Towards Noise-adaptive, Problem-adaptive (Accelerated) SGD

Sharan Vaswani [1]   **Benjamin Dubois-Taine** [2]   **Reza Babanezhad** [3]

Abstract

We aim to make stochastic gradient descent (SGD) adaptive to (i) the noise σ^2 in the stochastic gradients and (ii) problem-dependent constants. When minimizing smooth, strongly-convex functions with condition number κ, we prove that T iterations of SGD with exponentially decreasing step-sizes and knowledge of the smoothness can achieve an $\tilde{O}\left(\exp\left(-T/\kappa\right) + \sigma^2/T\right)$ rate, without knowing σ^2. In order to be adaptive to the smoothness, we use a stochastic line-search (SLS) and show (via upper and lower-bounds) that SGD with SLS converges at the desired rate, but only to a neighbourhood of the solution. On the other hand, we prove that SGD with an offline estimate of the smoothness converges to the minimizer. However, its rate is slowed down proportional to the estimation error. Next, we prove that SGD with Nesterov acceleration and exponential step-sizes (referred to as ASGD) can achieve the near-optimal $\tilde{O}\left(\exp\left(-T/\sqrt{\kappa}\right) + \sigma^2/T\right)$ rate, without knowledge of σ^2. When used with offline estimates of the smoothness and strong-convexity, ASGD still converges to the solution, albeit at a slower rate. Finally, we empirically demonstrate the effectiveness of exponential step-sizes coupled with a novel variant of SLS.

1. Introduction

We study unconstrained minimization of a finite-sum objective $f : \mathbb{R}^d \to \mathbb{R}$ prevalent in machine learning,

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w). \tag{1}$$

[1]Simon Fraser University [2]DI ENS, Ecole normale supérieure, Université PSL, CNRS, INRIA, 75005 Paris, France [3]SAIT AI lab, Montreal. Correspondence to: Sharan Vaswani <vaswani.sharan@gmail.com>.

For supervised learning, n represents the number of training examples and f_i is the loss on example i. We assume f to be a smooth, strongly-convex function and denote w^* to be the unique minimizer of the above problem.

We study stochastic gradient descent (SGD) and its accelerated variant for minimizing f (Robbins and Monro, 1951; Nemirovski and Yudin, 1983; Nesterov, 2004; Bottou et al., 2018). The empirical performance and the theoretical convergence of SGD is governed by the choice of its step-size, and there are numerous ways of selecting it. For example, Moulines and Bach (2011); Gower et al. (2019) use a *constant* step-size for convex and strongly convex functions. A constant step-size only guarantees convergence to a neighborhood of the solution. In order to converge to the exact minimizer, a common technique is to decrease the step-size at an appropriate rate, and such decreasing step-sizes have also been well-studied (Robbins and Monro, 1951; Ghadimi and Lan, 2012). The rate at which the step-size needs to be decayed depends on the function class under consideration. For example, when minimizing smooth, strongly-convex functions using T iterations of SGD, the step-size is decayed at an $O(1/k)$ rate where k is the iteration number. This results in an $\Theta(1/T)$ convergence rate for SGD and is optimal in the stochastic setting (Nguyen et al., 2018).

On the other hand, when minimizing a smooth, strongly-convex function with condition number κ, deterministic (full-batch) gradient descent (GD) with a *constant* step-size converges linearly at an $O(\exp(-T/\kappa))$ rate. Augmenting constant step-size GD with Nesterov acceleration can further improve the convergence rate to $\Theta(\exp(-T/\sqrt{\kappa}))$ which is optimal in the deterministic setting (Nesterov, 2004). Hence, the stochastic and deterministic algorithms use different step-size strategies to obtain the optimal rates in their respective settings.

Noise-adaptive SGD: Ideally, we want to design step-size schemes that make SGD adaptive to the noise in the stochastic gradients, matching the optimal convergence rates in both the deterministic and stochastic settings. Furthermore, in order for the algorithm to be practical, it should not require knowledge of the stochasticity (e.g. a bound on σ^2, the variance in stochastic gradients). Recently, Khaled and Richtárik (2020); Li et al. (2020)

achieve the $\tilde{O}\left(\exp(-T/\kappa) + \sigma^2/T\right)$ for smooth functions satisfying the Polyak-Lojasiewicz (PL) condition (Karimi et al., 2016), a generalization of strong-convexity. More importantly, these works are *noise-adaptive* and do not require the knowledge of σ^2. For this, Li et al. (2020) use SGD with an exponentially decreasing sequence of step-sizes, while Khaled and Richtárik (2020) use a constant then decaying step-size. There are two limitations with these works: (i) they require the knowledge of problem-dependent constants such as the smoothness and strong-convexity of the underlying function, and (ii) they do not match the optimal $\sqrt{\kappa}$ dependence (of the Nesterov accelerated method) in the linear convergence term, and are hence sub-optimal in the deterministic setting. We will address both these limitations in this work.

Towards noise and problem-adaptive SGD: Typically, SGD requires the knowledge of problem-dependent constants to set the step-size. In practice, it is difficult to estimate these quantities, and one can only obtain loose bounds on them. Consequently, there have been numerous methods (Duchi et al., 2011; Li and Orabona, 2019; Kingma and Ba, 2015; Bengio, 2015; Vaswani et al., 2019b; Loizou et al., 2021) that can adapt to the problem, and adjust the step-size on the fly. We term such methods as *problem-adaptive*. Unfortunately, it is unclear if such problem-adaptive methods can also be made noise-adaptive. On the other hand, as mentioned above, none of the noise-adaptive methods (Li et al., 2020; Khaled and Richtárik, 2020; Stich, 2019) are problem-adaptive. Amongst these, the noise-adaptive algorithm in Li et al. (2020) only requires knowledge of the smoothness constant and we try to relax this requirement.

Contribution: In Section 3.2, we use stochastic line-search (SLS) (Vaswani et al., 2019b) to estimate the smoothness constant on the fly. We prove that SGD in conjunction with exponentially decreasing step-sizes and SLS converges at the desired noise-adaptive rate but only to a *neighbourhood of the solution*. This neighbourhood depends on the noise and the error in estimating the smoothness. We prove a corresponding lower-bound that shows the necessity of this neighbourhood term. Our lower-bound shows that if the SGD step-size is adaptively set in an online fashion (using the sampled function), no decreasing sequence of step-sizes can converge to the minimizer.

Contribution: In Section 3.3, we consider estimating the smoothness constant in an offline fashion (before running the algorithm). We prove that SGD with an offline estimate of the smoothness and exponentially decreasing step-sizes converges to the solution, though its rate is slowed down by a factor proportional to the estimation error in the smoothness. In particular, our upper-bound shows that misestimating the smoothness constant can slow down the conver-

gence rate. We complement this result with a lower-bound that shows that this slowdown is unavoidable.

Our results thus demonstrate the difficulty of obtaining noise-adaptive rates while being adaptive to problem-dependent parameters.

Noise-adaptive SGD with Nesterov acceleration: We now turn to the second limitation of existing noise-adaptive methods, and aim to use Nesterov acceleration in order to obtain the optimal $\tilde{O}\left(\exp(-T/\sqrt{\kappa}) + \sigma^2/T\right)$ rate, without the knowledge of σ^2. The work in Jain et al. (2018); Arjevani et al. (2020) satisfies the desired criteria for quadratic functions. For general smooth, strongly-convex functions, Ghadimi and Lan (2013); Kulunchakov and Mairal (2019) obtain the desired rate, but require the knowledge of σ^2, and are consequently not noise-adaptive. Aybat et al. (2019) propose a multi-stage accelerated algorithm that does not require knowledge of σ^2. The authors use a dynamical systems analysis, and prove that their algorithm achieves the desired optimal rate *only for $T \geq 2\sqrt{\kappa}$*.

Contribution: In contrast, in Section 4, we use SGD with a stochastic variant of Nesterov acceleration (Cohen et al., 2018; Vaswani et al., 2019a) and the same exponentially decreasing step-sizes. We refer to the resulting method as Accelerated SGD (ASGD). Compared to Aybat et al. (2019), ASGD is a more natural extension of the deterministic Nesterov accelerated gradient method. Under a growth condition (Li et al., 2020; Khaled and Richtárik, 2020; Bottou et al., 2018) similar to Aybat et al. (2019), we use the standard estimating sequences analysis, and prove that ASGD *achieves the desired rate for all T without the knowledge of σ^2*. Hence, exponentially decreasing step-sizes result in noise-adaptivity for both SGD and ASGD.

Contribution: ASGD requires the knowledge of both the smoothness and strong-convexity parameters. As a step towards problem-adaptivity for ASGD, we analyze its convergence with offline estimates of these problem-dependent constants. To the best of our knowledge this is the first such result. We prove that, similar to SGD, misspecified ASGD converges to the minimizer, but its rate is slowed down by a factor proportional to the estimation errors.

Contribution: Finally, in Section 5, we evaluate the performance of different step-size schemes on strongly-convex supervised learning problems. We show that (A)SGD consistently out-perform existing noise-adaptive algorithms. We propose a novel variant of SLS that guarantees convergence to the minimizer and demonstrate its practical effectiveness in making (A)SGD problem-adaptive.

Additional contributions: In Appendix B.1.1, we show matching results for SGD on strongly star-convex functions (Hinder et al., 2020), a class of structured non-convex functions. Finally, we prove upper-bounds for non-

strongly-convex functions (Appendix B.1.2) and show that even when the smoothness constant is known, exponentially decreasing step-sizes converge to a neighbourhood of the solution. We give some justification as to why polynomial or exponentially decreasing step-sizes are unlikely to be noise-adaptive in this setting.

2. Problem setup and Background

We assume that f and each f_i are differentiable and lower-bounded by f^* and f_i^*, respectively. Throughout the paper, we assume that f is μ-strongly convex, and each f_i is convex. We also assume that each function f_i is L_i-smooth, implying that f is L-smooth with $L := \max_i L_i$ (see Appendix A for the necessary definitions) and define $\kappa := \frac{L}{\mu}$.

We use stochastic gradient descent (SGD) or SGD with Nesterov acceleration (Nesterov, 2004) (referred to as ASGD) to minimize f in Eq. (1). In each iteration $k \in [T]$, SGD selects a function f_{ik} (typically uniformly) at random, computes its gradient and takes a descent step. Specifically,

$$w_{k+1} = w_k - \gamma_k \alpha_k \nabla f_{ik}(w_k), \qquad (2)$$

where w_{k+1} and w_k are the SGD iterates, and $\nabla f_{ik}(\cdot)$ is the gradient of the loss function chosen at iteration k. Each stochastic gradient $\nabla f_{ik}(w)$ is unbiased, implying that $\mathbb{E}_i [\nabla f_i(w)] = \nabla f(w)$. The product of scalars $\eta_k := \gamma_k \alpha_k$ defines the *step-size* for iteration k. The step-size consists of two parts — γ_k, a problem-dependent scaling term that captures the (local) smoothness of the function; and α_k, a problem-independent term that controls the decay of the step-size. Typically, α_k is a decreasing sequence of k, and $\lim_{k \to \infty} \alpha_k = 0$. The choice of the α_k sequence depends on the properties of f, for example, for smooth, strongly-convex functions, α_k is typically set to be $O(1/k)$.

Throughout the paper, we will assume that T is known in advance. In order to obtain noise-adaptive rates, we consider exponentially decreasing step-sizes (Li et al., 2020) of the form $\alpha_k := \alpha^k$ where $\alpha := \left[\frac{\beta}{T}\right]^{1/T} \leq 1$ for a constant $\beta \geq 1$. These step-sizes lie between the constant step-size used in the deterministic setting and the $1/k$ decreasing step-sizes used in the stochastic setting, meaning that for $k \in [T]$, $\alpha_k \in \left[\frac{1}{k}, 1\right]$.

In the next section, we analyze the convergence of SGD with exponentially decreasing step-sizes for smooth, strongly-convex functions.

3. Towards noise & problem adaptive SGD

In this section, we consider approaches for developing noise and problem-adaptive SGD i.e. we aim to obtain the noise-adaptive rate matching Stich (2019); Li et al. (2020);

Khaled and Richtárik (2020), but do so without the knowledge of problem-dependent constants.

Instead of the typical assumption of finite gradient noise $z^2 := \mathbb{E}_i[\|\nabla f_i(w^*)\|^2] < \infty$, we assume a finite optimal objective difference. Specifically, we define the noise as $\sigma^2 := \mathbb{E}_i[f_i(w^*) - f_i^*] \geq 0$. This notion of noise has been used to study the convergence of constant step-size SGD in the *interpolation* setting for over-parameterized models (Zhang and Zhou, 2019; Loizou et al., 2021; Vaswani et al., 2020). Note that when interpolation is exactly satisfied, $\sigma = z = 0$. In general, if each function f_i is μ-strongly convex and L-smooth, then $\frac{1}{2L}z^2 \leq \sigma^2 \leq \frac{1}{2\mu}z^2$.

As a warm-up, we first assume knowledge of the smoothness constant in Section 3.1 and analyze the resulting SGD algorithm with exponentially decreasing step-sizes. In Section 3.2, we consider using a stochastic line-search (Vaswani et al., 2019b; 2020) in order to estimate the smoothness constant and set the step-size on the fly. Finally, in Section 3.3, we analyze the convergence of SGD when using an offline estimate of the smoothness.

3.1. Known smoothness

We use the knowledge of smoothness to set the problem-dependent part of the step-size for SGD, specifically, $\gamma_k = 1/L$. With an exponentially decreasing α_k-sequence, we prove the following theorem in Appendix C.1.

Theorem 1. Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\gamma_k = \frac{1}{L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$ converges as,

$$\mathbb{E}\|w_{T+1} - w^*\|^2 \leq \|w_1 - w^*\|^2 c_2 \exp\left(-\frac{T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)$$
$$+ \frac{8\sigma^2 c_2 \kappa}{\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2 T},$$

where $c_2 = \exp\left(\frac{1}{\kappa} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

Compared to Moulines and Bach (2011) that use polynomially decreasing step-sizes, exponential step-sizes result in a better trade-off between the bias (initial distance to the minimizer) and variance (noise) terms, achieving the desired $\tilde{O}\left(\exp(-T/\kappa) + \sigma^2/T\right)$ noise-adaptive rate. In Lemmas 3 and 4 in Appendix B.3, we show that no polynomially decreasing step-size can result in the desired noise-adaptive rate. In order to interpolate between the stochastic (mini-batch size equal to 1) and fully deterministic (mini-batch size equal to n) setting, we show the explicit dependence of σ^2 on the mini-batch size in Appendix B.2.

Since strongly-convex functions also satisfy the PL condition (Karimi et al., 2016), the above result can be deduced

from (Li et al., 2020). However, unlike (Li et al., 2020), our result does not require the growth condition and uses a weaker notion of noise. Moreover, we use a different proof technique, specifically, Li et al. (2020) use the smoothness inequality in the first step and obtain the rate in terms of the function suboptimality, $\mathbb{E}[f(w_T) - f^*]$. In contrast, our proof uses an expansion of the iterates to obtain the rate in terms of the distance to the minimizer, $\mathbb{E} \|w_{T+1} - w^*\|^2$. This change allows us to easily handle the case when the smoothness constant is unknown and needs to be estimated.

Next, we use stochastic line-search techniques to estimate the unknown smoothness and set the step-size on the fly.

3.2. Online estimation of unknown smoothness

In this section, we assume that the smoothness constant is unknown, and aim to estimate it and set the step-size in an *online* fashion. By online estimation, we mean that in iteration k, we use the knowledge of the sampled function i_k to set the step-size, i.e. setting γ_k depends on i_k. We only consider methods that use the knowledge of i_k in iteration k and are not allowed to access the other functions in f (for example, to compute the full-batch gradient at w_k). Methods based on a stochastic line-search (Vaswani et al., 2019b; 2020) or the stochastic Polyak step-size (Loizou et al., 2021; Berrada et al., 2020) satisfy this criterion.

We use stochastic line-search (SLS) to estimate the local Lipschitz constant and set γ_k, the problem-dependent part of the step-size. SLS is the stochastic analog of the traditional Armijo line-search (Armijo, 1966) used for deterministic gradient descent (Nocedal and Wright, 2006). In iteration k, SLS estimates the smoothness constant L_{ik} of the sampled function using f_{ik} and ∇f_{ik}. In particular, starting from a guess ($\gamma_{\max}$) of the step-size, SLS uses a backtracking procedure and returns the largest step-size γ_k that satisfies: $\gamma_k \leq \gamma_{\max}$ and,

$$f_{ik}(w_k - \gamma_k \nabla f_{ik}(w_k)) \leq f_{ik}(w_k) - c\,\gamma_k \|\nabla f_{ik}(w_k)\|^2. \tag{3}$$

Here, $c \in (0, 1)$ is a hyper-parameter to be determined theoretically. SLS guarantees that resulting the step-size γ_k lies in the $\left[\min\left\{\frac{2(1-c)}{L_{ik}}, \gamma_{\max}\right\}, \gamma_{\max}\right]$ range (Lemma 8). If the initial guess is large enough i.e. $\gamma_{\max} > 1/L_{ik}$, then the resulting step-size $\gamma_k \geq \frac{2(1-c)}{L_{ik}}$. Thus, with $c = 1/2$, SLS can be used to obtain an upper-bound on $1/L_{ik}$.

In the interpolation ($\sigma = 0$) setting, a constant step-size ($\alpha_k = 1$ for all k) suffices, and SGD with SLS achieves a linear rate of convergence (for $c \geq 1/2$) when minimizing smooth, strongly-convex functions (Vaswani et al., 2019b). In general, for a non-zero σ, using SGD with SLS and no step-size decay ($\alpha_k = 1$) results in

$O\left(\exp(-T/\kappa) + \gamma_{\max}\sigma^2\right)$ rate (Vaswani et al., 2020), implying convergence to a neighbourhood determined by the $\gamma_{\max}\sigma^2$ term.

In order to obtain a similar rate as Theorem 1 but without the knowledge of L, we set γ_k with SLS and use the same exponentially decreasing α_k-sequence. We prove the following theorem in Appendix C.2.

Theorem 2. Under the same assumptions as Theorem 1, SGD (Eq. (2)) with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and Eq. (3) with $c = 1/2$ converges as,

$$\mathbb{E} \|w_{T+1} - w^*\|^2 \leq \|w_1 - w^*\|^2 c_1 \exp\left(-\frac{T}{\kappa'} \frac{\alpha}{\ln(T/\beta)}\right)$$
$$+ \frac{8\sigma^2 c_1 (\kappa')^2 \gamma_{\max}}{e^2} \frac{(\ln(T/\beta))^2}{\alpha^2 T}$$
$$+ \frac{2\sigma^2 c_1 \kappa' \ln(T/\beta) \gamma_{\mathrm{err}}}{e\alpha},$$

where $\gamma_{\mathrm{err}} := \left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$, $\kappa' := \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

We observe that the first two terms are similar to those in Theorem 1. For $\gamma_{\max} \geq \frac{1}{L}$, $\kappa' = \kappa$ and the above theorem implies the same $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ rate of convergence. However, as $T \to \infty$, w_{T+1} does not converge to w^*, but rather to a neighbourhood determined by the last term $\frac{2\sigma^2 \kappa' c_1 \ln(T/\beta)}{e\alpha} \left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$. The neighbourhood thus depends on the noise σ^2 and γ_{err}, the estimation error (in the smoothness) of the initial guess.

