 [\mathbb{E}f(w_{k+1}) - f^*] + 2a_k^2 [f(w_k) - f^*] + 2 [b_{k+1}^2 \gamma_k \eta - a_{k+1}^2 + a_k^2] f^* + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho} \end{aligned}$$

Since  $[b_{k+1}^2 \gamma_k \eta - a_{k+1}^2 + a_k^2] = 0$ ,

$$b_{k+1}^2 \mathbb{E}[r_{k+1}^2] \leq b_k^2 r_k^2 - 2a_{k+1}^2 [\mathbb{E}f(w_{k+1}) - f^*] + 2a_k^2 [f(w_k) - f^*] + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho}$$

Denoting  $\mathbb{E}f(w_{k+1})$  as  $\phi_k$ ,

$$2a_{k+1}^2 [\phi_{k+1} - f^*] + b_{k+1}^2 \mathbb{E}[r_{k+1}^2] \leq 2a_k^2 [\phi_k - f^*] + b_k^2 \mathbb{E}[r_k^2] + \frac{2a_{k+1}^2 \sigma^2 \eta}{\rho}$$

By recursion,

$$\begin{aligned} 2a_{k+1}^2 [\phi_{k+1} - f^*] + b_{k+1}^2 \mathbb{E}[r_{k+1}^2] &\leq 2a_0^2 [f(x_0) - f^*] + b_0^2 \|x_0 - w^*\|^2 + \frac{2\sigma^2 \eta}{\rho} \sum_{i=0}^k [a_{i+1}^2] \\ 2a_{k+1}^2 [\phi_{k+1} - f^*] &\leq 2a_0^2 [f(x_0) - f^*] + b_0^2 \|x_0 - w^*\|^2 + \frac{2\sigma^2 \eta}{\rho} \sum_{i=0}^k [a_{i+1}^2] \\ 2b_{k+1}^2 \gamma_k^2 \rho \eta [\phi_{k+1} - f^*] &\leq 2a_0^2 [f(x_0) - f^*] + b_0^2 \|x_0 - w^*\|^2 + 2\sigma^2 \eta^2 \rho \sum_{i=0}^k [\gamma_i^2 b_{i+1}^2] \\ b_{k+1}^2 \gamma_k^2 [\mathbb{E}f(w_{k+1}) - f^*] &\leq \frac{a_0^2}{\rho \eta} [f(x_0) - f^*] + \frac{b_0^2}{2\rho \eta} \|x_0 - w^*\|^2 + \frac{\sigma^2 \eta}{\rho} \sum_{i=0}^k [\gamma_i^2 b_{i+1}^2] \end{aligned}$$

□

**Lemma 4.** Under the parameter setting according to Equations 11- 15, the following relation is true:

$$\gamma_k^2 - \gamma_k \left[ \frac{1}{\rho} - \mu \eta \gamma_{k-1}^2 \right] = \gamma_{k-1}^2$$

*Proof.*

$$\begin{aligned} \gamma_k &= \frac{1}{\rho} \left[ 1 + \frac{\beta_k (1-\alpha_k)}{\alpha_k} \right] && \text{(From equation 11)} \\ \gamma_k^2 - \frac{\gamma_k}{\rho} &= \frac{\gamma_k \beta_k (1-\alpha_k)}{\rho \alpha_k} \\ &= \frac{1}{\eta \rho} \frac{a_k^2}{b_{k+1}^2} && \text{(From equation 12)} \\ &= \frac{\beta_k}{\eta \rho} \frac{a_k^2}{b_k^2} && \text{(From equation 15)} \\ &= \frac{1 - \gamma_k \mu \eta}{\eta \rho} \frac{a_k^2}{b_k^2} && \text{(From equation 13)} \\ &= \frac{1 - \gamma_k \mu \eta}{\eta \rho} (\gamma_{k-1} \sqrt{\eta \rho})^2 && \text{(From equation 13)} \\ &= (1 - \gamma_k \mu \eta) \gamma_{k-1}^2 \\ \implies \gamma_k^2 - \gamma_k \left[ \frac{1}{\rho} - \mu \eta \gamma_{k-1}^2 \right] &= \gamma_{k-1}^2 && \text{(18)} \end{aligned}$$

□

### B.1.1 Strongly-convex case

We now consider the strongly-convex case,

Using Lemma 4,

$$\gamma_k^2 - \gamma_k \left[ \frac{1}{\rho} - \mu\eta\gamma_{k-1}^2 \right] = \gamma_{k-1}^2$$