When $\sigma^2 = 0$, this neighbourhood term disappears, and SGD converges to the minimizer despite the estimation error. This matches the result for SLS in the interpolation setting (Vaswani et al., 2019b). Conversely, when the smoothness is known and $\gamma_{\max}$ can be set equal to $\frac{1}{L}$, we also obtain convergence to the minimizer and recover the result of Theorem 1. In fact, if we can "guess" a value of $\gamma_{\max} \leq \frac{1}{L}$, it would result in the neighbourhood term becoming zero, thus ensuring convergence to the minimizer. In this case, the stochastic line-search does not decrease the step-size in any iteration, and the algorithm becomes the same as using a constant step-size equal to $\gamma_{\max}$. Finally, we contrast our result with the $\alpha_k = 1$ setting (Vaswani et al., 2020), and observe that instead of the dependence on $\gamma_{\max}$, our neighbourhood term depends on the estimation error in the smoothness. Next, we show the necessity of such a neighbourhood term.

3.2.1. LOWER BOUND ON QUADRATICS

In order to prove a lower-bound, we consider a pair of 1-dimensional quadratics $f_i(w) = 1/2\,(x_i w - y_i)^2$ for

$i = 1, 2$. Here, w, x_i, y_i are all scalars. The overall function to be minimized is $f(w) = (1/2) \cdot [f_1(w) + f_2(w)]$. We assume that $\|x_1\| \neq \|x_2\|$, and since $L_i = \|x_i\|^2$, this assumption implies different smoothness constants for the two functions. For a sufficiently large value of $\gamma_{\max}$ i.e. $\left(\gamma_{\max} \geq \frac{1}{\min_{i \in [2]} L_i}\right)$, using SLS with $c \geq 1/2$ (required for convergence) results in $\gamma_k \leq 1/L_{ik}$[1] (see Lemma 8). With these choices, we prove the following lower-bound.

Theorem 3. When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w) + f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w - 1)^2$ and $f_2(w) = \frac{1}{2}(2w + 1/2)^2$, setting γ_k using SLS with $\gamma_{\max} \geq 1$ and $c \geq 1/2$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if w^* is the minimizer of f and $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

The above result (proved in Appendix D.1) shows that using SGD with SLS to set γ_k and *any convergent sequence of* α_k (including the exponentially-decreasing sequence in Theorem 2) will necessarily result in convergence to a neighbourhood. The neighbourhood term can thus be viewed as the *price of misestimation* of the unknown smoothness constant. This result is in contrast to the conventional thinking that choosing an α_k sequence such that $\lim_{k \to \infty} \alpha_k = 0$ will always ensure convergence to the minimizer. Note that this result is not specific to SLS and would hold for other related methods (Loizou et al., 2021; Berrada et al., 2020).

Since the lower-bound holds for any convergent α_k sequence, a possible reason for this convergence to the neighbourhood is the correlation between i_k and the computation of γ_k. We investigate this hypothesis in the next section.

3.3. Offline estimation of unknown smoothness

In this section, we consider an offline estimation of the smoothness constant. By offline, we mean that in iteration k, γ_k is set *before* sampling i_k and cannot use any information about it. This ensures that γ_k is decorrelated with the sampled function i_k. The entire sequence of γ_k can even be chosen before running SGD.

For simplicity of calculations, we consider a fixed $\gamma_k = \gamma$ for all iterations. Here γ is an offline estimate of $\frac{1}{L}$, and can be obtained by any method. Without loss of generality, we assume that this offline estimate is off by a multiplicative factor ν that is $\gamma = \frac{\nu}{L}$ for some $\nu > 0$. Here ν quantifies

---

[1] For 1-dimensional quadratics, $\gamma_k = 1/L_{ik}$ for $c = 1/2$.

the estimation error in γ with $\nu = 1$ corresponding to an exact estimation of L. In practice, it is typically possible to obtain lower-bounds on the smoothness constant. Hence, the $\nu > 1$ regime is of practical interest. For SGD with $\gamma_k = \gamma = \frac{\nu}{L}$ and an exponentially decreasing α_k-sequence, we prove the following theorem in Appendix C.3.

Theorem 4. Under the same assumptions as Theorem 1, SGD (Eq. (2)) with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ converges as,

$$\Delta_{T+1} \leq \Delta_1 \, c_2 \exp\left(-\frac{\min\{\nu, 1\} \, T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right)$$

$$+ \max\{\nu^2, 1\} \frac{8 c_2 \kappa \, \ln(T/\beta)}{\mu \, e^2 \, \alpha^2 \, T} \left[2\sigma^2 \ln(T/\beta) + G \, [\ln(\nu)]_+\right]$$

where $c_2 = \exp\left(\frac{1}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $k_0 = \lfloor T \frac{[\ln(\nu)]_+}{\ln(T/\beta)} \rfloor$, $G = \max_{j \in [k_0]}\{f(w_j) - f^*\}$ and $\Delta_k := \|w_k - w^*\|^2$.

The above theorem implies an $\tilde{O}\left(\exp\left(-\frac{\min\{\nu, 1\} \, T}{\kappa}\right) + \frac{\max\{\nu^2, 1\} \left[\sigma^2 + G[\ln(\nu)]_+\right]}{T}\right)$ convergence to the minimizer. The first two terms are similar to that in Theorem 1 and imply an $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence to the minimizer. Analyzing the third term, we observe that when $\nu \leq 1$, the third term is zero (since $[\ln(\nu)]_+ = 0$), and the rate matches that of Theorem 1 up to constants that depend on ν. The third term depends on $[\max_{j \in k_0}\{f(w_j) - f^*\}]$ because if $\nu > 1$, the step-size $\gamma_k \alpha_k = \frac{\nu}{L} \alpha_k \geq \frac{1}{L}$ initially, and SGD diverges in this regime. Since α_k is an exponentially decreasing sequence, after $k_0 := T \frac{\ln(\nu)}{\ln(T/\beta)}$ iterations, $\frac{\nu}{L} \alpha_k \leq \frac{1}{L}$, the distance to the minimizer decreases after iteration k_0, eventually converging to the solution.

Furthermore, observe that the second term depends on $\tilde{O}\left(\max\{\nu^2, 1\}\right)$ meaning that if we misestimate the smoothness constant by a multiplicative factor of $\nu > 1$, it can slow down the convergence rate by an $O(\nu^2)$ factor. Finally, our theorem implies that even in the deterministic setting, misestimating L can slowdown the convergence rate to $O\left(\frac{\nu^2}{T}\right)$ instead of the usual linear rate of convergence. The third term can thus be viewed as the *price of misestimation* of the unknown smoothness constant. Unlike Theorem 2 where this price was convergence to a neighbourhood, here, the price of misestimation is slower convergence to the minimizer.

Moulines and Bach (2011) also considered the effect of misspecifying L but in conjunction with polynomially decreasing step-sizes. Specifically, they proved that using a step-size of $\frac{\nu}{L} \frac{1}{T^\theta}$ results in the following bounds that depend on γ and ν (Moulines and Bach, 2011, Theorem 1).

Below, we show their bounds for three common choices of $\theta = \{0, 1/2, 1\}$ and emphasize the effect of ν.

$$\Delta_{T+1} = O\left(\exp\left((\nu^2 - \nu/\kappa)\, T\right)(\Delta_1 + \sigma^2) + \nu\sigma^2\right)$$
$$\text{(When } \theta = 0)$$

$$= O\left(\exp\left(\nu^2 \ln(T) - \nu/\kappa\, \sqrt{T}\right)(\Delta_1 + \sigma^2) + \frac{\nu\sigma^2}{\sqrt{T}}\right)$$
$$\text{(When } \theta = \tfrac{1}{2})$$

$$= O\left(\exp\left(\nu^2 - \nu/\kappa\, \ln(T)\right)(\Delta_1 + \sigma^2) + \frac{\nu^2\sigma^2}{T^{\nu/2\kappa}}\right)$$
$$\text{(When } \theta = 1 \text{ and } \nu < 2\kappa)$$

$$= O\left(\exp\left(\nu^2 - \nu/\kappa\, \ln(T)\right)(\Delta_1 + \sigma^2) + \frac{\nu^2\sigma^2}{T}\right)$$
$$\text{(When } \theta = 1 \text{ and } \nu \geq 2\kappa)$$

Observe that for each regime, the convergence rate depends on $\exp(\nu)$. In contrast, the convergence rate in Theorem 4 depends on $O(\nu^2)$. This robustness towards misspecification can be viewed as an additional advantage of using exponentially decreasing step-sizes. In the next section, we justify the dependence on $[\ln(\nu)]_+$ in Theorem 4 by proving a corresponding lower-bound. It is unclear whether the ν^2 dependence in Theorem 4 is tight, and we leave verifying this for future work.

3.3.1. LOWER BOUND ON QUADRATICS

In this section, we consider gradient descent on a one-dimensional quadratic and study the effect of misestimating the smoothness constant by a factor of $\nu > 1$. We consider minimizing a single quadratic, ensuring that $\sigma^2 = 0$ and prove the following lower-bound in Appendix D.2.

Theorem 5. When minimizing a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, GD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$, satisfies

$$w_{k+1} - w^* = (w_1 - w^*)\prod_{i=1}^{k}(1 - \nu\alpha_i).$$

After $k' := \frac{T}{\ln(T/\beta)}\ln\left(\frac{\nu}{3}\right)$ iterations, we have that

$$|w_{k'+1} - w^*| \geq 2^{k'}|w_1 - w^*|.$$

Instantiating this lower-bound, suppose the estimate of L is off by a factor of $\nu = 10$, then $\ln\left(\frac{\nu}{3}\right) \geq 1$, which implies that $k' \geq \lfloor\frac{T}{\ln(T/\beta)}\rfloor$. In other words, we do not make any progress in the first $\frac{T}{\ln(T/\beta)}$ iterations, and at this point the optimality gap has been multiplied by a factor of $2^{T/\ln(T/\beta)}$ compared to the starting optimality gap. This simple example shows the slowdown in the rate of convergence by misestimating the smoothness.

4. Towards noise & problem adaptive ASGD

In this section, we will first aim to use SGD with Nesterov acceleration and obtain the optimal $\tilde{O}\left(\exp\left(\frac{-T}{\sqrt{\kappa\rho}}\right) + \frac{\sigma^2}{T}\right)$ rate without knowledge of σ^2. Subsequently, we will analyze the convergence of ASGD with offline estimates of the smoothness and strong-convexity parameters, quantifying the price of misspecification (similar to Section 3.3).

ASGD has two sequences $\{w_k, y_k\}$ and an additional extrapolation parameter b_k. ASGD computes the stochastic gradient at the extrapolated point y_k and takes a descent step in that direction. The update in iteration k is:

$$y_k = w_k + b_k(w_k - w_{k-1}), \tag{4}$$
$$w_{k+1} = y_k - \gamma_k\alpha_k\nabla f_{ik}(y_k). \tag{5}$$

For analyzing the convergence of ASGD, we will assume that the stochastic gradients satisfy a growth condition similar to Bottou et al. (2018); Li et al. (2020); Khaled and Richtárik (2020)[2] – there exists a (ρ, σ) with $\rho \geq 1$ and $\sigma \geq 0$, such that for all w,

$$\mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2 + \sigma^2. \tag{6}$$

In the deterministic setting (when using the full-gradient in Eq. (5)), $\rho = 1$ and $\sigma = 0$. Similarly, $\sigma = 0$ when the stochastic gradients satisfy the *strong-growth condition* when using over-parameterized models (Schmidt and Roux, 2013; Vaswani et al., 2019a). [3] Using this growth condition, we prove the following result in Appendix E.3.

Theorem 6. Under the same assumptions of Theorem 1 and (iii) the growth condition in Eq. (6), ASGD (Eqs. (4) and (5)) with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

$$\Delta_{T+1} \leq 2c_3 \exp\left(-\frac{T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\Delta_1$$
$$+ \frac{2\sigma^2 c_3}{\rho\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2 T},$$

where $\Delta_k := \mathbb{E}[f(w_k) - f^*]$ and $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$.

---

[2]The growth condition in Bottou et al. (2018, (Eq 4.8)) is $\mathbb{E}_i \|\nabla f(w) - \nabla f_i(w)\|^2 \leq M_V \|\nabla f(w)\|^2 + M$ for some constants M and M_V. Since the stochastic gradient is unbiased, this is the same as Eq. (6) with $\rho = M_V + 1$ and $\sigma^2 = M$.

[3]Note that the noise-model in Section 3 is equivalent (upto constants) to assuming that the variance of the stochastic gradients at the *optimum* is bounded. However, the noise-model in Eq. (6) is equivalent to assuming that the variance of the stochastic gradients at *any iterate* is bounded. Hence, the two noise-models are similar, though Eq. (6) makes a stronger assumption on the noise.

The above theorem implies that ASGD achieves an $\tilde{O}\left(\exp\left(\frac{-T}{\sqrt{\kappa\rho}}\right) + \frac{\sigma^2}{T}\right)$ convergence rate. This improves over the non-accelerated $\tilde{O}\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\sigma^2}{T}\right)$ noise-adaptive rate obtained in Theorem 1 and Stich (2019); Khaled and Richtárik (2020); Li et al. (2020)[4]. In the fully-deterministic setting ($\rho = 1$ and $\sigma = 0$), Theorem 6 implies an $\tilde{O}(\exp(-T/\sqrt{\kappa}))$ convergence to the minimizer, matching the optimal rate in the deterministic setting (Nesterov, 2004). Under the strong-growth condition (when $\sigma = 0$), ASGD improves (by a $\sqrt{\rho}$ factor) over the rate in Vaswani et al. (2019a) and matches (upto log factors) the rate in Mishkin (2020). In the general stochastic case, when $\sigma \neq 0$, Cohen et al. (2018); Vaswani et al. (2019a) use constant step-sizes ($\alpha_k = 1$), and prove convergence to a neighbourhood of the solution; whereas we show convergence to the minimizer at a rate governed by the $O(\sigma^2/T)$ term. To smoothly interpolate between the stochastic (batch size equal to 1) and fully deterministic (batch size equal to n) setting, we generalize the growth condition to show an explicit dependence on the batch size (Appendix B).

Comparing our result to that in Aybat et al. (2019), we note that they prove the accelerated noise-adaptive rate under two different assumptions on the noise. (i) Bounded variance of the stochastic gradients. This is a special case of the growth condition in Eq. (6) with $\rho = 1$. (ii) They also consider a weaker assumption which is equivalent to Eq. (6) up to constants (see Appendix G for a detailed comparison of the two assumptions, and the resulting upper-bounds). Using a dynamical systems perspective, Aybat et al. (2019, Corollary 3.8) prove the desired rate *only for* $T \geq 2\sqrt{\kappa}$, whereas our proof uses the more standard Nesterov's estimate sequences and our result holds for all T.

The result in Theorem 6 requires the knowledge of both μ and L and is thus not problem-adaptive. In the next section, we analyze the convergence of ASGD when it is used with offline estimates of L and μ.

4.1. Offline estimation of unknown smoothness & strong-convexity

Similar to Section 3.3, for simplicity, we will assume that $\gamma_k = \gamma = \frac{1}{\tilde{L}}$ where without loss of generality, $\frac{1}{\tilde{L}} = \frac{\nu_L}{L}$. Similarly, we use $\tilde{\mu}$ as the offline estimate of the strong-convexity, and assume that $\tilde{\mu} = \nu_\mu\mu$. We will only consider the case where we underestimate μ, and hence $\nu_\mu \leq 1$. This is the typical case in practice – for example, while

optimizing regularized convex loss functions in supervised learning (see Section 5 for empirical results), $\tilde{\mu}$ is set to the regularization strength, and thus underestimates the true strong-convexity parameter. The following theorem (proved in Appendix E.4.1) analyzes the effect of misspecifying L, μ on the ASGD convergence.

Theorem 7. Under the same assumptions as Theorem 6 and (iv) $\nu = \nu_L\nu_\mu \leq \rho\kappa$, ASGD (Eqs. (4) and (5)) with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho\tilde{L}} = \frac{\nu_L}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\tilde{\mu} = \nu_\mu\mu \leq \mu$, $r_k = \sqrt{\frac{\tilde{\mu}}{\rho\tilde{L}}}\left(\frac{\beta}{T}\right)^{k/2T} = \sqrt{\frac{\nu}{\rho\kappa}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

$$\Delta_{T+1} \leq 2c_3 \exp\left(-\frac{\sqrt{\min\{\nu, 1\}}T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\Delta_1$$
$$+ \frac{2c_3(\ln(T/\beta))^2}{e^2\alpha^2\mu T}\left[\frac{\sigma^2}{\rho} + G^2\min\{\frac{k_0}{T}, 1\}\right]\max\{\frac{\nu_L}{\nu_\mu}, \nu_L^2\},$$

where $\Delta_k := \mathbb{E}[f(w_k) - f^*]$, $c_3 = \exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $k_0 := \lfloor T\frac{[\ln(\nu_L)]_+}{\ln(T/\beta)}\rfloor$ and $G = \max_{j\in[k_0]} \|\nabla f(y_j)\|$.

The above theorem implies an $\tilde{O}\left(\exp\left(\frac{-T\sqrt{\min\{\nu, 1\}}}{\sqrt{\kappa\rho}}\right) + \left[\frac{\sigma^2 + G^2[\ln(\nu_L)]_+}{T}\right]\max\{\frac{\nu_L}{\nu_\mu}, \nu_L^2\}\right)$ convergence to the minimizer. Observe that (i) when the problem-dependent parameters are known ($\nu_\mu = \nu_L = 1$), we recover the rate of Theorem 6, (ii) if $\nu_\mu = 1$, and we misestimate L, similar to SGD (Theorem 4), ASGD converges to the minimizer at an $O(1/T)$ rate, even in the deterministic setting (when $\sigma = 0$), (iii) if $\nu_L = 1$, underestimating μ matches the rate in Theorem 6 upto (potentially large) constants, resulting in linear convergence when $\sigma = 0$, and (iv) compared to Theorem 6, the decrease in the bias term is slowed down by an $O\left(\exp(\sqrt{\min\{\nu_L\nu_\mu, 1\}})\right)$ factor, whereas the decrease in the variance is slowed by an $O\left(\max\{\frac{\nu_L}{\nu_\mu}, \nu_L^2\}\right)$ factor.

In the next section, we design an SLS variant that ensures convergence to the minimizer while empirically controlling the misestimation for both SGD and ASGD.

5. Experiments

For comparing different step-size choices, we consider two common supervised learning losses – squared loss for regression tasks and logistic loss for classification[5]. With

---

[4]Though we typically expect $\rho < \kappa$, in the worst-case stochastic setting, ρ can be as large as κ and we can not obtain an accelerated rate (with a dependence on $\sqrt{\kappa}$) in this case (Liu and Belkin, 2018). Mini-batching (see Appendix B.2) is one way to decrease the "effective" ρ. For a large enough batch-size $\rho < \kappa$ ensuring a $\sqrt{\kappa}$ dependence in the stochastic setting.

[5]The code to reproduce our experiments is available here: https://github.com/R3za/expsls
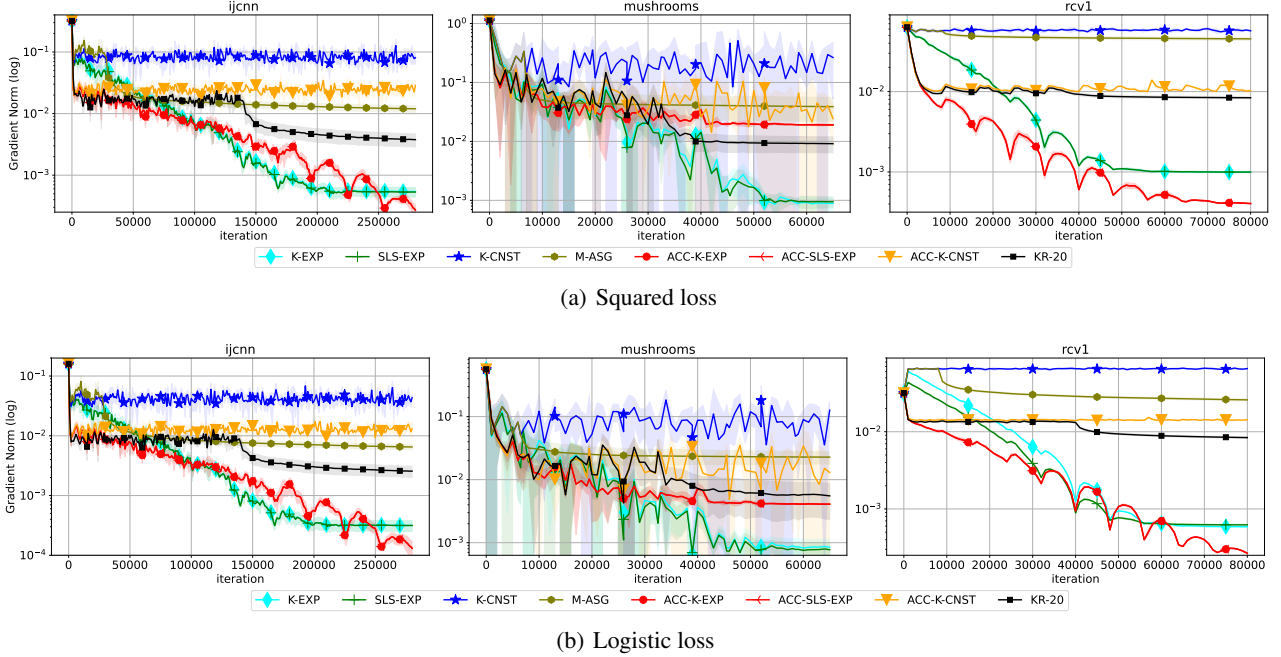
(a) Squared loss

(b) Logistic loss

Figure 1. Comparison for (a) squared loss and (b) logistic loss. Observe that exponentially decreasing step-sizes (i) result in more stable performance compared to using a constant step-size (for both SGD and ASGD) and (ii) consistently outperform the noise-adaptive methods in `KR-20` and `M-ASG`, and (iii) methods using the SLS in Eq. (7) match the performance of those with known smoothness.

a linear model and an ℓ_2 regularization equal to $\frac{\lambda}{2}\|w\|^2$, both objectives are strongly-convex. We use three standard datasets from LIBSVM (Chang and Lin, 2011) – *mushrooms*, *ijcnn* and *rcv1*, and use $\lambda = 0.01$. For each experiment, we consider 5 independent runs and plot the average result and standard deviation. We use the (full) gradient norm as the performance measure and plot it against the number of gradient evaluations.

For each dataset, we fix $T = 10n$, use a batch-size of 1 and compare the performance of the following optimization strategies: (i) the noise-adaptive "constant and then decay step-size" scheme in Khaled and Richtárik (2020, Theorem 3) (denoted as `KR-20` in the plots). Specifically, for $b = \max\{\frac{2L^2}{\mu}, 2\rho L\}$, we use a constant step-size equal to $1/b$ when $T < b/\mu$ or $k < \lceil T/2 \rceil$. Otherwise we set the step-size at iteration k to be $\frac{2}{\mu((2b/\mu)+k-\lceil T/2 \rceil)}$, (ii) constant step-size SGD with $\gamma_k = \frac{1}{L}$ and $\alpha_k = 1$ for all k (denoted as `K-CNST` in the plots) (iii) SGD with an exponentially decreasing step-size with knowledge of smoothness (Li et al., 2020) i.e. $\gamma_k = \frac{1}{L}$ and $\alpha_k = \alpha^k$ for $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$ (denoted as `K-EXP`), (iv) Accelerated SGD (ASGD) with a constant step-size ($\alpha_k = 1$ for all k) (Vaswani et al., 2019a; Cohen et al., 2018) (denoted as `ACC-K-CNST`), (v) ASGD with exponentially decreasing step-sizes, (Section 4) denoted as `ACC-K-EXP` and (vi) Multistage ASGD in Aybat et al. (2019) (denoted as `M-ASG`) with parameters as in Corollary 3.8. Specifically we ensure that $T > 2\sqrt{\kappa}$,

set $T_1 = T/C, T_k > 2^k[\sqrt{\kappa}\log(2^{(p+2)})], \alpha_1 = 1/L$ and $\alpha_k = 1/(2^{2k}L)$ where p and C are hyper-parameters.

None of the above strategies are problem-adaptive, and all of them require the knowledge of the smoothness constant L. Additionally, `KR-20` and the ASGD variants also require knowledge of ρ, the parameter of the growth condition in Eq. (6), while the ASGD variants and `M-ASG` require knowledge of μ. If x_i is the feature vector corresponding to example i, then we obtain theoretical upper-bounds on the smoothness and set $L = \max_i \|x_i\|^2 + \lambda$ for the squared-loss and $L = \max_i \frac{1}{4}\|x_i\|^2 + \lambda$ for the logistic loss. Similarly, we set $\mu = \lambda$ for both the squared and logistic loss. Note that this underestimates the true strong-convexity parameter, and is in line with Theorem 7. To set ρ, we use a grid search over $\{10, 100, 1000\}$. Similarly, to set p and C for `M-ASG`, we use a grid search over $\{1, 2, 4\}$ and $\{2, 10, 100\}$ respectively. For each method, we plot the variant that results in the smallest gradient norm.

Using a stochastic line-search (SLS) to estimate L can result in convergence to the neighbourhood (Section 3.2) because of the correlations between i_k and γ_k. To alleviate this, and still be problem-adaptive, we design a *decorrelated conservative* variant of SLS: at iteration k of SGD, we set γ_k using a stochastic line-search on the *previously sampled function* i_{k-1} (we can use a randomly sampled j_k as well). This ensures that there is no correlation between i_k and computing γ_k. The overall procedure can be

described as follows: starting from γ_{k-1} (the conservative aspect), with $\gamma_0 = \gamma_{\max}$, find the largest step-size γ_k that satisfies, for a random or previously sampled index j_k,

$$f_{j_k}(w_k - \gamma_k \nabla f_{j_k}(w_k)) \leq f_{j_k}(w_k) - c\gamma_k \|\nabla f_{j_k}(w_k)\|^2, \tag{7}$$

and update w_k according to Eq. (2). The above procedure with $c = 1/2$ ensures that $\gamma_k \in [\min\{\gamma_{k-1}, 1/L\}, \gamma_{k-1}]$. Since there is no correlation between γ_k and i_k, we can treat γ_k as an offline estimate of the smoothness, meaning that $\gamma_k = \nu_k/L$ for some $\nu_k > 0$. Moreover, since we are using a conservative line-search, $\gamma_k \in [\min\{\gamma_{k-1}, 1/L\}, \gamma_{k-1}]$, meaning that $\nu_k \leq \nu_{k-1} \leq \nu_1$. Hence the maximum misspecification in the smoothness is given by $\nu_1 > 1$, which is governed by line-search in the first iteration. Given this, we can use a similar analysis as Theorem 4, upper-bounding ν_k by ν_1 for each k and obtaining the corresponding result in terms of ν_1.

We use this variant of SLS with exponentially decreasing step-sizes for both SGD and ASGD, and denote the resulting variants as SLS-EXP and ACC-SLS-EXP respectively. We emphasize that this strategy is both noise-adaptive and problem-adaptive.

From Fig. 1, we observe that exponentially decreasing step-sizes (i) result in more stable performance compared to the constant step-size variants (for both SGD and ASGD) and (ii) consistently outperform the noise-adaptive methods, KR-20 and M-ASG. We also observe that (iii) methods (SLS-EXP and ACC-SLS-EXP) using the SLS condition in Eq. (7) consistently match the performance of those with known smoothness (K-EXP and ACC-K-EXP).

6. Conclusion

We used exponentially decreasing step-sizes to make SGD noise-adaptive, and considered two strategies for problem-adaptivity. Using upper and lower-bounds, we quantified the price of problem-adaptivity – estimating the smoothness in an online fashion results in convergence to a neighbourhood of the solution, while an offline estimation results in a slower convergence to the minimizer. We then developed an accelerated variant of SGD (ASGD) and proved that it achieves the near-optimal convergence rate. We analyzed the effect of misspecifying the strong-convexity and smoothness parameters for ASGD. Finally, we empirically demonstrated the effectiveness of (A)SGD with exponential step-sizes coupled with a novel variant of SLS.

7. Acknowledgements

References

Arjevani, Y., Shamir, O., and Srebro, N. (2020). A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pages 111–132. PMLR.

Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*.

Aybat, N. S., Fallah, A., Gurbuzbalaban, M., and Ozdaglar, A. (2019). A universally optimal multistage accelerated stochastic gradient method. *Advances in neural information processing systems*, 32:8525–8536.

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 *International Conference on Machine Learning*, pages 799–809. PMLR.

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

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cohen, M., Diakonikolas, J., and Orecchia, L. (2018). On acceleration with noise-corrupted gradients. In *International Conference on Machine Learning*, pages 1019–1028. PMLR.

Duchi, J. C., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492.

Ghadimi, S. and Lan, G. (2013). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089.

Gower, R., Sebbouh, O., and Loizou, N. (2021). Sgd for structured nonconvex functions: Learning rates, mini-batching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR.

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

Hinder, O., Sidford, A., and Sohoni, N. (2020). Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on Learning Theory*, pages 1894–1938. PMLR.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. (2018). Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

Khaled, A. and Richtárik, P. (2020). Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*.

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

Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR.

Kulunchakov, A. and Mairal, J. (2019). Estimate sequences for variance-reduced stochastic composite optimization. In *International Conference on Machine Learning*, pages 3541–3550. PMLR.

Levy, K. Y., Yurtsever, A., and Cevher, V. (2018). Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems, NeurIPS*.

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

Li, X., Zhuang, Z., and Orabona, F. (2020). A second look at exponential and cosine step sizes: Simplicity, convergence, and performance. *arXiv preprint arXiv:2002.05273*.

Liu, C. and Belkin, M. (2018). Accelerating sgd with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*.

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

Loizou, N., Vaswani, S., Laradji, I. H., and Lacoste-Julien, S. (2021). Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR.

Lucas, J., Bae, J., Zhang, M. R., Fort, S., Zemel, R., and Grosse, R. (2021). Analyzing monotonic linear interpolation in neural network loss landscapes. *arXiv preprint arXiv:2104.11044*.

Mishkin, A. (2020). *Interpolation, growth conditions, and stochastic gradient descent*. PhD thesis, University of British Columbia.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24:451–459.

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

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media.

Nguyen, P. H., Nguyen, L. M., and van Dijk, M. (2018). Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in sgd. *arXiv preprint arXiv:1810.04723*.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

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

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

Stich, S. U. (2019). Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*.

Vaswani, S., Bach, F., and Schmidt, M. (2019a). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR.

Vaswani, S., Laradji, I. H., Kunstner, F., Meng, S. Y., Schmidt, M., and Lacoste-Julien, S. (2020). Adaptive gradient methods converge faster with over-parameterization (and you can do a line-search).

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., and Lacoste-Julien, S. (2019b). Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32:3732–3745.

Zhang, L. and Zhou, Z.-H. (2019). Stochastic approximation of smooth and strongly convex functions: Beyond the $o(1/t)$ convergence rate. In *Conference on Learning Theory*, pages 3160–3179. PMLR.

Organization of the Appendix

A. Definitions

Our main assumptions are that each individual function f_i is differentiable, has a finite minimum f_i^*, and is L_i-smooth, meaning that for all v and w,

$$f_i(v) \leq f_i(w) + \langle \nabla f_i(w), v - w \rangle + \frac{L_i}{2} \|v - w\|^2, \qquad \text{(Individual Smoothness)}$$

which also implies that f is L-smooth, where L is the maximum smoothness constant of the individual functions. A consequence of smoothness is the following bound on the norm of the stochastic gradients,

$$\|\nabla f_i(w)\|^2 \leq 2L(f_i(w) - f_i^*).$$

We also assume that each f_i is convex, meaning that for all v and w,

$$f_i(v) \geq f_i(w) - \langle \nabla f_i(w), w - v \rangle, \qquad \text{(Convexity)}$$

Depending on the setting, we will also assume that f is μ strongly-convex, meaning that for all v and w,

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\mu}{2} \|v - w\|^2, \qquad \text{(Strong Convexity)}$$

B. Additional theoretical results

In this section, we relax the strong-convexity assumption to handle broader function classes in Appendix B.1 and prove results that help provide an explicit dependence on the mini-batch size (Appendix B.2) and in Appendix B.3 show that polynomially decreasing step-sizes cannot obtain the desired noise-adaptive rate.

B.1. Relaxing the assumptions

In this section, we extend our theoretical results to a richer class of functions - strongly quasar-convex functions (Hinder et al., 2020) in Appendix B.1.1, and (non-strongly) convex functions in Appendix B.1.2.

B.1.1. EXTENSION TO STRONGLY STAR-CONVEX FUNCTIONS

We consider the class of smooth, non-convex, but strongly star-convex functions (Hinder et al., 2020; Gower et al., 2021), a subset of strongly quasar-convex functions. Quasar-convex functions are unimodal along lines that pass through a global minimizer i.e. the function monotonically decreases along the line to the minimizer, and monotonically increases thereafter. In addition to this, strongly quasar-convex functions also have curvature near the global minimizer. Importantly, this property is satisfied for neural networks for common architectures and learning problems (Lucas et al., 2021; Kleinberg et al., 2018).

Formally, a function is (ζ, μ) strongly quasar-convex if it satisfies the following for all w and minimizers w^*,

$$f(w^*) \geq f(w) + \frac{1}{\zeta}\langle \nabla f(w), w^* - w \rangle + \frac{\mu}{2}\|w - w^*\|^2. \tag{8}$$

Strongly star-convex functions are a subset of this class of functions with $\zeta = 1$. If L is known, it is straightforward to show that the results of Theorem 1 carry over to the strongly star-convex functions and we obtain the similar $O\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ rate. In the case when L is not known, it was recently shown that SGD with a stochastic Polyak step-size (Gower et al., 2021) results in linear convergence to the minimizer on strongly star-convex functions under interpolation and achieves an $O\left(\exp(-T) + \gamma_{\max}\sigma^2\right)$ convergence rate in general. The proposed stochastic Polyak step-size (SPS) does not require knowledge of L, and matches the rate achieved for strongly-convex functions (Loizou et al., 2021). However, SPS requires knowledge of f_i^*, which is usually zero for machine learning models under interpolation but difficult to get a handle on in the general case.

Consequently, we continue to use SLS to estimate the smoothness constant. Our proofs only use strong-convexity between w and a minimizer w^*, and hence we can extend all our results from strongly-convex functions, to structured non-convex functions satisfying the strongly star-convexity property, matching the rates in Theorem 2 and Theorem 4. Finally, we note that given knowledge of ζ, there is no fundamental limitation in extending all our results to strongly quasar-convex functions. In the next section, we relax the strong-convexity assumption in a different way - by considering convex functions without curvature.

B.1.2. HANDLING (NON-STRONGLY)-CONVEX FUNCTIONS

In this section, we analyze the behaviour of exponentially decreasing step-sizes on convex functions (without strong-convexity). As a starting point, we assume that L is known, and the algorithm is only required to adapt to the noise σ^2. In the following theorem (proved in Appendix C.4), we show that SGD with an exponentially decreasing step-size is not guaranteed to converge to the minimizer, but to a neighbourhood of the solution.

Theorem 8. Assuming (i) convexity and (ii) L_i-smoothness of each f_i, SGD with step-size $\eta_k = \frac{1}{2L}\alpha_k$ has the following convergence rate,

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\|w_1 - w^*\|^2}{\sum_{k=1}^T \alpha_k} + \sigma^2 \frac{\sum_{k=1}^T \alpha_k^2}{\sum_{k=1}^T \alpha_k} \tag{9}$$

where $\bar{w}_{T+1} = \frac{\sum_{k=1}^T \alpha_k w_k}{\sum_{k=1}^T \alpha_k}$. For $\alpha_k = \left[\frac{\beta}{T}\right]^{k/T}$, the convergence rate is given by,

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\ln(T/\beta)\|w_1 - w^*\|^2}{\alpha T - 2\beta} + \sigma^2 \frac{T}{T - \beta}$$

We thus see that even with the knowledge of L, SGD converges to a neighbourhood of the solution at an $O(1/T)$ rate. We contrast our result to AdaGrad (Duchi et al., 2011; Levy et al., 2018) that adapts the step-sizes as the algorithm progresses (as opposed to using a predetermined sequence of step-sizes like in our case), is able to adapt to the noise, and achieves an $O\left(\frac{1}{T} + \frac{\sigma^2}{\sqrt{T}}\right)$ rate.