If  $\gamma_k = C$ , then

$$\begin{aligned} \gamma_k &= \frac{1}{\sqrt{\mu\eta\rho}} \\ \beta_k &= 1 - \sqrt{\frac{\mu\eta}{\rho}} \\ b_{k+1} &= \frac{b_0}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \\ a_{k+1} &= \frac{1}{\sqrt{\mu\eta\rho}} \cdot \sqrt{\eta\rho} \cdot \frac{b_0}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} = \frac{b_0}{\sqrt{\mu}} \cdot \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}} \end{aligned}$$

If  $b_0 = \sqrt{\mu}$ ,

$$a_{k+1} = \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)/2}}$$

The above equation implies that  $a_0 = 1$ . This gives us the parameter settings used in Theorem 1.

Using the result of Lemma 3 and the above relations, we obtain the following inequality. Note that  $\phi_{k+1} = \mathbb{E}[f(w_{k+1})]$ .

$$\begin{aligned} \frac{\mu}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(k+1)}} \cdot \frac{1}{\mu\eta\rho} [\phi_{k+1} - f^*] &\leq \frac{1}{\rho\eta} [f(x_0) - f^*] + \frac{\mu}{2\rho\eta} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{\rho} \cdot \frac{1}{\mu\eta\rho} \sum_{i=0}^k \frac{\mu}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(i+1)}} \\ \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k} [\phi_{k+1} - f^*] &\leq [f(x_0) - f^*] + \frac{\mu}{2} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{\rho} \sum_{i=0}^k \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{(i+1)}} \\ \frac{1}{\left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k} [\phi_{k+1} - f^*] &\leq \left[ f(x_0) - f^* + \frac{\mu}{2} \|x_0 - w^*\|^2 \right] + \frac{\sigma^2\sqrt{\eta}}{\sqrt{\rho\mu}} \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^{-k} \\ [\phi_{k+1} - f^*] &\leq \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k \left[ f(x_0) - f^* + \frac{\mu}{2} \|x_0 - w^*\|^2 \right] + \frac{\sigma^2\sqrt{\eta}}{\sqrt{\rho\mu}} \end{aligned}$$

### B.1.2 Proof of Theorem 1

We use the above relation to complete the proof for Theorem 1. Substituting  $\eta = \frac{1}{\rho L}$  and  $\sigma = 0$ , we obtain the following:

$$[\mathbb{E}[f(w_{k+1})] - f^*] \leq \left(1 - \sqrt{\frac{\mu\eta}{\rho}}\right)^k \left[f(x_0) - f^* + \frac{\mu}{2} \|x_0 - w^*\|^2\right]$$

### B.1.3 Convex case

We now use the above lemmas to first prove the convergence rate in the convex case. In this case,  $\mu = 0$  and the result of Lemma 4 can be written as:

$$\begin{aligned} \gamma_k^2 - \frac{\gamma_k}{\rho} - \gamma_{k-1}^2 &= 0 \\ \implies \gamma_k &= \frac{\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} + 4\gamma_{k-1}^2}}{2} \end{aligned}$$

Let  $\gamma_0 = 0$ . From equation 13, for all  $k$ ,

$$\begin{aligned} \beta_k &= 1 \\ b_{k+1} = b_k = b_0 &= 1 && \text{(From equation 15)} \\ a_{k+1} = \gamma_k \sqrt{\eta\rho} b_0 &\implies a_{k+1} = \gamma_k \sqrt{\eta\rho} && \text{(From equation 14)} \end{aligned}$$

The above equation implies that  $a_0 = 0$ . This gives us the parameter settings used in Theorem 2.

Using the result of Lemma 3 by setting  $\mu = 0$  and the above relations, we obtain the following inequality. Note that  $\phi_{k+1} = \mathbb{E}[f(w_{k+1})]$ .

$$\gamma_k^2 [\phi_{k+1} - f^*] \leq \frac{1}{2\rho\eta} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{\rho} \sum_{i=1}^{k-1} [\gamma_i^2]$$

By induction,  $\gamma_i \geq \frac{i}{2\rho}$ ,

$$\begin{aligned} \frac{k^2}{4\rho^2} [\phi_{k+1} - f^*] &\leq \frac{1}{2\rho\eta} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{4\rho^3} \sum_{i=1}^{k-1} [i^2] \\ [\phi_{k+1} - f^*] &\leq \frac{2\rho}{k^2\eta} \|x_0 - w^*\|^2 + \frac{\sigma^2\eta}{k^2\rho} \sum_{i=1}^{k-1} [i^2] \\ [\phi_{k+1} - f^*] &\leq \frac{2\rho}{k^2\eta} \|x_0 - w^*\|^2 + \frac{k\sigma^2\eta}{\rho} \end{aligned}$$