In order to be noise-adaptive and match the AdaGrad rate, we can use Eq. (9) to infer that a sufficient condition is for the α_k-sequence to satisfy the following inequalities, (i) $\alpha_k \geq C_1 T$ and (ii) $\alpha_k^2 \leq C_2\sqrt{T}$ where C_1, C_2 are constants. Unfortunately, in Lemmas 9 and 10, we prove that it is not possible for *any* polynomially or exponentially-decreasing sequence to satisfy these sufficient conditions. While we do not have a formal lower-bound in the convex case, it seems unlikely that these α_k-sequences can result in the desired rate, and we conjecture a possible lower-bound. Finally, we note that to the best of our knowledge, the only predetermined (non-adaptive) step-size that achieves the AdaGrad rate is $\min\left\{\frac{1}{2L}, \frac{1}{\sigma\sqrt{T}}\right\}$ (Ghadimi and Lan, 2012). We also conjecture a lower-bound that shows that there is no predetermined

sequence of step-sizes (that does not use knowledge of σ^2) that is noise-adaptive and can achieve the $O\left(\frac{1}{T} + \frac{\sigma^2}{\sqrt{T}}\right)$ rate.

B.2. Dependence on the mini-batch size

In this section, we prove two results in order to explicitly model the dependence on the mini-batch size. We denote a mini-batch as $\mathcal{B}$, its size as $B \in [1, n]$ and the corresponding mini-batch gradient as $\nabla f_{\mathcal{B}}(w) = \frac{1}{B}\sum_{f_i \in \mathcal{B}} \nabla f_i(w)$. The mini-batch gradient is also unbiased i.e. $\mathbb{E}_{\mathcal{B}}[\nabla f_{\mathcal{B}}(w)] = \nabla f(w)$, implying that all the proofs remain unchanged, but we need to use a different growth condition for the ASGD proofs in Section 4 and a different definition of σ for the SGD proofs in Section 3. We refine these quantities here, and show the explicit dependence on the mini-batch size.

Lemma 1. *If*
$$\mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2 + \sigma^2,$$
then,
$$\mathbb{E}_{\mathcal{B}} \|\nabla f_{\mathcal{B}}(w))\|^2 \leq \left((\rho - 1)\frac{n - B}{nB} + 1\right) \|\nabla f(w)\|^2 + \frac{n - B}{nB}\sigma^2.$$

Proof.

$$\mathbb{E}_{\mathcal{B}} \|\nabla f_{\mathcal{B}}(w))\|^2 = \mathbb{E}_{\mathcal{B}} \|\nabla f_{\mathcal{B}}(w) - \nabla f(w) + \nabla f(w)\|^2 = \mathbb{E}_{\mathcal{B}} \|\nabla f_{\mathcal{B}}(w) - \nabla f(w)\|^2 + \|\nabla f(w)\|^2$$

$$\text{(Since } \mathbb{E}_{\mathcal{B}}[\nabla f_{\mathcal{B}}(w)] = \nabla f(w))$$

Since we are sampling the batch with replacement, using (Lohr, 2019),

$$\leq \frac{n - B}{nB}\left(E_i \|\nabla f_i(w)\|^2 - \|\nabla f(w)\|^2\right) + \|\nabla f(w)\|^2$$

$$\leq \frac{n - B}{nB}\left((\rho - 1)\|\nabla f(w)\|^2 + \sigma^2\right) + \|\nabla f(w)\|^2 \qquad \text{(Using the growth condition)}$$

$$\implies \mathbb{E}_{\mathcal{B}} \|\nabla f_{\mathcal{B}}(w))\|^2 \leq \left((\rho - 1)\frac{n - B}{nB} + 1\right) \|\nabla f(w)\|^2 + \frac{n - B}{nB}\sigma^2.$$

$$\square$$

Lemma 2. *If*
$$\sigma^2 := \mathbb{E}[f_i(w^*) - f_i^*],$$
and each function f_i is μ strongly-convex and L-smooth, then
$$\sigma_B^2 := \mathbb{E}_{\mathcal{B}}[f_{\mathcal{B}}(w^*) - f_{\mathcal{B}}^*] \leq \frac{L}{\mu}\frac{n - B}{nB}\sigma^2.$$

Proof.

By strong-convexity of f_i,

$$\mathbb{E}_{\mathcal{B}}[f_{\mathcal{B}}(w^*) - f_{\mathcal{B}}^*] \leq \frac{1}{2\mu}\mathbb{E}_{\mathcal{B}} \|\nabla f_{\mathcal{B}}(w^*))\|^2$$

Since we are sampling the batch with replacement, using (Lohr, 2019),

$$\leq \frac{1}{2\mu}\frac{n - B}{nB}\mathbb{E}_i \|\nabla f_i(w^*)\|^2$$

$$\leq \frac{L}{\mu}\frac{n - B}{nB}\mathbb{E}[f_i(w^*) - f_i^*] \qquad \text{(By smoothness of } f_i)$$

$$\implies \sigma_B^2 \leq \frac{L}{\mu}\frac{n-B}{nB}\sigma^2.$$

$\square$

B.3. Polynomially decaying stepsizes

In this section, we analyze polynomially decreasing step-sizes, namely when $\eta_k = \frac{\eta}{(k+1)^\delta}$ for some constants $\eta > 0$ and $0 \leq \delta \leq 1$. We argue that even with knowledge of the smoothness constant, these step-sizes fail to converge at the desired noise-adaptive rate even on simple quadratics. In particular, the next lemma shows that gradient descent (GD) applied to a strongly-convex quadratic with a polynomially decreasing step-size fails to obtain the usual linear rate of the form $O(\rho^{-T})$ for some $\rho < 1$.

> **Lemma 3.** *When using T iterations of GD to minimize a one-dimensional quadratic $f(w) = \frac{1}{2}(xw - y)^2$, setting $\eta_k = \frac{1}{L}\frac{1}{(k+1)^\delta}$ for some $0 < \delta \leq 1$ results in the following lower bounds. If $\delta = 1$,*
>
> $$w_{T+1} - w^* = (w_1 - w^*)\frac{1}{T+1}$$
>
> *If $0 < \delta < 1$, $w_1 - w^* > 0$ and T is large enough,*
>
> $$w_{T+1} - w^* \geq (w_1 - w^*)\left(1 - \frac{1}{2^\delta}\right)^{\lfloor 2^{1/\delta}\rfloor - 1} 4^{\frac{2^\delta - 1}{1-\delta}} 4^{-\frac{(T+1)^{1-\delta}}{1-\delta}}$$

Proof. Observe that $w^* = y/x$ and $L = x^2$. The GD iteration with $\eta_k = \frac{1}{L}\frac{1}{(k+1)^\delta}$ reads

$$w_{k+1} = w_k - \frac{1}{L}\frac{1}{(k+1)^\delta}\left(x^2 w_k - xy\right) = w_k\left(1 - \frac{1}{(k+1)^\delta}\right) + \frac{y}{x}\frac{1}{(k+1)^\delta} = w_k\left(1 - \frac{1}{(k+1)^\delta}\right) + w^*\frac{1}{(k+1)^\delta}$$

and thus

$$w_{k+1} - w^* = (w_k - w^*)\left(1 - \frac{1}{(k+1)^\delta}\right) \Rightarrow w_{T+1} - w^* = (w_1 - w^*)\prod_{k=1}^{T}\left(1 - \frac{1}{(k+1)^\delta}\right)$$

If $\delta = 1$,

$$w_{T+1} - w^* = (w_1 - w^*)\prod_{k=1}^{T}\frac{k}{k+1} = (w_1 - w^*)\frac{1}{T+1}$$

If $0 < \delta < 1$ and $w_1 - w^* > 0$,

$$w_{T+1} - w^* = (w_1 - w^*)\prod_{k=1}^{T}\left(1 - \frac{1}{(k+1)^\delta}\right) = (w_1 - w^*)\prod_{k=1}^{T}\left(1 - \frac{2}{2(k+1)^\delta}\right)$$

We wish to use the inequality $1 - \frac{2x}{2} \geq 2^{-2x}$ which is true for all $x \in [0, 1/2]$. In our case it holds for

$$\frac{1}{(k+1)^\delta} \leq \frac{1}{2} \Rightarrow k \geq 2^{1/\delta} - 1$$

Let $k_0 = \lfloor 2^{1/\delta}\rfloor$. Then for $T \geq k_0$,

$$w_{T+1} - w^* = (w_1 - w^*)\prod_{k=1}^{k_0-1}\left(1 - \frac{1}{(k+1)^\delta}\right)\prod_{k=k_0}^{T}\left(1 - \frac{2}{2(k+1)^\delta}\right)$$

Now, for $k \le k_0 - 1$, we have that $\frac{1}{(k+1)^\delta} \le \frac{1}{2^\delta}$ and thus

$$\prod_{k=1}^{k_0-1} \left(1 - \frac{1}{(k+1)^\delta}\right) \ge \left(1 - \frac{1}{2^\delta}\right)^{k_0-1} = \left(1 - \frac{1}{2^\delta}\right)^{\lfloor 2^{1/\delta} \rfloor - 1}$$

For $k \ge k_0$, we have $1 - \frac{2}{2(k+1)^\delta} \ge 2^{-2\frac{1}{(k+1)^\delta}}$ and thus

$$\prod_{k=k_0}^{T} \left(1 - \frac{2}{2(k+1)^\delta}\right) \ge 2^{-2\sum_{k=k_0}^{T}\frac{1}{(k+1)^\delta}} = 2^{-2\left(\sum_{k=1}^{T+1}\frac{1}{k^\delta} - \sum_{k=1}^{k_0}\frac{1}{k^\delta}\right)} \ge 2^{-2\sum_{k=1}^{T+1}\frac{1}{k^\delta}}$$

Using the bound in the proof of Lemma 9, we have

$$\sum_{k=1}^{T+1} \frac{1}{k^\delta} \le 1 + \frac{1}{1-\delta}\left((T+1)^{1-\delta} - 1\right)$$

Putting this together we have that

$$2^{-2\sum_{k=1}^{T+1}\frac{1}{k^\delta}} \ge 2^{-2\left(1 + \frac{1}{1-\delta}\left((T+1)^{1-\delta}-1\right)\right)} = \frac{4^{1/(1-\delta)}}{4} 4^{-\frac{(T+1)^{1-\delta}}{1-\delta}} = 4^{\frac{2\delta-1}{1-\delta}} 4^{-\frac{(T+1)^{1-\delta}}{1-\delta}}$$

Putting everything together we get that

$$w_{T+1} - w^* \ge (w_1 - w^*)\left(1 - \frac{1}{2^\delta}\right)^{\lfloor 2^{1/\delta} \rfloor - 1} 4^{\frac{2\delta-1}{1-\delta}} 4^{-\frac{(T+1)^{1-\delta}}{1-\delta}}$$

$\square$

The next lemma shows that when $\delta = 0$, namely when the step-size is constant, SGD applied to the sum of two quadratics fails to converge to the minimizer.

Lemma 4. *When using SGD to minimize the sum* $f(w) = \frac{f_1(w)+f_2(w)}{2}$ *of two one-dimensional quadratics:* $f_1(w) = \frac{1}{2}(w-1)^2$ *and* $f_2(w) = \frac{1}{2}(2w+\frac{1}{2})^2$ *with a constant step-size* $\eta = \frac{1}{L}$*, the following holds: whenever* $|w_k - w^*| < \frac{1}{8}$*, the next iterate satisfies* $|w_{k+1} - w^*| > \frac{1}{8}$*.*

Proof. First observe that $w^* = 0$ and that $L = 4$. The updates then read

$$\text{If } i_k = 1: \quad w_{k+1} = w_k - \eta(w_k - 1) = w_k\left(1 - \frac{1}{4}\right) + \frac{1}{4} = \frac{3}{4}w_k + \frac{1}{4}$$

$$\text{If } i_k = 2: \quad w_{k+1} = w_k - \eta 2\left(2w_k + \frac{1}{2}\right) = w_k\left(1 - \frac{4}{4}\right) - \frac{1}{4} = -\frac{1}{4}$$

Suppose that $|w_k - w^*| = |w_k| < \frac{1}{8}$. We want to show that $|w_{k+1}| > \frac{1}{8}$. We can separate the analyses in three cases. If $w_k \in (-\frac{1}{8}, 0)$ and $i_k = 1$ then

$$w_{k+1} = \frac{3}{4}w_k + \frac{1}{4} > -\frac{3}{4} \times \frac{1}{8} + \frac{1}{4} = \frac{5}{32} > \frac{1}{8}$$

If $w_k \in (0, \frac{1}{8})$ and $i_k = 1$ then

$$w_{k+1} = \frac{3}{4}w_k + \frac{1}{4} > \frac{1}{8}$$

If $i_k = 2$ then

$$w_{k+1} = -\frac{1}{4} < -\frac{1}{8}$$

implying that in each case, $|w_{k+1}| > 1/8$. $\qquad\square$

C. Upper-bound Proofs for Section 3

C.1. Proof of Theorem 1

Theorem 1. Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\gamma_k = \frac{1}{L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$ converges as,

$$\mathbb{E} \left\| w_{T+1} - w^* \right\|^2 \le \left\| w_1 - w^* \right\|^2 c_2 \exp\left(-\frac{T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right)$$
$$+ \frac{8\sigma^2 c_2 \kappa}{\mu e^2} \frac{(\ln(T/\beta))^2}{\alpha^2 T},$$

where $c_2 = \exp\left(\frac{1}{\kappa} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

Proof.

$$\left\| w_{k+1} - w^* \right\|^2 = \left\| w_k - \eta_k \nabla f_{ik}(w_k) - w^* \right\|^2$$
$$= \left\| w_k - w^* \right\|^2 - 2\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \eta_k^2 \left\| \nabla f_{ik}(w_k) \right\|^2$$
$$= \left\| w_k - w^* \right\|^2 - 2\gamma_k \alpha_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \gamma_k^2 \alpha_k^2 \left\| \nabla f_{ik}(w_k) \right\|^2$$
$$\left\| w_{k+1} - w^* \right\|^2 \le \left\| w_k - w^* \right\|^2 - 2\gamma_k \alpha_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \gamma_k^2 \alpha_k^2 \, 2L[f_{ik}(w_k) - f_{ik}^*] \qquad \text{(Smoothness)}$$
$$= \left\| w_k - w^* \right\|^2 - \frac{2}{L}\alpha_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \frac{2}{L}\alpha_k^2 \left[f_{ik}(w_k) - f_{ik}(w^*)\right] + \frac{2}{L}\alpha_k^2 \left[f_{ik}(w^*) - f_{ik}^*\right]$$
$$\text{(Since } \gamma_k = 1/L.\text{)}$$

Taking expectation w.r.t i_k,

$$\mathbb{E} \left\| w_{k+1} - w^* \right\|^2 \le \mathbb{E} \left\| w_k - w^* \right\|^2 - \frac{2}{L}\alpha_k \langle \nabla f(w_k), w_k - w^* \rangle + \frac{2}{L}\alpha_k^2 \left[f(w_k) - f(w^*)\right] + \frac{2}{L}\alpha_k^2 \sigma^2$$
$$\le \mathbb{E} \left\| w_k - w^* \right\|^2 - \frac{2}{L}\alpha_k \langle \nabla f(w_k), w_k - w^* \rangle + \frac{2}{L}\alpha_k \left[f(w_k) - f(w^*)\right] + \frac{2}{L}\alpha_k^2 \sigma^2 \qquad \text{(Since } \alpha_k \le 1\text{)}$$
$$\mathbb{E} \left\| w_{k+1} - w^* \right\|^2 \le \left(1 - \frac{\mu \alpha_k}{L}\right) \mathbb{E} \left\| w_k - w^* \right\|^2 + \frac{2}{L}\alpha_k^2 \sigma^2 \qquad \text{(By } \mu\text{-strong convexity of } f)$$

Unrolling the recursion starting from w_1 and using the exponential step-sizes,

$$\mathbb{E} \left\| w_{T+1} - w^* \right\|^2 \le \left\| w_1 - w^* \right\|^2 \prod_{k=1}^{T}\left(1 - \frac{\mu \alpha^k}{L}\right) + \frac{2\sigma^2}{L} \sum_{k=1}^{T} \left[\prod_{i=k+1}^{T} \alpha^{2k}\left(1 - \frac{\mu \alpha^i}{L}\right)\right]$$

Writing $\Delta_k = \mathbb{E} \left\| w_k - w^* \right\|^2$

$$\Delta_{T+1} \le \Delta_1 \underbrace{\exp\left(-\frac{\mu}{L}\sum_{k=1}^{T} \alpha^k\right)}_{:=A} + \frac{2\sigma^2}{L} \sum_{k=1}^{T} \alpha^{2k} \underbrace{\exp\left(-\frac{\mu}{L}\sum_{i=k+1}^{T} \alpha^i\right)}_{:=B_t}$$

Using Lemma 5 to lower-bound A, we obtain $A \geq \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}$. The first term in the above expression can then be bounded as,

$$\Delta_1 \exp\left(-\frac{\mu}{L}A\right) = \Delta_1\, c_2 \exp\left(-\frac{T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right),$$

where $\kappa = \frac{L}{\mu}$ and $c_2 = \exp\left(\frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$. Using Lemma 6 to upper-bound B_t, we obtain $B_t \leq \frac{4\kappa^2 c_2 (\ln(T/\beta))^2}{e^2\alpha^2 T}$, thus bounding the second term. Putting everything together,

$$\Delta_{T+1} \leq \Delta_1\, c_2 \exp\left(-\frac{T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right) + \frac{8\sigma^2 c_2 \kappa^2}{Le^2}\frac{(\ln(T/\beta))^2}{\alpha^2 T}$$

$\square$

C.2. Proof of Theorem 2

> **Theorem 2.** Under the same assumptions as Theorem 1, SGD (Eq. (2)) with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and Eq. (3) with $c = 1/2$ converges as,
>
> $$\mathbb{E}\left\|w_{T+1} - w^*\right\|^2 \leq \left\|w_1 - w^*\right\|^2 c_1 \exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right)$$
> $$+ \frac{8\sigma^2 c_1 (\kappa')^2 \gamma_{\max}}{e^2}\frac{(\ln(T/\beta))^2}{\alpha^2 T}$$
> $$+ \frac{2\sigma^2 c_1 \kappa' \ln(T/\beta)\, \gamma_{\mathrm{err}}}{e\alpha},$$
>
> where $\gamma_{\mathrm{err}} := \left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$,
> $\kappa' := \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

Proof.

$$\left\|w_{k+1} - w^*\right\|^2 \leq \left\|w_k - w^*\right\|^2 - 2\gamma_k\alpha_k\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + \gamma_k\alpha_k^2\left[\frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \qquad \text{(By Lemma 8)}$$

Setting $c = 1/2$,

$$= \left\|w_k - w^*\right\|^2 - 2\gamma_k\alpha_k\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + 2\gamma_k\alpha_k^2\left[f_{ik}(w_k) - f_{ik}^*\right]$$
$$= \left\|w_k - w^*\right\|^2 - 2\gamma_k\alpha_k\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + 2\gamma_k\alpha_k^2\left[f_{ik}(w_k) - f_{ik}(w^*)\right] + 2\gamma_k\alpha_k^2\left[f_{ik}(w^*) - f_{ik}^*\right]$$

Adding, subtracting $2\gamma_k\alpha_k[f_{ik}(w_k) - f_{ik}(w^*)]$,

$$= \left\|w_k - w^*\right\|^2 + 2\gamma_k\alpha_k\left[-\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + [f_{ik}(w_k) - f_{ik}(w^*)]\right] - 2\gamma_k\alpha_k[f_{ik}(w_k) - f_{ik}(w^*)]$$
$$+ 2\gamma_k\alpha_k^2\left[f_{ik}(w_k) - f_{ik}(w^*)\right] + 2\gamma_k\alpha_k^2\left[f_{ik}(w^*) - f_{ik}^*\right]$$
$$\leq \left\|w_k - w^*\right\|^2 + 2\gamma_{\min}\alpha_k\left[-\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + [f_{ik}(w_k) - f_{ik}(w^*)]\right]$$
$$- 2\gamma_k(\alpha_k - \alpha_k^2)[f_{ik}(w_k) - f_{ik}(w^*)] + 2\gamma_{\max}\alpha_k^2\left[f_{ik}(w^*) - f_{ik}^*\right]$$

where we used convexity of f_{ik} to ensure that $-\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + [f_{ik}(w_k) - f_{ik}(w^*)] \leq 0$. Taking expectation,

$$\mathbb{E}\left\|w_{k+1} - w^*\right\|^2 \leq \left\|w_k - w^*\right\|^2 + 2\gamma_{\min}\alpha_k\left[-\langle\nabla f(w_k), w_k - w^*\rangle + [f(w_k) - f(w^*)]\right]$$

$$- (\alpha_k - \alpha_k^2)\mathbb{E}\left[2\gamma_k[f_{ik}(w_k) - f_{ik}(w^*)]\right] + 2\gamma_{\max}\alpha_k^2\sigma^2$$

$$\mathbb{E}\left\|w_k - w^*\right\|^2 \le (1 - \alpha_k\gamma_{\min}\mu)\left\|w_k - w^*\right\|^2 - (\alpha_k - \alpha_k^2)\mathbb{E}\left[2\gamma_k[f_{ik}(w_k) - f_{ik}(w^*)]\right] + 2\gamma_{\max}\alpha_k^2\sigma^2$$

Since $\alpha_k \le 1$, and $\alpha_k - \alpha_k^2 \ge 0$, let us analyze $-\mathbb{E}[[\gamma_k[f_{ik}(w_k) - f_{ik}(w^*)]]]$.

$$
\begin{aligned}
-\mathbb{E}[[\gamma_k[f_{ik}(w_k) - f_{ik}(w^*)]]] &= -\mathbb{E}[[\gamma_k[f_{ik}(w_k) - f_{ik}^*]] - \mathbb{E}[[\gamma_k[f_{ik}^* - f_{ik}(w^*)]] \\
&\le -\mathbb{E}[[\gamma_{\min}[f_{ik}(w_k) - f_{ik}^*]] - \mathbb{E}[[\gamma_{\max}[f_{ik}^* - f_{ik}(w^*)]] &&(\gamma_k \le \gamma_{\max}) \\
&= -\mathbb{E}[[\gamma_{\min}[f_{ik}(w_k) - f_{ik}^*]] + \gamma_{\max}\sigma^2 \\
&= -\mathbb{E}[[\gamma_{\min}[f_{ik}(w_k) - f_{ik}(w^*)]] - \mathbb{E}[[\gamma_{\min}[f_{ik}(w^*) - f_{ik}^*]] + \gamma_{\max}\sigma^2 \\
&= -\gamma_{\min}[f(w_k) - f(w^*)] - \gamma_{\min}\sigma^2 + \gamma_{\max}\sigma^2 \\
&\le (\gamma_{\max} - \gamma_{\min})\sigma^2
\end{aligned}
$$

Putting this relation back,

$$
\begin{aligned}
\mathbb{E}\left\|w_k - w^*\right\|^2 &\le (1 - \alpha_k\gamma_{\min}\mu)\left\|w_k - w^*\right\|^2 + 2(\alpha_k - \alpha_k^2)(\gamma_{\max} - \gamma_{\min})\sigma^2 + 2\gamma_{\max}\alpha_k^2\sigma^2 \\
&\le (1 - \alpha_k\gamma_{\min}\mu)\left\|w_k - w^*\right\|^2 + 2\alpha_k(\gamma_{\max} - \gamma_{\min})\sigma^2 + 2\gamma_{\max}\alpha_k^2\sigma^2.
\end{aligned}
$$

Setting $\kappa' = \max\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\}$ we get that $1 - \alpha_k\gamma_{\min}\mu \le 1 - \frac{1}{\kappa'}$. Writing $\Delta_k = \mathbb{E}\left\|w_k - w^*\right\|^2$ and unrolling the recursion we get

$$
\begin{aligned}
\Delta_{T+1} &\le \left(\prod_{k=1}^{T}\left(1 - \frac{1}{\kappa'}\alpha^k\right)\right)\Delta_1 + 2\gamma_{\max}\sigma^2\sum_{k=1}^{T}\alpha^{2k}\prod_{i=t+1}^{T}\left(1 - \frac{1}{\kappa'}\alpha^i\right) + 2\sigma^2\sum_{k=1}^{T}\alpha^k(\gamma_{\max} - \gamma_{\min})\prod_{i=k+1}^{T}\left(1 - \frac{1}{\kappa'}\alpha^i\right) \\
&\le \Delta_1\exp\left(-\frac{1}{\kappa'}\underbrace{\sum_{k=1}^{T}\alpha^k}_{:=A}\right) + 2\gamma_{\max}\sigma^2\underbrace{\sum_{k=1}^{T}\alpha^{2k}\exp\left(-\frac{1}{\kappa'}\sum_{i=k+1}^{T}\alpha^i\right)}_{:=B_t} \\
&\quad + 2\sigma^2(\gamma_{\max} - \gamma_{\min})\underbrace{\sum_{k=1}^{T}\alpha^k\exp\left(-\frac{1}{\kappa'}\sum_{i=k+1}^{T}\alpha^i\right)}_{:=C_t}
\end{aligned}
$$

Using Lemma 5 to lower-bound A, we obtain $A \ge \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}$. The first term in the above expression can then be bounded as,

$$\Delta_1\exp\left(-\frac{1}{\kappa'}A\right) \le \Delta_1 c_1\exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right),$$

where $c_1 = \exp\left(\frac{1}{\kappa'}\frac{2\beta}{\ln(T/\beta)}\right)$. Using Lemma 6 to upper-bound B_t, we obtain $B_t \le \frac{4(\kappa')^2 c_1(\ln(T/\beta))^2}{e^2\alpha^2 T}$, thus bounding the second term. Using Lemma 7 to upper-bound C_t, we obtain $C_t \le c_1\frac{\kappa'\ln(T/\beta)}{e\alpha}$, thus bounding the third term. Finally, by Lemma 8 we have that $\gamma_{\min} \ge \min\left\{\gamma_{\max}, \frac{1}{L}\right\}$.

Putting everything together,

$$\Delta_{T+1} \le \Delta_1 c_1\exp\left(-\frac{T}{\kappa'}\frac{\alpha}{\ln(T/\beta)}\right) + \frac{8\sigma^2 c_1(\kappa')^2\gamma_{\max}}{e^2}\frac{(\ln(T/\beta))^2}{\alpha^2 T} + \frac{2c_1\sigma^2\kappa'\ln(T/\beta)}{e\alpha}\left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$$

$\square$

C.3. Proof of Theorem 4

Theorem 4. Under the same assumptions as Theorem 1, SGD (Eq. (2)) with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ converges as,

$$\Delta_{T+1} \le \Delta_1\, c_2 \exp\left(-\frac{\min\{\nu, 1\}\, T}{\kappa}\, \frac{\alpha}{\ln(T/\beta)}\right)$$

$$+ \max\{\nu^2, 1\}\frac{8c_2\kappa}{\mu\, e^2}\frac{\ln(T/\beta)}{\alpha^2\, T}\left[2\sigma^2 \ln(T/\beta) + G\,[\ln(\nu)]_+\right]$$

where $c_2 = \exp\left(\frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $k_0 = \lfloor T\frac{[\ln(\nu)]_+}{\ln(T/\beta)}\rfloor$, $G = \max_{j \in [k_0]}\{f(w_j) - f^*\}$ and $\Delta_k := \|w_k - w^*\|^2$.

Proof. Following the steps from the proof of Theorem 1,

$$\|w_{k+1} - w^*\|^2 \le \|w_k - w^*\|^2 - 2\gamma_k\alpha_k\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + 2L\gamma_k^2\alpha_k^2\left[f_{ik}(w_k) - f_{ik}(w^*)\right] + 2L\gamma_k^2\alpha_k^2\left[f_{ik}(w^*) - f_{ik}^*\right]$$

Taking expectation wrt i_k, and since both γ_k and α_k are independent of i_k,

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \le \|w_k - w^*\|^2 - 2\gamma_k\alpha_k\langle\nabla f(w_k), w_k - w^*\rangle + 2L\gamma_k^2\alpha_k^2\left[f(w_k) - f^*\right] + 2L\gamma_k^2\alpha_k^2\sigma^2$$

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \le (1 - \mu\gamma_k\alpha_k)\|w_k - w^*\|^2 + 2L\gamma_k^2\alpha_k^2\sigma^2 + [f(w_k) - f^*]\left(2L\gamma_k^2\alpha_k^2 - 2\gamma_k\alpha_k\right)$$

$$\text{(By strong convexity)}$$

Let us separately consider the $\nu \le 1$ and $\nu > 1$ case.

For the $\nu \le 1$ case, $(2L\gamma_k^2\alpha_k^2 - 2\gamma_k\alpha_k) = \frac{2\nu^2\alpha_k^2}{L} - \frac{2\nu\alpha_k}{L} \le \frac{2\nu\alpha_k}{L} - \frac{2\nu\alpha_k}{L} = 0$. Hence, the above equation can be simplified as:

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \le \left(1 - \frac{\mu\nu\alpha_k}{L}\right)\|w_k - w^*\|^2 + \frac{2\nu^2\alpha_k^2}{L}\sigma^2$$

Proceeding in the same way as the proof of Theorem 1, define $\Delta_k = \mathbb{E}\|w_k - w^*\|^2$ and unroll the recursion,

$$\Delta_{T+1} \le \Delta_1 \prod_{k=1}^{T}(1 - \frac{\mu\nu}{L}\alpha_k) + \left(\frac{2\nu^2\sigma^2}{L}\right)\sum_{k=1}^{T}\alpha_k^2 \prod_{i=k+1}^{T}(1 - \frac{\mu\nu}{L}\alpha_i)$$

Bounding the first term similar to Lemma 5,

$$\prod_{k=1}^{T}(1 - \frac{\mu\nu}{L}\alpha_k) \le \exp\left(-\frac{\mu\nu}{L}\frac{\alpha - \alpha^{T+1}}{1 - \alpha}\right) \le \exp\left(-\frac{\mu\nu}{L}\frac{\alpha T - 2\beta}{\ln(T/\beta)}\right) = \exp\left(-\frac{\nu T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\exp\left(\frac{\nu}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$$

Bounding the second term similar to Lemma 6,

$$\sum_{k=1}^{T}\alpha_k^2 \prod_{i=k+1}^{T}(1 - \frac{\mu\nu}{L}\alpha_i) \le \sum_{k=1}^{T}\alpha_k^2 \exp\left(-\frac{\mu\nu}{L}\sum_{i=k+1}^{T}\alpha^i\right)$$

$$= \sum_{k=1}^{T}\alpha_k^2 \exp\left(-\frac{\nu}{\kappa}\frac{\alpha^{k+1} - \alpha^{T+1}}{1 - \alpha}\right)$$

$$= \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right) \sum_{k=1}^{T} \alpha_k^2 \exp\left(-\frac{\nu\alpha^{k+1}}{\kappa(1-\alpha)}\right)$$

$$\leq \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right) \sum_{k=1}^{T} \alpha_k^2 \left(\frac{2(1-\alpha)\kappa}{\nu e \alpha^{k+1}}\right)^2$$

$$= \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right) \frac{4(1-\alpha)^2\kappa^2}{\nu^2 e^2 \alpha^2} T$$

$$\leq \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right) \frac{4\kappa^2}{\nu^2 e^2 \alpha^2} \frac{\ln(T/\beta)^2}{T}$$

$$\leq \exp\left(\frac{\nu}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right) \frac{4\kappa^2}{\nu^2 e^2 \alpha^2} \frac{\ln(T/\beta)^2}{T}$$

Putting everything together, we obtain that,

$$\Delta_{T+1} \leq \Delta_1 \exp\left(-\frac{\nu T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right) \exp\left(\frac{\nu}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right) + \frac{2\sigma^2}{L} \exp\left(\frac{\nu}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right) \frac{4\kappa^2}{e^2 \alpha^2} \frac{\ln(T/\beta)^2}{T}$$

$$\implies \Delta_{T+1} \leq \Delta_1 c_2 \exp\left(-\frac{\nu T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right) + \frac{2\sigma^2 c_2}{L} \frac{4\kappa^2}{e^2 \alpha^2} \frac{\ln(T/\beta)^2}{T} \qquad \text{(Since } \nu \leq 1\text{)}$$

For the $\nu > 1$ case,

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \leq \left(1 - \frac{\mu\nu\alpha_k}{L}\right)\|w_k - w^*\|^2 + \frac{2\nu^2\alpha_k^2\sigma^2}{L} + [f(w_k) - f^*]\left(2L\gamma_k^2\alpha_k^2 - 2\gamma_k\alpha_k\right)$$

$$\leq \left(1 - \frac{\mu\alpha_k}{L}\right)\|w_k - w^*\|^2 + \frac{2\nu^2\alpha_k^2\sigma^2}{L} + [f(w_k) - f^*]\left(\frac{2\nu^2\alpha_k^2}{L} - \frac{2\nu\alpha_k}{L}\right) \qquad \text{(Since } \nu > 1\text{)}$$

For the last term to be negative, we require $\alpha_k \leq \frac{1}{\nu}$. By definition of α_k, this will happen after $k \geq k_0 := T\frac{\ln(\nu)}{\ln(T/\beta)}$ iterations. However, until k_0 iterations, we observe that $(2L\gamma_k^2\alpha_k^2 - 2\gamma_k\alpha_k) \leq \frac{2\nu(\nu-1)}{L}\alpha_k^2$.

For the $k < k_0$ regime,

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \leq \left(1 - \frac{\mu\alpha_k}{L}\right)\|w_k - w^*\|^2 + 2L\gamma_k^2\alpha_k^2\sigma^2 + \max_{j\in[k_0]}\{f(w_j) - f^*\}\frac{2\nu(\nu-1)}{L}\alpha_k^2$$

Writing $\Delta_k = \mathbb{E}\|w_k - w^*\|^2$, and unrolling the recursion for the first k_0 iterations we get

$$\Delta_{k_0} \leq \Delta_1 \prod_{k=1}^{k_0-1}\left(1 - \frac{\mu}{L}\alpha_k\right) + \left(\underbrace{2\frac{\nu^2}{L}\sigma^2 + \max_{j\in[k_0]}\{f(w_j) - f^*\}\frac{2\nu(\nu-1)}{L}}_{:=c_5}\right) \sum_{k=1}^{k_0-1}\alpha_k^2 \prod_{i=k+1}^{k_0-1}\left(1 - \frac{\mu}{L}\alpha_i\right)$$

Bounding the first term similar to [Lemma 5](#),

$$\prod_{k=1}^{k_0-1}\left(1 - \frac{\mu}{L}\alpha_k\right) \leq \exp\left(-\frac{\mu}{L}\frac{\alpha - \alpha^{k_0}}{1-\alpha}\right)$$

Bounding the second term similar to [Lemma 6](#),

$$\sum_{k=1}^{k_0-1}\alpha_k^2 \prod_{i=k+1}^{k_0-1}\left(1 - \frac{\mu}{L}\alpha_i\right) \leq \sum_{k=1}^{k_0-1}\alpha_k^2 \exp\left(-\frac{\mu}{L}\sum_{i=k+1}^{k_0-1}\alpha^i\right)$$

$$= \sum_{k=1}^{k_0-1} \alpha_k^2 \exp\left(-\frac{1}{\kappa}\frac{\alpha^{k+1}-\alpha^{k_0}}{1-\alpha}\right)$$

$$= \exp\left(\frac{\alpha^{k_0}}{\kappa(1-\alpha)}\right)\sum_{k=1}^{k_0-1} \alpha_k^2 \exp\left(-\frac{\alpha^{k+1}}{\kappa(1-\alpha)}\right)$$

$$\leq \exp\left(\frac{\alpha^{k_0}}{\kappa(1-\alpha)}\right)\sum_{k=1}^{k_0-1} \alpha_k^2 \left(\frac{2(1-\alpha)\kappa}{e\alpha^{k+1}}\right)^2$$

$$\leq \exp\left(\frac{\alpha^{k_0}}{\kappa(1-\alpha)}\right)\frac{4(1-\alpha)^2\kappa^2}{e^2\alpha^2}k_0$$

$$\leq \exp\left(\frac{\alpha^{k_0}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\frac{k_0\ln(T/\beta)^2}{T^2}$$

Putting everything together, we obtain,

$$\Delta_{k_0} \leq \Delta_1 \exp\left(-\frac{\mu}{L}\frac{\alpha-\alpha^{k_0}}{1-\alpha}\right) + c_5 \exp\left(\frac{\alpha^{k_0}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\frac{k_0\ln(T/\beta)^2}{T^2}$$

Now let us consider the regime $k \geq k_0$ where $\alpha_k \leq \frac{1}{\nu}$, so that we have

$$\mathbb{E}\|w_{k+1}-w^*\|^2 \leq \left(1-\frac{\mu\alpha_k}{L}\right)\|w_k-w^*\|^2 + \frac{2\nu^2\sigma^2}{L}\alpha_k^2$$

Writing $\Delta_k = \mathbb{E}\|w_k-w^*\|^2$, and unrolling the recursion from $k = k_0$ to T,

$$\Delta_{T+1} \leq \Delta_{k_0}\prod_{k=k_0}^{T}(1-\frac{\mu}{L}\alpha_k) + \frac{2\nu^2\sigma^2}{L}\sum_{k=k_0}^{T}\alpha_k^2\prod_{i=k+1}^{T}(1-\frac{\mu}{L}\alpha_i)$$

Bounding the first term similar to [Lemma 5](#),

$$\prod_{k=k_0}^{T}(1-\frac{\mu}{L}\alpha_k) \leq \exp\left(-\frac{\mu}{L}\sum_{k=k_0}^{T}\alpha_k\right) = \exp\left(\frac{-\mu}{L}\frac{\alpha^{k_0}-\alpha^{T+1}}{1-\alpha}\right)$$