### B.1.4 Proof of Theorem 2

We use the above relation to complete the proof for Theorem 2. Substituting  $\eta = \frac{1}{\rho L}$  and  $\sigma = 0$ , we obtain the following:

$$[\mathbb{E}[f(w_{k+1})] - f^*] \leq \frac{2\rho^2 L}{k^2} \|x_0 - w^*\|^2$$

## B.2 Proof of Theorem 3

*Proof.* Recall the stochastic gradient descent update,

$$w_{k+1} = w_k - \eta \nabla f(w_k, z_k) \quad (19)$$

By Lipschitz continuity of the gradient,

$$\begin{aligned} f(w_{k+1}) - f(w_k) &\leq \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \|w_{k+1} - w_k\|^2 \\ &\leq -\eta \langle \nabla f(w_k), \nabla f(w_k, z_k) \rangle + \frac{L\eta^2}{2} \|\nabla f(w_k, z_k)\|^2 \end{aligned}$$

Taking expectation wrt  $z_k$  and using equations 9, 10

$$\begin{aligned} \mathbb{E}[f(w_{k+1}) - f(w_k)] &\leq -\eta \|\nabla f(w_k)\|^2 + \frac{L\rho\eta^2}{2} \|\nabla f(w_k)\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \mathbb{E}[f(w_{k+1}) - f(w_k)] &\leq \left[ -\eta + \frac{L\rho\eta^2}{2} \right] \|\nabla f(w_k)\|^2 + \frac{L\eta^2\sigma^2}{2} \end{aligned}$$

If  $\eta \leq \frac{1}{\rho L}$ ,

$$\begin{aligned} \mathbb{E}[f(w_{k+1}) - f(w_k)] &\leq \left( \frac{-\eta}{2} \right) \|\nabla f(w_k)\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \implies \|\nabla f(w_k)\|^2 &\leq \left( \frac{2}{\eta} \right) \mathbb{E}[f(w_k) - f(w_{k+1})] + L\eta\sigma^2 \end{aligned} \quad (20)$$

Taking expectation wrt  $z_0, z_1, \dots, z_{t-1}$  and summing from  $k = 0$  to  $t-1$ ,

$$\begin{aligned} \sum_{k=0}^{t-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \left( \frac{2}{\eta} \right) \sum_{k=0}^{t-1} \mathbb{E} [f(w_k) - f(w_{k+1})] + L\eta t\sigma^2 \\ \implies \sum_{k=0}^{t-1} \min_{k=0,1,\dots,t-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \left( \frac{2}{\eta} \right) \sum_{k=0}^{t-1} \mathbb{E} [f(w_k) - f(w_{k+1})] + L\eta\sigma^2 \\ \min_{k=0,1,\dots,t-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \left( \frac{2}{\eta t} \right) [f(w_0) - \mathbb{E}[f(w_t)]] + L\eta\sigma^2 \\ \min_{k=0,1,\dots,t-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \left( \frac{2}{\eta t} \right) [f(w_0) - f(w^*)] + L\eta\sigma^2 \end{aligned}$$

If  $\sigma = 0$ ,

$$\begin{aligned} \min_{k=0,1,\dots,t-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \left( \frac{2}{\eta t} \right) [f(w_0) - f(w^*)] \\ \implies \min_{k=0,1,\dots,t-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \left( \frac{2\rho L}{t} \right) [f(w_0) - f(w^*)] \quad (\text{Setting } \eta = \frac{1}{\rho L}) \end{aligned}$$

□

## B.3 Proof of Theorem 4

*Proof.* Similar to the proof of Theorem 3, we can use the SGD update and Lipschitz continuity of the gradient to obtain the following equation for the stepsize  $\eta \leq \frac{1}{\rho L}$ :

$$\mathbb{E}[f(w_{k+1}) - f(w_k)] \leq \left( \frac{-\eta}{2} \right) \|\nabla f(w_k)\|^2 + \frac{L\eta^2\sigma^2}{2}$$

We now use the PL inequality with constant  $\mu$  as follows:

$$\|\nabla f(w_k)\|^2 \geq 2\mu [f(w_k) - f^*]$$

Combining the above two inequalities,

$$\mathbb{E}[f(w_{k+1}) - f(w_k)] \leq -\eta\mu [f(w_k) - f^*] + \frac{L\eta^2\sigma^2}{2}$$