Bounding the second term similar to [Lemma 6](#),

$$\sum_{k=k_0}^{T}\alpha_k^2\prod_{i=k+1}^{T}(1-\frac{\mu}{L}\alpha_i) \leq \sum_{k=k_0}^{T}\alpha_k^2\exp\left(-\frac{\mu}{L}\sum_{i=k+1}^{T}\alpha^i\right)$$

$$= \sum_{k=k_0}^{T}\alpha_k^2\exp\left(-\frac{1}{\kappa}\frac{\alpha^{k+1}-\alpha^{T+1}}{1-\alpha}\right)$$

$$= \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right)\sum_{k=k_0}^{T}\alpha_k^2\exp\left(-\frac{\alpha^{k+1}}{\kappa(1-\alpha)}\right)$$

$$\leq \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right)\sum_{k=k_0}^{T}\alpha_k^2\left(\frac{2(1-\alpha)\kappa}{e\alpha^{k+1}}\right)^2$$

$$= \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4(1-\alpha)^2\kappa^2}{e^2\alpha^2}(T-k_0+1)$$

$$\leq \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{e^2\alpha^2}\frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}$$

Putting everything together,

$$\Delta_{T+1} \leq \Delta_{k_0} \exp\left(\frac{-\mu}{L} \frac{\alpha^{k_0} - \alpha^{T+1}}{1-\alpha}\right) + \frac{2\nu^2\sigma^2}{L} \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right) \frac{4\kappa^2}{e^2\alpha^2} \frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}$$

Combining the above bounds for the two regimes, we get,

$$\Delta_{T+1} \leq \exp\left(\frac{-\mu}{L} \frac{\alpha^{k_0} - \alpha^{T+1}}{1-\alpha}\right) \left(\Delta_1 \exp\left(-\frac{\mu}{L} \frac{\alpha - \alpha^{k_0}}{1-\alpha}\right) + c_5 \exp\left(\frac{\mu\alpha^{k_0}}{L(1-\alpha)}\right) \frac{4\kappa^2}{e^2\alpha^2} \frac{k_0 \ln(T/\beta)^2}{T^2}\right)$$

$$+ \frac{2\nu^2\sigma^2}{L} \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right) \frac{4\kappa^2}{e^2\alpha^2} \frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}$$

$$= \Delta_1 \exp\left(\frac{-\mu}{L} \frac{\alpha - \alpha^{T+1}}{1-\alpha}\right) + c_5 \exp\left(\frac{\mu}{L} \frac{\alpha^{T+1}}{1-\alpha}\right) \frac{4\kappa^2}{2e^2\alpha^2} \frac{k_0 \ln(T/\beta)^2}{T^2}$$

$$+ \frac{2\nu^2\sigma^2}{L} \exp\left(\frac{\alpha^{T+1}}{\kappa(1-\alpha)}\right) \frac{4\kappa^2}{e^2\alpha^2} \frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}$$

Using Lemma 5 to bound the first term, and noting that $\frac{\alpha^{T+1}}{1-\alpha} \leq \frac{2\beta}{\ln(T/\beta)}$

$$\Delta_{T+1} \leq \Delta_1 c_2 \exp\left(-\frac{T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right) + c_5 \frac{4c_2\kappa^2}{e^2\alpha^2} \frac{k_0 \ln(T/\beta)^2}{T^2} + \frac{2\nu^2\sigma^2}{L} \frac{4c_2\kappa^2}{e^2\alpha^2} \frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}$$

where $\kappa = \frac{L}{\mu}$ and $c_2 = \exp\left(\frac{1}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right)$.

Putting in the value of c_5 and k_0, and rearranging, we get

$$\Delta_{T+1} \leq \Delta_1 c_2 \exp\left(-\frac{T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right) + \frac{4\nu^2\sigma^2}{LT} \frac{4c_2\kappa^2 \ln(T/\beta)^2}{e^2\alpha^2}$$

$$+ \left[\max_{j\in[k_0]}\{f(w_j) - f^*\} \frac{2\nu(\nu-1)}{L}\right] \frac{4c_2\kappa^2}{\nu^2 e^2\alpha^2} \frac{[\ln(\nu)]_+ \ln(T/\beta)}{T}$$

Combining the statements from $\nu \leq 1$ and $\nu > 1$ gives us the theorem statement.

□

C.4. Proof of Theorem 8

Theorem 8. Assuming (i) convexity and (ii) L_i-smoothness of each f_i, SGD with step-size $\eta_k = \frac{1}{2L}\alpha_k$ has the following convergence rate,

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\|w_1 - w^*\|^2}{\sum_{k=1}^{T}\alpha_k} + \sigma^2 \frac{\sum_{k=1}^{T}\alpha_k^2}{\sum_{k=1}^{T}\alpha_k} \tag{9}$$

where $\bar{w}_{T+1} = \frac{\sum_{k=1}^{T}\alpha_k w_k}{\sum_{k=1}^{T}\alpha_k}$. For $\alpha_k = \left[\frac{\beta}{T}\right]^{k/T}$, the convergence rate is given by,

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\ln(T/\beta)\|w_1 - w^*\|^2}{\alpha T - 2\beta} + \sigma^2 \frac{T}{T-\beta}$$

Proof. Following the proof of Theorem 1,

$$\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\gamma_k\alpha_k\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + 2L\gamma_k^2\alpha_k^2\left[f_{ik}(w_k) - f_{ik}(w^*)\right]$$
$$+ \frac{2}{L}\gamma_k^2\alpha_k^2\left[f_{ik}(w_k) - f_{ik}^*\right]$$

$$\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - \frac{\alpha_k}{L}\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + \frac{\alpha_k^2}{2L}\left[f_{ik}(w_k) - f_{ik}(w^*)\right] + \frac{\alpha_k^2}{2L}\left[f_{ik}(w_k) - f_{ik}^*\right]$$
$$(\gamma_k = \tfrac{1}{2L} \text{ for all } k.)$$

$$\leq \|w_k - w^*\|^2 - \frac{\alpha_k}{L}\left[f_{ik}(w_k) - f_{ik}(w^*)\right] + \frac{\alpha_k^2}{2L}\left[f_{ik}(w_k) - f_{ik}(w^*)\right] + \frac{\alpha_k^2}{2L}\left[f_{ik}(w_k) - f_{ik}^*\right]$$
$$(\text{By convexity})$$

Taking expectation,

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

$$\leq \|w_k - w^*\|^2 - \frac{\alpha_k}{2L}\left[f(w_k) - f(w^*)\right] + \frac{\alpha_k^2}{2L}\sigma^2 \qquad (\text{Since } f(w_k) - f(w^*) \geq 0 \text{ and } \alpha_k \leq 1)$$

Rearranging and summing from $k = 1$ to T,

$$\sum_{k=1}^{T}\alpha_k[f(w_k) - f(w^*)] \leq 2L\|w_1 - w^*\|^2 + \sigma^2\sum_{k=1}^{T}\alpha_k^2$$

By averaging and using Jensen. Denote $\bar{w}_{T+1} = \frac{\sum_{k=1}^{T}\alpha_k w_k}{\sum_{k=1}^{T}\alpha_k}$,

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\|w_1 - w^*\|^2}{\sum_{k=1}^{T}\alpha_k} + \sigma^2\frac{\sum_{k=1}^{T}\alpha_k^2}{\sum_{k=1}^{T}\alpha_k}$$

Next, we bound $\sum_{k=1}^{T}\alpha_k$ and $\sum_{k=1}^{T}\alpha_k^2$ for the exponentially-decreasing α_k sequence, when $\alpha_k = \left[\frac{\beta}{T}\right]^{k/T}$. From Lemma 5, we know that,

$$\sum_{k=1}^{T}\alpha_k \geq \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}.$$

Bounding the ratio $\frac{\sum_{k=1}^{T} \alpha_k^2}{\sum_{k=1}^{T} \alpha_k} = \frac{\sum_{k=1}^{T} \alpha^{2k}}{\sum_{k=1}^{T} \alpha^k}$ where $\alpha = \left[\frac{\beta}{T}\right]^{1/T}$,

$$\frac{\sum_{k=1}^{T} \alpha^{2k}}{\sum_{k=1}^{T} \alpha^k} \leq \frac{\alpha^2}{1 - \alpha^2} \frac{1 - \alpha}{\alpha - \alpha^{T+1}}$$

$$= \frac{\alpha}{1 + \alpha} \frac{1}{1 - \alpha^T} \leq \frac{1}{1 - \alpha^T} = \frac{T}{T - \beta}$$

Putting everything together,

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L \ln(T/\beta) \|w_1 - w^*\|^2}{\alpha T - 2\beta} + \sigma^2 \frac{T}{T - \beta}$$

$\square$

C.5. Additional lemmas for upper-bound proofs

Lemma 5.

$$A := \sum_{t=1}^{T} \alpha^t \geq \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}$$

Proof.

$$\sum_{t=1}^{T} \alpha^t = \frac{\alpha - \alpha^{T+1}}{1 - \alpha} = \frac{\alpha}{1 - \alpha} - \frac{\alpha^{T+1}}{1 - \alpha}$$

We have

$$\frac{\alpha^{T+1}}{1 - \alpha} = \frac{\alpha\beta}{T(1 - \alpha)} = \frac{\beta}{T} \cdot \frac{1}{1/\alpha - 1} \leq \frac{\beta}{T} \cdot \frac{2}{\ln(1/\alpha)} = \frac{\beta}{T} \cdot \frac{2}{\frac{1}{T}\ln(T/\beta)} = \frac{2\beta}{\ln(T/\beta)} \tag{10}$$

where in the inequality we used Lemma 16 and the fact that $1/\alpha > 1$. Plugging back into A we get,

$$A \geq \frac{\alpha}{1 - \alpha} - \frac{2\beta}{\ln(T/\beta)}$$

$$\geq \frac{\alpha}{\ln(1/\alpha)} - \frac{2\beta}{\ln(T/\beta)} \qquad (1 - x \leq \ln(\tfrac{1}{x}))$$

$$= \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}$$

$\square$

Lemma 6. *For* $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$ *and any* $\kappa > 0$,

$$\sum_{k=1}^{T} \alpha^{2k} \exp\left(-\frac{1}{\kappa} \sum_{i=k+1}^{T} \alpha^i\right) \leq \frac{4\kappa^2 c_2 (\ln(T/\beta))^2}{e^2 \alpha^2 T}$$

where $c_2 = \exp\left(\frac{1}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right)$

Proof. First observe that,

$$\sum_{i=k+1}^{T} \alpha^i = \frac{\alpha^{k+1} - \alpha^{T+1}}{1 - \alpha}$$

We have

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha\beta}{T(1-\alpha)} = \frac{\beta}{T} \cdot \frac{1}{1/\alpha - 1} \leq \frac{\beta}{T} \cdot \frac{2}{\ln(1/\alpha)} = \frac{\beta}{T} \cdot \frac{2}{\frac{1}{T}\ln(T/\beta)} = \frac{2\beta}{\ln(T/\beta)}$$

where in the inequality we used 16 and the fact that $1/\alpha > 1$. These relations imply that,

$$\sum_{i=k+1}^{T} \alpha^i \geq \frac{\alpha^{k+1}}{1-\alpha} - \frac{2\beta}{\ln(T/\beta)}$$

$$\implies \exp\left(-\frac{1}{\kappa}\sum_{i=k+1}^{T} \alpha^i\right) \leq \exp\left(-\frac{1}{\kappa}\frac{\alpha^{k+1}}{1-\alpha} + \frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right) = c_2 \exp\left(-\frac{1}{\kappa}\frac{\alpha^{k+1}}{1-\alpha}\right)$$

We then have

$$\sum_{k=1}^{T} \alpha^{2k} \exp\left(-\frac{1}{\kappa}\sum_{i=k+1}^{T} \alpha^i\right) \leq c_2 \sum_{k=1}^{T} \alpha^{2k} \exp\left(-\frac{1}{\kappa}\frac{\alpha^{k+1}}{1-\alpha}\right)$$

$$\leq c_2 \sum_{k=1}^{T} \alpha^{2k} \left(\frac{2(1-\alpha)\kappa}{e\alpha^{k+1}}\right)^2 \qquad \text{(Lemma 17)}$$

$$= \frac{4\kappa^2 c_2}{e^2\alpha^2} T(1-\alpha)^2$$

$$\leq \frac{4\kappa^2 c_2}{e^2\alpha^2} T(\ln(1/\alpha))^2$$

$$= \frac{4\kappa^2 c_2 (\ln(T/\beta))^2}{e^2\alpha^2 T}$$

$\square$

Lemma 7. *For* $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$ *and any* $\kappa > 0$,

$$\sum_{k=1}^{T} \alpha^k \exp\left(-\frac{1}{\kappa}\sum_{i=k+1}^{T} \alpha^i\right) \leq c_2 \frac{\kappa \ln(T/\beta)}{e\alpha}$$

for $c_2 = \exp\left(\frac{1}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$

Proof. Proceeding in the same way as Lemma 6, we obtain the following inequality,

$$\sum_{k=1}^{T} \alpha^k \exp\left(-\frac{1}{\kappa}\sum_{i=k+1}^{T} \alpha^i\right) \leq c_2 \sum_{k=1}^{T} \alpha^k \exp\left(-\frac{1}{\kappa}\frac{\alpha^{k+1}}{1-\alpha}\right)$$

Further bounding this term,

$$\sum_{k=1}^{T} \alpha^k \exp\left(-\frac{1}{\kappa}\sum_{i=t+1}^{T} \alpha^i\right) \leq c_2 \sum_{k=1}^{T} \alpha^k \frac{(1-\alpha)\kappa}{e\alpha^{k+1}} \qquad \text{(Lemma 17)}$$

$$\leq c_2 (1-\alpha) \frac{\kappa T}{e\alpha}$$

$$\leq c_2 \ln(1/\alpha) \frac{\kappa T}{e\alpha}$$

$$= c_2 \frac{\kappa \ln(T/\beta)}{e\alpha}$$

$$\square$$

Lemma 8. *If f_i is L_i-smooth, stochastic lines-searches ensures that*

$$\gamma \left\| \nabla f_i(w) \right\|^2 \leq \frac{1}{c}(f_i(w) - f_i^*), \qquad \text{and} \qquad \min\left\{ \gamma_{\max}, \frac{2(1-c)}{L_i} \right\} \leq \gamma \leq \gamma_{\max}.$$

Moreover, if f_i is a one-dimensional quadratic,

$$\gamma = \min\left\{ \gamma_{\max}, \frac{2(1-c)}{L_i} \right\}$$

Proof. Recall that if f_i is L_i-smooth, then for an arbitrary direction d,

$$f_i(w - d) \leq f_i(w) - \langle \nabla f_i(w), d \rangle + \frac{L_i}{2} \left\| d \right\|^2.$$

For the stochastic line-search, $d = \gamma \nabla f_i(w)$. The smoothness and the line-search condition are then

$$\text{Smoothness: } f_i(w - \gamma \nabla f_i(w)) - f_i(w) \leq \left(\frac{L_i}{2} \gamma^2 - \gamma \right) \left\| \nabla f_i(w) \right\|^2,$$

$$\text{Line-search: } f_i(w - \gamma \nabla f_i(w)) - f_i(w) \leq -c\gamma \left\| \nabla f_i(w) \right\|^2.$$

The line-search condition is looser than smoothness if

$$\left(\tfrac{L_i}{2} \gamma^2 - \gamma \right) \left\| \nabla f_i(w) \right\|^2 \leq -c\gamma \left\| \nabla f_i(w) \right\|^2.$$

The inequality is satisfied for any $\gamma \in [a, b]$, where a, b are values of γ that satisfy the equation with equality, $a = 0, b = 2(1-c)/L_i$, and the line-search condition holds for $\gamma \leq 2(1-c)/L_i$. As the line-search selects the largest feasible step-size, $\gamma \geq 2(1-c)/L_i$. If the step-size is capped at $\gamma_{\max}$, we have $\eta \geq \min\{\gamma_{\max}, 2(1-c)/L_i\}$, and the proof for the stochastic line-search is complete.

From the previous discussion, observe that if $\gamma > \frac{2(1-c)}{L_i}$, then we have

$$\left(\tfrac{L_i}{2} \gamma^2 - \gamma \right) \left\| \nabla f_i(w) \right\|^2 > -c\gamma \left\| \nabla f_i(w) \right\|^2.$$

If f is a one-dimensional quadratic, the smoothness inequality is actually an equality, and thus

$$f_i(w - \gamma \nabla f_i(w)) - f_i(w) = \left(\frac{L_i}{2} \gamma^2 - \gamma \right) \left\| \nabla f_i(w) \right\|^2$$

So if $\gamma > \frac{2(1-c)}{L_i}$,

$$f_i(w - \gamma \nabla f_i(w)) - f_i(w) \geq -c\gamma \left\| \nabla f_i(w) \right\|^2$$

and the line-search condition does not hold. This implies that for one-dimensional quadratics $\gamma = \min\{\gamma_{\max}, \frac{2(1-c)}{L_i}\}$ $\square$

D. Lower-bound proofs for Section 3

D.1. Proof of Theorem 3

Theorem 3. When using T iterations of SGD to minimize the sum $f(w) = \frac{f_1(w) + f_2(w)}{2}$ of two one-dimensional quadratics, $f_1(w) = \frac{1}{2}(w-1)^2$ and $f_2(w) = \frac{1}{2}\left(2w + \frac{1}{2}\right)^2$, setting γ_k using SLS with $\gamma_{\max} \geq 1$ and $c \geq \frac{1}{2}$, any convergent sequence of α_k results in convergence to a neighbourhood of the solution. Specifically, if w^* is the minimizer of f and $w_1 > 0$, then,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right).$$

Proof. For SLS with a general $c \geq \frac{1}{2}$ on quadratics, we know that $\gamma_k = \frac{2(1-c)}{L_{i_k}}$ (see Lemma 8 for a formal proof). Recall that we consider two one-dimensional quadratics $f_i(w) = \frac{1}{2}(wx_i - y_i)^2$ for $i \in \{1, 2\}$ such that $x_1 = 1$, $y_1 = 1$, $x_2 = 2$, $y_2 = -\frac{1}{2}$. Specifically,

$$f_1(w) = \frac{1}{2}(w-1)^2 \Rightarrow L_1 = 1$$

$$f_2(w) = \frac{1}{2}(2w + \frac{1}{2})^2 \Rightarrow L_2 = 4$$

$$f(w) = \frac{1}{4}(w-1)^2 + \frac{1}{4}(2w + \frac{1}{2})^2 = \frac{5}{4}w^2 + \frac{1}{4} + \frac{1}{16} \Rightarrow w^* = 0$$

If $i_k = 1$,

$$w_{k+1} = w_k - \alpha_k 2(1-c)(w_k - 1) = 2(1-c)\alpha_k + (1 - 2(1-c)\alpha_k)w_k$$

If $i_k = 2$,

$$w_{k+1} = w_k - 2(1-c)\alpha_k \frac{2}{4}(2w_k + \frac{1}{2}) = (1 - 2(1-c)\alpha_k)w_k - \frac{1}{4}2(1-c)\alpha_k$$