If  $\sigma = 0$ ,

$$\begin{aligned} \mathbb{E}[f(w_{k+1}) - f(w_k)] &\leq -\eta\mu [f(w_k) - f^*] \\ \implies \mathbb{E}[f(w_{k+1}) - f^*] &\leq (1 - \eta\mu) [f(w_k) - f^*] \end{aligned}$$

Substituting  $\eta = \frac{1}{\rho L}$ ,

$$\begin{aligned} \mathbb{E}[f(w_{k+1}) - f^*] &\leq \left(1 - \frac{\mu}{\rho L}\right) [f(w_k) - f^*] \\ \implies \mathbb{E}[f(w_{k+1}) - f^*] &\leq \left(1 - \frac{\mu}{\rho L}\right)^k [f(w_0) - f^*] \end{aligned} \tag{21}$$

□

#### B.4 Proof of Theorem 5

*Proof.*

$$\begin{aligned} \|w_{k+1} - w^*\|^2 &= \|w_k - \eta\nabla f(w_k, z) - w^*\|^2 \\ &= \|w_k - w^*\|^2 - 2\eta\langle \nabla f(w_k, z), w_k - w^* \rangle + \eta^2 \|\nabla f(w_k, z)\|^2 \\ \mathbb{E}_z[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\eta\mathbb{E}[\langle \nabla f(w_k, z), w_k - w^* \rangle] + \eta^2\mathbb{E}[\|\nabla f(w_k, z)\|^2] \\ &= \|w_k - w^*\|^2 - 2\eta\langle \nabla f(w_k), w_k - w^* \rangle + \eta^2\mathbb{E}[\|\nabla f(w_k, z)\|^2] \\ &\quad (\text{From the unbiasedness of stochastic gradients.}) \\ &\leq \|w_k - w^*\|^2 - 2\eta\langle \nabla f(w_k), w_k - w^* \rangle + 2\rho\eta^2L[f(w_k) - f^*] \\ &\leq \|w_k - w^*\|^2 + 2\eta \left[ f^* - f(w_k) - \frac{\mu}{2} \|w_k - w^*\|^2 \right] + 2\rho\eta^2L[f(w_k) - f^*] \\ &\quad (\text{From equation 6}) \\ &= (1 - \mu\eta) \|w_k - w^*\|^2 + (2\eta^2\rho L - 2\eta) [f(w_k) - f^*] \\ \|w_{k+1} - w^*\|^2 &\leq \left(1 - \frac{\mu}{\rho L}\right) \|w_k - w^*\|^2 \\ \implies \|w_{k+1} - w^*\|^2 &\leq \left(1 - \frac{\mu}{\rho L}\right)^k \|x_0 - w^*\|^2 \end{aligned} \tag{Setting } \eta = \frac{1}{\rho L}$$

□

#### B.5 Proof of Theorem 6

*Proof.*

By convexity,

$$f(w_k) \leq f(w^*) + \langle \nabla f(w_k), w_k - w^* \rangle$$

For any  $\beta \leq 1$ ,

$$f(w_k) \leq \beta f(w_k) + (1 - \beta)f(w^*) + (1 - \beta)\langle \nabla f(w_k), w_k - w^* \rangle$$

By Lipschitz continuity of  $\nabla f(f)$ ,

$$\begin{aligned} f(w_{k+1}) &\leq f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \|w_{k+1} - w_k\|^2 \\ \implies f(w_{k+1}) &\leq f(w_k) - \eta \langle \nabla f(w_k), \nabla f(w_k, z) \rangle + \frac{\eta^2 L}{2} \|\nabla f(w_k, z)\|^2 \end{aligned}$$

From the above equations,

$$f(w_{k+1}) \leq \beta f(w_k) + (1 - \beta)f(w^*) + (1 - \beta)\langle \nabla f(w_k), w_k - w^* \rangle - \eta \langle \nabla f(w_k), \nabla f(w_k, z) \rangle + \frac{\eta^2 L}{2} \|\nabla f(w_k, z)\|^2$$

Note that,

$$\begin{aligned} \frac{1}{2\eta} (\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2) &= \frac{1}{2\eta} (\|w_k - w^*\|^2 - \|w_k - \eta \nabla f(w_k, z) - w^*\|^2) \\ &= \frac{1}{2\eta} (\|w_k - w^*\|^2 - \|w_k - w^*\|^2 - \eta^2 \|\nabla f(w_k, z)\|^2 + 2\eta \langle w_k - w^*, \nabla f(w_k, z) \rangle) \\ \frac{1}{2\eta} (\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2) &= \frac{-\eta}{2} \|\nabla f(w_k, z)\|^2 + \langle w_k - w^*, \nabla f(w_k, z) \rangle \\ \implies \langle w_k - w^*, \nabla f(w_k, z) \rangle &= \frac{1}{2\eta} (\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2) + \frac{\eta}{2} \|\nabla f(w_k, z)\|^2 \end{aligned}$$

Taking expectation

$$\begin{aligned} \mathbb{E} [\langle w_k - w^*, \nabla f(w_k, z) \rangle] &= \frac{1}{2\eta} (\|w_k - w^*\|^2 - \mathbb{E} [\|w_{k+1} - w^*\|^2]) + \frac{\eta}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \\ \implies \langle w_k - w^*, \nabla f(w_k) \rangle &= \frac{1}{2\eta} (\|w_k - w^*\|^2 - \mathbb{E} [\|w_{k+1} - w^*\|^2]) + \frac{\eta}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \end{aligned}$$

Using the above equations,

$$\begin{aligned} f(w_{k+1}) &\leq \beta f(w_k) + (1 - \beta)f(w^*) + \frac{1 - \beta}{2\eta} (\|w_k - w^*\|^2 - \mathbb{E} [\|w_{k+1} - w^*\|^2]) + \frac{(1 - \beta)(\eta)}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \\ &\quad - \eta \langle \nabla f(w_k), \nabla f(w_k, z) \rangle + \frac{\eta^2 L}{2} \|\nabla f(w_k, z)\|^2 \end{aligned}$$

Taking expectation,

$$\begin{aligned} \mathbb{E}[f(w_{k+1})] &\leq \beta f(w_k) + (1 - \beta)f(w^*) + \frac{1 - \beta}{2\eta} (\|w_k - w^*\|^2 - \mathbb{E} [\|w_{k+1} - w^*\|^2]) + \frac{(1 - \beta)(\eta)}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \\ &\quad - \eta \langle \nabla f(w_k), \mathbb{E} [\nabla f(w_k, z)] \rangle + \frac{\eta^2 L}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \\ &= \beta f(w_k) + (1 - \beta)f(w^*) + \frac{1 - \beta}{2\eta} (\|w_k - w^*\|^2 - \mathbb{E} [\|w_{k+1} - w^*\|^2]) + \frac{(1 - \beta)(\eta)}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \\ &\quad - \eta \|\nabla f(w_k)\|^2 + \frac{\eta^2 L}{2} \mathbb{E} [\|\nabla f(w_k, z)\|^2] \end{aligned}$$

The term  $-\eta \|\nabla f(w_k)\|^2 \leq 0$

$$\begin{aligned} \implies \mathbb{E}[f(w_{k+1})] &\leq \beta f(w_k) + (1 - \beta)f(w^*) + \frac{1 - \beta}{2\eta} \left( \|w_k - w^*\|^2 - \mathbb{E}[\|w_{k+1} - w^*\|^2] \right) \\ &\quad + \frac{(1 - \beta)(\eta)}{2} \mathbb{E}[\|\nabla f(w_k, z)\|^2] + \frac{\eta^2 L}{2} \mathbb{E}[\|\nabla f(w_k, z)\|^2] \\ \mathbb{E}[f(w_{k+1})] - f(w^*) &\leq \beta(f(w_k) - f(w^*)) + \frac{1 - \beta}{2\eta} \left( \|w_k - w^*\|^2 - \mathbb{E}[\|w_{k+1} - w^*\|^2] \right) \\ &\quad + \left( \frac{(1 - \beta)(\eta)}{2} + \frac{\eta^2 L}{2} \right) \mathbb{E}[\|\nabla f(w_k, z)\|^2] \end{aligned}$$

From equation 6,

$$\begin{aligned} \mathbb{E}[f(w_{k+1})] - f(w^*) &\leq \beta(f(w_k) - f(w^*)) + \frac{1 - \beta}{2\eta} \left( \|w_k - w^*\|^2 - \mathbb{E}[\|w_{k+1} - w^*\|^2] \right) \\ &\quad + (\rho(1 - \beta)\eta L + \eta^2 \rho L^2)(f(w_k) - f(w^*)) \end{aligned}$$