Then

$$\mathbb{E}w_{k+1} = (1 - 2(1-c)\alpha_k)w_k + \frac{1}{2}2(1-c)\alpha_k - \frac{1}{8}2(1-c)\alpha_k = (1 - 2(1-c)\alpha_k)w_k + \frac{3}{8}2(1-c)\alpha_k$$

and

$$\mathbb{E}w_T = \mathbb{E}(w_T - w^*) = (w_1 - w^*)\prod_{k=1}^{T}(1 - 2(1-c)\alpha_k) + \frac{3}{8}\sum_{k=1}^{T}2(1-c)\alpha_k \prod_{i=k+1}^{T}(1 - 2(1-c)\alpha_i)$$

Using Lemma 18 and the fact that $2(1-c)\alpha_k \leq 1$ for all k, we have that if $w_1 - w^* = w_1 > 0$,

$$\mathbb{E}(w_T - w^*) \geq \min\left(w_1, \frac{3}{8}\right)$$

$\square$

D.2. Proof of Theorem 5

Theorem 5. When minimizing a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, GD with $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$, satisfies

$$w_{k+1} - w^* = (w_1 - w^*) \prod_{i=1}^{k}(1 - \nu\alpha_i).$$

After $k' := \frac{T}{\ln(T/\beta)} \ln\left(\frac{\nu}{3}\right)$ iterations, we have that

$$|w_{k'+1} - w^*| \geq 2^{k'}|w_1 - w^*|.$$

Proof. One has $w^* = \frac{y}{x}$ and $L = x^2$. Therefore

$$
\begin{aligned}
w_{k+1} - w^* &= w_k - w^* - \alpha_k\eta_k x\,(xw_k - y)\\
&= w_k - w^* - \alpha_k\frac{\nu}{L}Lw_k + \alpha_k\frac{\nu}{L}xy\\
&= w_k - w^* - \alpha_k\nu w_k + \alpha_k\rho w^* = (1 - \nu\alpha_k)(w_k - w^*)
\end{aligned}
$$

Iterating gives the first part of the result. Now, for $k \leq k'$, we have

$$1 - \nu\alpha^k \leq 1 - \nu\alpha^{k'} \leq 1 - \nu\alpha^{\frac{T}{\ln(T/\beta)}(\ln\nu - \ln 3)} = 1 - \nu\left(\frac{\beta}{T}\right)^{\frac{1}{\ln(T/\beta)}(\ln\nu - \ln 3)} = 1 - \nu\left(\frac{3}{\nu}\right) = -2$$

and thus

$$|w_{k'+1} - w^*| = |w_1 - w^*|\prod_{i=1}^{k'}|1 - \nu\alpha_k| \geq |w_1 - w^*|2^{k'}$$

$\square$

D.3. Lemmas for convex setting

Lemma 9. *The polynomial stepsize defined as $\alpha_k = (1/k)^\delta$ for some $0 \leq \delta \leq 1$ cannot satisfy $\sum_{k=1}^{T}\alpha_k \geq C_1 T$ and $\sum_{k=1}^{T}\alpha_k^2 \leq C_2\sqrt{T}$ for positive constants C_1 and C_2.*

Proof. If $\delta = 0$, $\alpha_k = 1$ for all k, and then $\sum_{k=1}^{T}\alpha_k^2 = T$. If $\delta = 1$, then $\sum_{k=1}^{T}\alpha_k = \Theta(\ln T)$. If $0 < \delta < 1$, basic calculus shows that

$$\int_{1}^{T+1}\frac{1}{x^\delta} \leq \sum_{k=1}^{T}\frac{1}{k^\delta} \leq 1 + \int_{1}^{T}\frac{1}{x^\delta}$$

and thus

$$\frac{1}{1-\delta}\left((T+1)^{1-\delta} - 1\right) \leq \sum_{k=1}^{T}\frac{1}{k^\delta} \leq 1 + \frac{1}{1-\delta}\left(T^{1-\delta} - 1\right)$$

which shows that $\sum_{k=1}^{T}\alpha_k = \Theta(T^{1-\delta})$, and thus we cannot have $\sum_{k=1}^{T}\alpha_k \geq C_1 T$ for all T. $\square$

Lemma 10. *The exponential stepsize defined as $\alpha_k = \alpha^k$ for some $\alpha < 1$ cannot satisfy $\sum_{k=1}^{T} \alpha_k \geq C_1 T$ and $\sum_{k=1}^{T} \alpha_k^2 \leq C_2 \sqrt{T}$ for positive constants C_1 and C_2.*

Proof. Suppose by contradiction that the exponential stepsize satisfies the two conditions. Then

$$C_2\sqrt{T} \geq \sum_{k=1}^{T} \alpha_k^2 = \sum_{k=1}^{T} \alpha^{2k} = \sum_{k=1}^{2T} \alpha^k - \sum_{k=1}^{T} \alpha^{2k-1} = \sum_{k=1}^{2T} \alpha^k - \frac{1}{\alpha}\sum_{k=1}^{T} \alpha^{2k}$$

By assumption, $\sum_{k=1}^{2T} \alpha^k \geq C_1 2T$ and $\sum_{k=1}^{T} \alpha^{2k} \leq C_2\sqrt{T}$. Therefore

$$\sum_{k=1}^{2T} \alpha^k - \frac{1}{\alpha}\sum_{k=1}^{T} \alpha^{2k} \geq 2C_1 T - \frac{1}{\alpha}C_2\sqrt{T}$$

But then we obtain

$$C_2\sqrt{T} \geq 2C_1 T - \frac{1}{\alpha}C_2\sqrt{T}$$

which is a contradiction by taking T to infinity. $\qquad\square$

E. Proofs for Section 4

E.1. Reformulation

Let us consider a general ASGD update whose parameters satisfy the following conditions.

$$r_k^2 = (1 - r_k)r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k\mu\eta_k. \tag{11}$$

$$b_k = \frac{(1 - r_{k-1})r_{k-1}\frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2\frac{\eta_k}{\eta_{k-1}}}, \tag{12}$$

It can be verified that setting $\eta_k = \gamma_k\alpha_k = \frac{1}{\rho L}\left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ satisfies Eq. (11). We first show that the update in Eq. (4)-Eq. (5) satisfying the conditions in Eq. (12) and Eq. (11) can be written in an equivalent form more amenable to the analysis.

Lemma 11. *The following update:*

$$y_k = w_k - \frac{r_k q_k}{q_k + r_k\mu}(w_k - z_k) \tag{13}$$

$$w_{k+1} = y_k - \eta_k \nabla f_{ik}(y_k) \tag{14}$$

$$z_{k+1} = w_k + \frac{1}{r_k}[w_{k+1} - w_k] \tag{15}$$

where,

$$q_{k+1} = (1 - r_k)q_k + r_k\mu \tag{16}$$

$$r_k^2 = q_{k+1}\eta_k \tag{17}$$

$$z_{k+1} = \frac{1}{q_{k+1}}\left[(1 - r_k)q_k z_k + r_k\mu y_k - r_k\nabla f_{ik}(y_k)\right] \tag{18}$$

is equivalent to the update in Eq. (4)-Eq. (5).

Proof.

First we check the consistency of the update (Eq. (15)) and definition (Eq. (18)) of z_k. Using Eq. (18),

$$z_{k+1} = \frac{1}{q_{k+1}}\left[(1 - r_k)q_k z_k + r_k\mu y_k - r_k\nabla f_{ik}(y_k)\right]$$

$$= -\frac{(1 - r_k)}{r_k}w_k - \frac{r_k}{q_{k+1}}\nabla f_{ik}(y_k) + y_k\left[\frac{(1 - r_k)(q_k + r_k\mu)}{q_{k+1}r_k} + \frac{r_k\mu}{q_{k+1}}\right]$$

$$= -\frac{(1 - r_k)}{r_k}w_k - \frac{r_k}{q_{k+1}}\nabla f_{ik}(y_k) + y_k\left[\frac{q_k(1 - r_k) + (r_k\mu - r_k^2\mu)}{q_{k+1}r_k} + \frac{r_k^2\mu}{q_{k+1}r_k}\right]$$

$$= -\frac{(1 - r_k)}{r_k}w_k - \frac{r_k}{q_{k+1}}\nabla f_{ik}(y_k) + y_k\left[\frac{(q_{k+1} - r_k\mu) + (r_k\mu - r_k^2\mu) + r_k^2\mu}{q_{k+1}r_k}\right] \quad \text{(From Eq. (16))}$$

$$= w_k - \frac{w_k}{r_k} + \frac{1}{r_k}[y_k - \eta_k\nabla f_{ik}(y_k)] \quad \text{(From Eq. (17))}$$

$$z_{k+1} = w_k + \frac{1}{r_k}[w_{k+1} - w_k] \quad \text{(From Eq. (14))}$$

which recovers Eq. (15) showing that the definition of z_k and its update is consistent.

Now we check the equivalence of Eq. (11) and Eq. (16)-Eq. (17). Eliminating q_k using Eq. (16)-Eq. (17),

$$\frac{r_k^2}{\eta_k} = (1 - r_k)\frac{r_{k-1}^2}{\eta_{k-1}} + r_k\mu$$

Multiplying by η_k recovers Eq. (11).

Since Eq. (5) and Eq. (14) are equivalent, we need to establish the equivalence of Eq. (4) and the updates in Eq. (13)-Eq. (15). From Eq. (15)

$$z_k = w_{k-1} + \frac{1}{r_{k-1}}[w_k - w_{k-1}] \implies z_k - w_k = \frac{1 - r_{k-1}}{r_{k-1}}(w_k - w_{k-1})$$

Starting from Eq. (13) and using the above relation to eliminate z_k,

$$y_k = w_k + \frac{r_k q_k}{q_k + r_k\mu}\frac{1 - r_{k-1}}{r_{k-1}}[w_k - w_{k-1}]$$

which is in the same form as Eq. (4). We now eliminate q_k from $\frac{r_k q_k}{q_k + r_k\mu}\frac{1-r_{k-1}}{r_{k-1}}$. From Eq. (16) and Eq. (17),

$$\frac{r_k^2}{\eta_k} = (1 - r_k)q_k + r_k\mu \implies q_k + r_k\mu = \frac{r_k^2}{\eta_k} + r_k q_k$$

Using this relation,

$$\frac{r_k q_k}{q_k + r_k\mu}\frac{1 - r_{k-1}}{r_{k-1}} = \frac{q_k\eta_k}{r_k + q_k\eta_k}\frac{1 - r_{k-1}}{r_{k-1}}$$

Using Eq. (17), observe that $\eta_k q_k = \frac{\eta_k}{\eta_{k-1}}\eta_{k-1}q_k = \frac{\eta_k}{\eta_{k-1}}r_{k-1}^2$. Using this relation,

$$\frac{r_k q_k}{q_k + r_k\mu}\frac{1 - r_{k-1}}{r_{k-1}} = \frac{\frac{\eta_k}{\eta_{k-1}}r_{k-1}^2}{r_k + \frac{\eta_k}{\eta_{k-1}}r_{k-1}^2}\frac{1 - r_{k-1}}{r_{k-1}} = \frac{(1 - r_{k-1})r_{k-1}}{r_k\frac{\eta_{k-1}}{\eta_k} + r_{k-1}^2} = b_k$$

which establishes the equivalence to Eq. (4) and completes the proof.

$$\square$$

E.2. Estimating sequences

Similar to (Nesterov, 2004; Mishkin, 2020), we will use the estimating sequence $\{\phi_k, \lambda_k\}_{k=1}^\infty$ such that $\lambda_k \in (0, 1)$ and

$$\lambda_1 = 1 \quad ; \quad \lambda_{k+1} = (1 - r_k)\lambda_k, \tag{19}$$

$$\phi_k(w) = [\inf_w \phi_k(w)] + \frac{q_k}{2}\|w - z_k\|^2, \tag{20}$$

and satisfies the following update condition

$$\phi_k(w) \le (1 - \lambda_k)f(w) + \lambda_k\phi_1(w) \tag{21}$$

The above definitions impose the following update for $\phi_k^* := [\inf_w \phi_k(w)]$,

$$\phi_{k+1}^* = (1 - r_k)\phi_k^* + r_k \left[f(y_k) - \frac{r_k}{2q_{k+1}} \|\nabla f(y_k)\|^2 + \frac{(1 - r_k) q_k}{q_{k+1}} \left(\frac{\mu}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) \right] \quad (22)$$

Finally note that the definition of ϕ_k can be used to rewrite Eq. (13) as

$$y_k = w_k - \frac{r_k}{q_k + r_k \mu} \nabla \phi_k(w_k). \quad (23)$$

E.3. Proof of Theorem 6

Given the definitions in Appendix E.2, we first prove the descent lemma for $\eta_k = \frac{1}{\rho L} \alpha_k$, where $\alpha_k \leq 1$ is the exponentially decreasing step-size.

Lemma 12. *Using the update in Eq. (14) with $\eta_k = \frac{1}{\rho L} \alpha_k$, we obtain the following descent condition.*

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2} \|\nabla f(y_k)\|^2 + \frac{1}{2\rho^2 L} \alpha_k^2 \sigma^2$$

Proof. By smoothness, and the update in Eq. (14),

$$f(w_{k+1}) \leq f(y_k) - \eta_k \langle \nabla f(y_k), \nabla f_{i_k}(y_k) \rangle + \frac{L}{2} \eta_k^2 \|\nabla f_{i_k}(y_k)\|^2$$

Taking expectation w.r.t. i_k,

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \eta_k \|\nabla f(y_k)\|^2 + \frac{L}{2} \eta_k^2 \mathbb{E}[\|\nabla f_{i_k}(y_k)\|^2] \qquad (\eta_k \text{ is independent of } i_k.)$$

$$\leq \mathbb{E}[f(y_k)] - \eta_k \|\nabla f(y_k)\|^2 + \frac{\rho L}{2} \eta_k^2 \mathbb{E}[\|\nabla f(y_k)\|^2] + \frac{L}{2} \eta_k^2 \sigma^2 \qquad \text{(From Eq. (6))}$$

$$= \mathbb{E}[f(y_k)] - \eta_k \|\nabla f(y_k)\|^2 + \frac{\eta_k \alpha_k}{2} \mathbb{E}[\|\nabla f(y_k)\|^2] + \frac{1}{2\rho^2 L} \alpha_k^2 \sigma^2$$

$$\leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2} \|\nabla f(y_k)\|^2 + \frac{1}{2\rho^2 L} \alpha_k^2 \sigma^2$$

$\square$

The main part of the proof is to show that ϕ_k^* is an upper-bound on $f(w_k)$ (upto a factor governed by the noise term $\mathcal{N}_k$ depending on σ^2) for all k and is proved in the following lemma.

Lemma 13. *For the estimating sequences defined in Appendix E.2 and the updates in Eq. (13)-Eq. (18), for all k,*

$$\mathbb{E}[\phi_k^*] := \mathbb{E}[\inf_w \phi_k(w)] \geq \mathbb{E}[f(w_k)] - \mathcal{N}_k$$

where $\mathcal{N}_k := \frac{\sigma^2}{2\rho^2 L} \sum_{j=1}^k \alpha_j^2 \prod_{i=j+1}^k (1 - r_i)$.

Proof. We will prove the lemma by induction. For $k = 1$, we define $\phi_1^* = f(w_1)$, and since $\mathcal{N}_k \geq 0$ for all k, $\mathbb{E}[\phi_1^*] \geq f(w_1) - \mathcal{N}_1$, thus satisfying the base-case for the induction. For the induction, we will use the fact that $\mathcal{N}_{k+1} = (1 - r_k)\mathcal{N}_k + \frac{2\sigma^2}{\rho^2 L} \alpha_k^2$.

Assuming the induction hypothesis, $\mathbb{E}[\phi_k^*] \geq \mathbb{E}[f(w_k)] - \mathcal{N}_k$, we use Eq. (22) to prove the statement for $k + 1$ as follows. Taking expectations w.r.t to the randomness in $j = 1$ to k,

$$\mathbb{E}[\phi_{k+1}^*] = (1 - r_k)\mathbb{E}[\phi_k^*] + r_k \mathbb{E} \left[f(y_k) - \frac{r_k}{2q_{k+1}} \|\nabla f(y_k)\|^2 + \frac{(1 - r_k) q_k}{q_{k+1}} \left(\frac{\mu}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) \right]$$

$$\geq (1 - r_k)\mathbb{E}[f(w_k) - \mathcal{N}_k]$$
$$+ r_k\mathbb{E}\left[f(y_k) - \frac{r_k}{2q_{k+1}}\|\nabla f(y_k)\|^2 + \frac{(1 - r_k)q_k}{q_{k+1}}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)\right]$$

$$\text{(by the induction hypothesis)}$$

$$= (1 - r_k)\mathbb{E}[f(w_k)] + r_k\mathbb{E}[f(y_k)] - \frac{r_k^2}{2q_{k+1}}\mathbb{E}\|\nabla f(y_k)\|^2$$
$$+ \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right) - (1 - r_k)\mathcal{N}_k$$
$$= (1 - r_k)\mathbb{E}[f(w_k)] + r_k\mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\mathbb{E}\|\nabla f(y_k)\|^2$$
$$+ \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right) - (1 - r_k)\mathcal{N}_k \qquad \text{(Using Eq. (17))}$$

By convexity, $f(w_k) \geq f(y_k) + \langle\nabla f(y_k), w_k - y_k\rangle$,

$$\geq (1 - r_k)\mathbb{E}[f(y_k) + \langle\nabla f(y_k), w_k - y_k\rangle] + r_k\mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\mathbb{E}\|\nabla f(y_k)\|^2$$
$$+ \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right) - (1 - r_k)\mathcal{N}_k$$
$$= \mathbb{E}\left[f(y_k) - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2\right] + \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)$$
$$+ (1 - r_k)\mathbb{E}[\langle\nabla f(y_k), w_k - y_k\rangle] - (1 - r_k)\mathcal{N}_k$$

By Lemma 12,

$$\geq \mathbb{E}\left[f(w_{k+1}) - \frac{1}{2\rho^2 L}\alpha_k^2\sigma^2\right] + \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)$$
$$+ (1 - r_k)\mathbb{E}[\langle\nabla f(y_k), w_k - y_k\rangle] - (1 - r_k)\mathcal{N}_k$$
$$= \mathbb{E}[f(w_{k+1})] + \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right) + (1 - r_k)\mathbb{E}[\langle\nabla f(y_k), w_k - y_k\rangle]$$
$$- \left[(1 - r_k)\mathcal{N}_k + \frac{1}{2\rho^2 L}\alpha_k^2\sigma^2\right]$$

Since $\mathcal{N}_{k+1} = \left[(1 - r_k)\mathcal{N}_k + \frac{1}{2\rho^2 L}\alpha_k^2\sigma^2\right]$,

$$\mathbb{E}[\phi_{k+1}^*] \geq \mathbb{E}[f(w_{k+1})] - \mathcal{N}_{k+1} + (1 - r_k)\mathbb{E}[\langle\nabla f(y_k), w_k - y_k\rangle]$$
$$+ \frac{r_k(1 - r_k)q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)$$

Now we show that $(1 - r_k)\mathbb{E}\left[\langle\nabla f(y_k), w_k - y_k\rangle + \frac{r_k q_k}{q_{k+1}}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)\right] \geq 0$. For this, we use Eq. (13)

$$y_k = w_k - \frac{q_k r_k}{q_k + r_k\mu}(w_k - z_k)$$
$$\implies z_k - y_k = z_k - w_k + \frac{q_k r_k}{q_k + r_k\mu}(w_k - z_k) = \left(1 - \frac{q_k r_k}{q_k + r_k\mu}\right)(z_k - w_k)$$
$$= \left(\frac{q_k(1 - r_k) + r_k\mu}{q_k + r_k\mu}\right)(z_k - w_k) = \left(\frac{q_{k+1}}{q_k + r_k\mu}\right)(z_k - w_k) \qquad \text{(By Eq. (16))}$$

$$\implies \frac{r_k q_k}{q_{k+1}} \langle \nabla f(y_k), z_k - y_k \rangle = \left\langle \nabla f(y_k), \left(-\frac{r_k\, q_k}{q_k + r_k\mu}\right)(w_k - z_k) \right\rangle = \langle \nabla f(y_k), y_k - w_k \rangle$$

Using this relation to simplify,

$$(1 - r_k)\mathbb{E}\left[\langle \nabla f(y_k), w_k - y_k \rangle + \frac{r_k\, q_k}{q_{k+1}}\left(\frac{\mu}{2}\|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle\right)\right]$$

$$= (1 - r_k)\mathbb{E}\left[\frac{r_k\, q_k}{q_{k+1}}\frac{\mu}{2}\|y_k - z_k\|^2 + (1 - r_k)\left[\langle \nabla f(y_k), w_k - y_k \rangle + \langle \nabla f(y_k), y_k - w_k \rangle\right]\right]$$

$$= (1 - r_k)\mathbb{E}\left[\frac{r_k\, q_k}{q_{k+1}}\frac{\mu}{2}\|y_k - z_k\|^2\right] \geq 0 \qquad\qquad \text{(Since } r_k \leq 1.)$$

Putting everything together,

$$\mathbb{E}[\phi_{k+1}^*] \geq \mathbb{E}\left[f(w_{k+1})\right] - \mathcal{N}_{k+1}$$

and we conclude that $\mathbb{E}[\phi_k^*] \geq \mathbb{E}\left[f(w_k)\right] - \mathcal{N}_k$ for all k by induction. $\qquad\square$

We now use the above lemma to prove Theorem 6.