Let us choose  $1 - \beta = \eta L$ ,

$$\begin{aligned} \mathbb{E}[f(w_{k+1})] - f(w^*) &\leq \beta(f(w_k) - f(w^*)) + \frac{1 - \beta}{2\eta} \left( \|w_k - w^*\|^2 - \mathbb{E}[\|w_{k+1} - w^*\|^2] \right) + 2\rho\eta^2 L^2(f(w_k) - f(w^*)) \\ \mathbb{E}[f(w_{k+1})] - f(w^*) &\leq (\beta + 2\rho\eta^2 L^2)(f(w_k) - f(w^*)) + \frac{L}{2} \left( \|w_k - w^*\|^2 - \mathbb{E}[\|w_{k+1} - w^*\|^2] \right) \end{aligned}$$

Let  $\delta_{k+1} = \mathbb{E}[f(w_{k+1})] - f(w^*)$  and  $\Delta_{k+1} = \mathbb{E}[\|w_{k+1} - w^*\|^2]$

$$\implies \delta_{k+1} \leq (\beta + 2\rho\eta^2 L^2) \delta_k + \frac{L}{2} [\Delta_k - \Delta_{k+1}]$$

Summing from  $i = 0$  to  $k - 1$ ,

$$\begin{aligned} \sum_{i=0}^{k-1} \delta_{i+1} &\leq (\beta + 2\rho\eta^2 L^2) \sum_{i=0}^{k-1} \delta_i + \frac{L}{2} \sum_{i=0}^{k-1} [\Delta_i - \Delta_{i+1}] \\ \implies \sum_{i=0}^{k-1} \delta_{i+1} &\leq (\beta + 2\rho\eta^2 L^2) \sum_{i=0}^{k-1} \delta_i + \frac{L}{2} \Delta_0 \\ \implies \sum_{i=1}^k \delta_i &\leq \frac{(\beta + 2\rho\eta^2 L^2) \delta_0 + \frac{L}{2} \Delta_0}{(1 - \beta - 2\rho\eta^2 L^2)} \end{aligned}$$

Let  $\bar{w}_k = \frac{[\sum_{i=1}^k w_i]}{k}$ . By Jensen's inequality,

$$\begin{aligned} \mathbb{E}[f(\bar{w}_k)] &\leq \frac{\sum_{i=1}^k \mathbb{E}[f(w_i)]}{k} \\ \implies \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \sum_{i=1}^k \delta_i \\ \implies \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{(\beta + 2\rho\eta^2 L^2) \delta_0 + \frac{L}{2} \Delta_0}{(1 - \beta - 2\rho\eta^2 L^2) k} \\ \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{(1 - \eta L + 2\rho\eta^2 L^2)[f(w_0) - f(w^*)] + \frac{L}{2} \|w_0 - w^*\|^2}{(\eta L - 2\rho\eta^2 L^2) k} \quad (\text{Since } 1 - \beta = \eta L) \end{aligned}$$

If  $\eta = \frac{1}{4\rho L}$ ,

$$\begin{aligned}\mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{\frac{7}{8\rho} [f(w_0) - f(w^*)] + \frac{L}{2} \|w_0 - w^*\|^2}{\frac{1}{8\rho} k} \\ \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{7 [f(w_0) - f(w^*)] + 4\rho L \|w_0 - w^*\|^2}{k} \\ \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{(7L/2) \|w_0 - w^*\|^2 + 4\rho L \|w_0 - w^*\|^2}{k} \\ \implies \mathbb{E}[f(\bar{w}_k)] - f(w^*) &\leq \frac{4(1+\rho) \|w_0 - w^*\|^2}{k}\end{aligned}$$

□

## B.6 Proof for Proposition 1

*Proof.*

For the first part, we use the PL inequality which states the for all  $w$ ,

$$2[f(w) - f(w^*)] \leq \frac{1}{\mu} \|\nabla f(w)\|^2$$

Combining this with the WGC gives us the desired result

For the converse, we use smoothness and the convexity of  $f(\cdot)$ . Specifically, for all points  $a, b$ ,

$$f(a) - f(b) \geq \langle f(b), a - b \rangle + \frac{1}{2L} \|\nabla f(a) - \nabla f(b)\|^2$$

Substituting  $a = w$  and  $b = w^*$  and rearranging,

$$\|\nabla f(w)\|^2 \leq 2L \cdot [f(w) - f(w^*)]$$

Combining this with the SGC gives us the desired result.

□

## B.7 Proof for Proposition 2

*Proof.*

$$\mathbb{E}_i \|\nabla f_i(w)\|^2 = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w)\|^2 \tag{22}$$

By Lipschitz continuity of  $\nabla f_i(w)$  and convexity,

$$f_i(w) - f_i(w^*) \geq \langle \nabla f_i(w^*), w - w^* \rangle + \frac{1}{2L_i} \|\nabla f_i(w) - \nabla f_i(w^*)\|^2$$

For all  $i$ ,  $\nabla f_i(w^*) = \nabla f(w^*) = 0$ . Hence,

$$\begin{aligned} f_i(w) - f_i(w^*) &\geq \frac{1}{2L_i} \|\nabla f_i(w)\|^2 \\ \implies \|\nabla f_i(w)\|^2 &\leq 2L_i [f_i(w) - f_i(w^*)] \end{aligned}$$

Using Equation 22,

$$\begin{aligned} \mathbb{E}_i \|\nabla f_i(w)\|^2 &\leq \sum_{i=1}^n \left[ \frac{2L_i}{n} [f_i(w) - f_i(w^*)] \right] \\ &\leq \frac{2L_{\max}}{n} \sum_{i=1}^n [f_i(w) - f_i(w^*)] \\ \mathbb{E}_i \|\nabla f_i(w)\|^2 &\leq 2L_{\max} [f(w) - f(w^*)] \end{aligned} \tag{23}$$

□

## B.8 Proof for Lemma 1

*Proof.* Let  $a = y \cdot x$ . For the squared-hinge loss, the strong growth condition is equivalent to

$$\begin{aligned} \mathbb{E}[(1 - w^\top a)_+^2] &\leq \rho \|\mathbb{E}[(1 - w^\top a)_+]\|^2 \\ \|\mathbb{E}[(1 - w^\top a)_+ a]\| &\geq \frac{1}{\|w_*\|} \mathbb{E}[(1 - w^\top a)_+ a^\top w_*] \\ &\geq \tau \mathbb{E}[(1 - w^\top a)_+] \end{aligned}$$

We thus need to upper bound  $\mathbb{E}[(1 - w^\top a)_+^2]$  by a constant  $c$  times  $(\mathbb{E}[(1 - w^\top a)_+])^2$ . We must have  $c \geq 1$  (as a consequence of Jensen's inequality). Then we have  $\rho = c/\tau^2$ . Next, we prove that if the distribution of  $a$  is uniform over  $\kappa$  values, then  $c = \kappa$ .

Consider a random variable  $A \in \mathbb{R}_+$  taking  $\kappa$  values  $a_1, \dots, a_\kappa$  with probabilities  $p_1, \dots, p_\kappa$ . Then  $(\mathbb{E}A)^2 = \sum_{i,j} p_i p_j a_i a_j \geq \sum_i a_i^2 p_i^2 \geq \min_i p_i \sum_i a_i^2 p_i$ , □

## B.9 Proof for Lemma 2

*Proof.* Let  $a = y \cdot x$ .

$$\begin{aligned} \mathbb{P}(a^\top w \leq 0) &\leq \mathbb{P}((1 - a^\top w)_+^2 \geq 1) \\ &\leq \mathbb{E}(1 - a^\top w)_+^2 \\ \implies \mathbb{P}(a^\top w \leq 0) &\leq \mathbb{E}f(w, a) \end{aligned}$$

□

## C Additional experimental results

In this section, we propose to use a line-search heuristic for both constant step-size SGD and its accelerated variant. For SGD, we use the line-search proposed in SAG [31]: start with an initial estimate  $\hat{L} = 1$  and in each iteration, we double the estimate when the condition  $f_k \left( w_k - \frac{1}{\hat{L}} \nabla f_k(w_k) \right) \leq f_k(w_k) - \frac{1}{2\hat{L}} \|\nabla f_k(w_k)\|^2$  is not satisfied. We denote this variant as SGD(LS) and the corresponding variant that uses a  $1/L$  step-size as SGD(T). For the accelerated case, we use the same line-search procedure as above, but search for an appropriate value of  $\rho L$ . We denote the accelerated variant with and without line-search as Acc-SGD(LS) and Acc-SGD(T) respectively.

We make the following observations: (i) Accelerated SGD in conjunction with our line-search heuristic is stable across datasets. (ii) Acc-SGD(LS) either matches or outperforms Acc-SGD(T). (iii) In some cases, SGD(LS) can result in faster empirical convergence as compared to the accelerated variants. We plan to investigate better line-search methods for both SGD [31] and Acc-SGD [21] in the future.

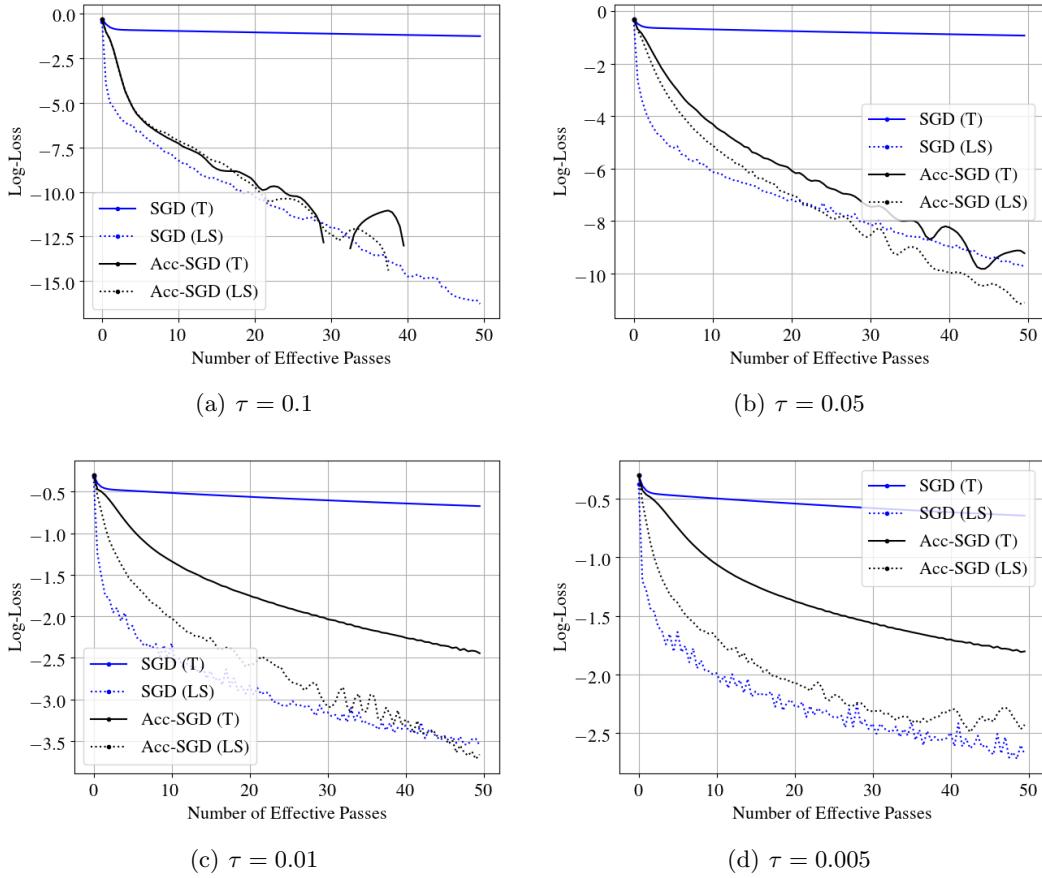


Figure 3: Comparison of SGD and variants of accelerated SGD on a synthetic linearly separable dataset with margin  $\tau$ .

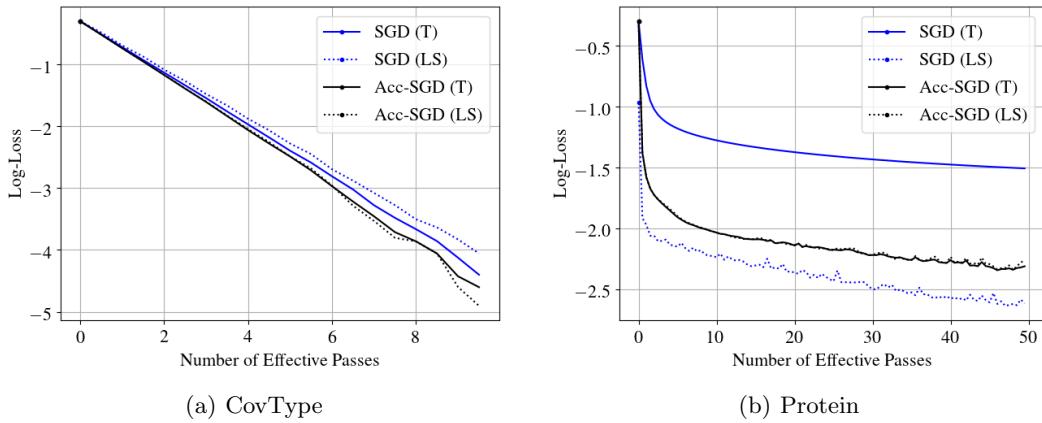


Figure 4: Comparison of SGD and accelerated SGD for learning a linear classifier with RBF features on the (a) CovType and (b) Protein datasets.