Theorem 6. Under the same assumptions of Theorem 1 and (iii) the growth condition in Eq. (6), ASGD (Eqs. (4) and (5)) with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1-r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}$ converges as,

$$\Delta_{T+1} \leq 2c_3 \exp\left(-\frac{T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\Delta_1$$
$$+ \frac{2\sigma^2 c_3}{\rho\mu e^2}\frac{(\ln(T/\beta))^2}{\alpha^2 T},$$

where $\Delta_k := \mathbb{E}[f(w_k) - f^*]$ and $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$.

Proof. Using the reformulation in Lemma 11 gives us $q_k = \mu$ for all k and $z_1 = w_1$. For the estimating sequences defined in Appendix E.2, using Lemma 13, we know that the (reformulated) updates satisfy the following relation,

$$\mathbb{E}[f(w_{T+1})] \leq \mathbb{E}[\phi_{T+1}^*] + \mathcal{N}_{T+1} \leq \mathbb{E}[\phi_{T+1}(w^*)] + \mathcal{N}_{T+1}$$

From Eq. (21), we know that for all w and k,

$$\phi_k(w) \leq (1 - \lambda_k)f(w) + \lambda_k\phi_1(w)$$

Using these relations,

$$\mathbb{E}[f(w_{T+1})] \leq (1 - \lambda_T)f^* + \lambda_T\phi_1(w^*) + \mathcal{N}_{T+1}$$
$$\implies \mathbb{E}[f(w_{T+1}) - f^*] \leq \lambda_T\left[\phi_1(w^*) - f^*\right] + \mathcal{N}_{T+1}$$

By Eq. (20),

$$\leq \lambda_T\left[\phi_1^* + \frac{q_1}{2}\|w^* - z_1\|^2 - f^*\right] + \mathcal{N}_{T+1}$$

Choosing $\phi_1^* = f(w_1)$,

$$\leq \lambda_T \left[f(w_1) - f^* + \frac{q_1}{2} \|w^* - z_1\|^2 \right] + \mathcal{N}_T$$

Since $z_1 = w_1$, $q_1 = \mu$,

$$\implies \mathbb{E}[f(w_{T+1}) - f^*] \leq \lambda_T \left[f(w_1) - f^* + \frac{\mu}{2} \|w^* - w_1\|^2 \right] + \frac{2\sigma^2}{\rho^2 L} \sum_{j=1}^{T} \alpha_j^2 \prod_{i=j+1}^{T} (1 - r_i)$$

Using the fact that $\lambda_1 = 1$ and $\lambda_{k+1} = (1 - r_k)\lambda_k$, we know that that $\lambda_T = \prod_{k=1}^{T} (1 - r_k)$, and

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq \left[\prod_{k=1}^{T} (1 - r_k) \right] \left[f(w_1) - f^* + \frac{\mu}{2} \|w^* - w_1\|^2 \right] + \frac{\sigma^2}{2\rho^2 L} \sum_{j=1}^{T} \alpha_j^2 \prod_{i=j+1}^{T} (1 - r_i).$$

Now our task is to upper-bound bound the $1 - r_k$ terms. From Eq. (17), we know that

$$r_k = \sqrt{q_{k+1}\eta_k} = \sqrt{\frac{q_{k+1}}{\rho L}} \sqrt{\alpha_k} \geq \sqrt{\frac{q_{k+1}}{\rho L}} \alpha_k \qquad \text{(Since } \alpha_k \leq 1 \text{ for all } k\text{)}$$

$$\implies (1 - r_k) \leq \left(1 - \sqrt{\frac{q_{k+1}}{\rho L}} \alpha_k \right)$$

Since $q_k = \mu$ for all k, putting everything together,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq \left[\prod_{k=1}^{T} \left(1 - \sqrt{\frac{1}{\rho \kappa}} \alpha_k \right) \right] \left[f(w_1) - f^* + \frac{\mu}{2} \|w^* - w_1\|^2 \right] + \frac{\sigma^2}{2\rho^2 L} \sum_{j=1}^{T} \alpha_j^2 \prod_{i=j+1}^{T} \left(1 - \sqrt{\frac{1}{\rho \kappa}} \alpha_i \right)$$

Denoting $\Delta_k = \mathbb{E}[f(w_k) - f^*]$, using the exponential step-size $\alpha_k = \alpha^{k/T} = \left(\frac{1}{T} \right)^{k/T}$ and that $f(w_1) - f^* \geq \frac{\mu}{2} \|w^* - w_1\|^2$,

$$\Delta_{T+1} \leq 2 \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \sum_{k=1}^{T} \alpha^k \right) \Delta_1 + \frac{\sigma^2}{2\rho^2 L} \sum_{k=1}^{T} \alpha^{2k} \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \sum_{i=k+1}^{T} \alpha^i \right)$$

Using Lemma 5, we can bound the first term as

$$2 \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \sum_{k=1}^{T} \alpha^k \right) \Delta_1 \leq 2 \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \left(\frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)} \right) \right) \Delta_1$$

$$= 2c_3 \exp \left(-\frac{T}{\sqrt{\kappa \rho}} \frac{\alpha}{\ln(T/\beta)} \right) [f(w_1) - f^*]$$

where $c_3 = \exp \left(\frac{2\beta}{\sqrt{\rho \kappa} \ln(T/\beta)} \right)$. We can now bound the second term by a proof similar to Lemma 6. Indeed we have

$$\sum_{k=1}^{T} \alpha^{2k} \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \sum_{i=k+1}^{T} \alpha^i \right) = \sum_{k=1}^{T} \alpha^{2k} \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \frac{\alpha^{k+1} - \alpha^{T+1}}{1 - \alpha} \right)$$

$$= \exp \left(\frac{1}{\sqrt{\rho \kappa}} \frac{\alpha^{T+1}}{1 - \alpha} \right) \sum_{k=1}^{T} \alpha^{2k} \exp \left(-\sqrt{\frac{1}{\rho \kappa}} \frac{\alpha^{k+1}}{1 - \alpha} \right)$$

$$\leq \exp \left(\frac{1}{\sqrt{\rho \kappa}} \frac{\alpha^{T+1}}{1 - \alpha} \right) \sum_{k=1}^{T} \alpha^{2k} \left(\frac{2(1 - \alpha)\sqrt{\rho \kappa}}{e \alpha^{k+1}} \right)^2 \qquad \text{(Lemma 17)}$$

$$= \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{\alpha^{T+1}}{1-\alpha}\right) \frac{4\rho\kappa}{e^2\alpha^2} T(1-\alpha)^2$$

$$\leq \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{\alpha^{T+1}}{1-\alpha}\right) \frac{4\rho\kappa}{e^2\alpha^2} T\ln(1/\alpha)^2$$

$$= \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{\alpha^{T+1}}{1-\alpha}\right) \frac{4\rho\kappa \ln(T/\beta)^2}{e^2\alpha^2 T}$$

Finally,

$$\exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{\alpha^{T+1}}{1-\alpha}\right) \leq \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right) = c_3$$

where the inequality comes from the bound in Eq. (10) in the proof of Lemma 5. Putting everything together we obtain

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2c_3 \exp\left(-\frac{T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right) [f(w_1) - f^*] + \frac{2\sigma^2 c_3 \kappa \ln(T/\beta)^2}{\rho L e^2 \alpha^2 T}.$$

$\square$

E.4. Proof for misspecified ASGD

In this section, we assume that L and μ are misspecified by positive ν_L and ν_μ and we set $\nu = \nu_L \nu_\mu$. In particular, we will use $\tilde{L}$ and $\tilde{\mu}$, offline estimates of the smoothness and strong-convexity parameters. W.l.o.g we will assume that $\tilde{\mu} = \nu_\mu \mu$ and $\tilde{L} = \frac{L}{\nu_L}$. Importantly, we will assume that $\nu_\mu \leq 1$ i.e. we will only consider the more practical case where we underestimate the strong-convexity. Since $\nu_\mu \leq 1$, f is also $\tilde{\mu}$ strongly-convex. Hence, all the equations of (11) hold where we replace μ with $\tilde{\mu}$. Similarly in the definition of ϕ_k^* in (22) we can replace μ with $\tilde{\mu}$.

With this estimated value, the extrapolation parameter b_k is computed as follows:

$$r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k \tilde{\mu} \eta_k. \tag{24}$$

$$b_k = \frac{(1 - r_{k-1}) r_{k-1} \frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}}}, \tag{25}$$

where $\eta_k = \gamma_k \alpha_k = \frac{1}{\rho\tilde{L}} \alpha_k = \frac{\nu_L}{\rho L}\left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\tilde{\mu}}{\rho\tilde{L}}}\left(\frac{\beta}{T}\right)^{k/2T} = \sqrt{\frac{\nu_\mu \mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T} = \sqrt{\frac{\nu}{\rho\kappa}}\left(\frac{\beta}{T}\right)^{k/2T}$ satisfy the above equations. It can be verified that the reformulation in Appendix E.1 does not use the specific form of r_k and η_k, and only relies on the consistency of the update above. Hence, the update can be be reformulated as a 3 variable sequence as in Appendix E.1 with a different choice of η_k and r_k, but with q_k and z_k defined analogously in terms of η_k and r_k.

Similarly, it can be verified that the definition of the estimating sequences in Appendix E.2 also does not depend on the specific value of r_k and η_k, and hence we use the same definition of ϕ_k.

E.4.1. Proof of Theorem 7

Given the definitions in Appendix E.2, we first prove the following descent lemma with the misspecified step-size.

Lemma 14. *Using the update in Eq. (14) with* $\eta_k = \frac{\nu_L}{\rho L}\alpha_k$, *and defining* $k_0 := T\frac{[\ln(\nu_L)]_+}{\ln(T/\beta)}$ *and* $G = \max_{j\in[k_0]}\|\nabla f(y_j)\|$, *we obtain the following descent lemma.*

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2 + \mathcal{I}\{k<k_0\}\frac{G^2\nu_L^2}{2\rho L}\alpha_k^2$$

Proof. By smoothness, and the update in Eq. (14),

$$f(w_{k+1}) \leq f(y_k) - \eta_k\langle\nabla f(y_k), \nabla f_{ik}(y_k)\rangle + \frac{L}{2}\eta_k^2\|\nabla f_{ik}(y_k)\|^2$$

Taking expectation w.r.t. i_k,

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \eta_k\|\nabla f(y_k)\|^2 + \frac{L}{2}\eta_k^2\mathbb{E}[\|\nabla f_{ik}(y_k)\|^2] \quad (\eta_k \text{ is independent of the randomness in } i_k.)$$

$$\leq \mathbb{E}[f(y_k)] - \eta_k\|\nabla f(y_k)\|^2 + \frac{\rho L}{2}\eta_k^2\|\nabla f(y_k)\|^2 + \frac{L}{2}\eta_k^2\sigma^2 \quad (\text{By the growth condition in Eq. (6)})$$

$$= \mathbb{E}[f(y_k)] - \eta_k\|\nabla f(y_k)\|^2 + \frac{\eta_k\nu_L\alpha_k}{2}\|\nabla f(y_k)\|^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2$$

$$= \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 - \frac{\eta_k}{2}(1-\nu_L\alpha_k)\|\nabla f(y_k)\|^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2$$

For $k \geq k_0$, $1 - \nu_L\alpha_k \geq 0$, implying that

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2.$$

For $k < k_0$,

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 + \frac{\eta_k\nu_L\alpha_k}{2}\|\nabla f(y_k)\|^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2$$

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 + \frac{\alpha_k^2\nu_L^2}{2\rho L}\|\nabla f(y_k)\|^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2 \quad (\text{Since } \eta_k = \frac{\nu_L}{\rho L}\alpha_k,)$$

Since $G = \max_{j\in[k_0]}\|\nabla f(y_j)\|$, we can further upper-bound the RHS by

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\|\nabla f(y_k)\|^2 + \frac{G^2\nu_L^2}{2\rho L}\alpha_k^2 + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2.$$

$\square$

Let us now prove the equivalent of Lemma 13 using the above modified descent lemma.

Lemma 15. *For the estimating sequences defined in Appendix E.2 and the updates in Eq. (13)-Eq. (18),*

$$\mathbb{E}[\phi_k^*] := \mathbb{E}[\inf_w \phi_k(w)] \geq \mathbb{E}[f(w_k)] - \mathcal{N}_k$$

where $\mathcal{N}_k := \frac{\sigma^2 \nu_L^2}{2\rho^2 L} \sum_{j=1}^k \alpha_j^2 \prod_{i=j+1}^k (1 - r_i) + \left(\frac{G^2 \nu_L^2}{2\rho L} \right) \sum_{j=1}^{\min\{k_0,k\}-1} \alpha_j^2 \prod_{i=j+1}^k (1 - r_i)$

Proof. We will prove the lemma by induction. For $k = 1$, we define $\phi_1^* = f(w_1)$, and since $\mathcal{N}_k \geq 0$ for all k, $\mathbb{E}[\phi^*] \geq f(w_1) - \mathcal{N}_1$, thus satisfying the base-case for the induction. For the induction, we will use the fact that $\mathcal{N}_{k+1} = (1 - r_k)\mathcal{N}_k + \frac{2\nu_L^2 \sigma^2}{\rho^2 L} \alpha_k^2 + \mathcal{I}\{k < k_0\} \frac{G^2 \nu_L^2}{2\rho L} \alpha_k^2$.

Assuming the induction hypothesis, $\mathbb{E}[\phi_k^*] \geq \mathbb{E}[f(w_k)] - \mathcal{N}_k$, we use Eq. (22) to prove the statement for $k + 1$ as follows. Taking expectations w.r.t to the randomness in $j = 1$ to k,

$$\mathbb{E}[\phi_{k+1}^*] = (1 - r_k)\mathbb{E}[\phi_k^*] + r_k \mathbb{E}\left[f(y_k) - \frac{r_k}{2q_{k+1}} \|\nabla f(y_k)\|^2 + \frac{(1 - r_k) q_k}{q_{k+1}} \left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) \right]$$

$$\geq (1 - r_k)\mathbb{E}[f(w_k) - \mathcal{N}_k]$$

$$+ r_k \mathbb{E}\left[f(y_k) - \frac{r_k}{2q_{k+1}} \|\nabla f(y_k)\|^2 + \frac{(1 - r_k) q_k}{q_{k+1}} \left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) \right]$$

(by the induction hypothesis)

$$= (1 - r_k)\mathbb{E}[f(w_k)] + r_k \mathbb{E}[f(y_k)] - \frac{r_k^2}{2q_{k+1}} \mathbb{E}\|\nabla f(y_k)\|^2$$

$$+ \frac{r_k(1 - r_k) q_k}{q_{k+1}} \mathbb{E}\left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right)$$

$$- (1 - r_k)\mathcal{N}_k$$

$$= (1 - r_k)\mathbb{E}[f(w_k)] + r_k \mathbb{E}[f(y_k)] - \frac{\eta_k}{2} \mathbb{E}\|\nabla f(y_k)\|^2$$

$$+ \frac{r_k(1 - r_k) q_k}{q_{k+1}} \mathbb{E}\left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) - (1 - r_k)\mathcal{N}_k$$

(Using Eq. (17))

By convexity, $f(w_k) \geq f(y_k) + \langle \nabla f(y_k), w_k - y_k \rangle$,

$$\geq (1 - r_k)\mathbb{E}[f(y_k) + \langle \nabla f(y_k), w_k - y_k \rangle] + r_k \mathbb{E}[f(y_k)] - \frac{\eta_k}{2} \mathbb{E}\|\nabla f(y_k)\|^2$$

$$+ \frac{r_k(1 - r_k) q_k}{q_{k+1}} \mathbb{E}\left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) - (1 - r_k)\mathcal{N}_k$$

$$= \mathbb{E}\left[f(y_k) - \frac{\eta_k}{2} \|\nabla f(y_k)\|^2 \right] + \frac{r_k(1 - r_k) q_k}{q_{k+1}} \mathbb{E}\left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right)$$

$$+ (1 - r_k)\mathbb{E}[\langle \nabla f(y_k), w_k - y_k \rangle] - (1 - r_k)\mathcal{N}_k$$

By Lemma 14,

$$\geq \mathbb{E}\left[f(w_{k+1}) - \frac{\nu_L^2}{2\rho^2 L} \alpha_k^2 \sigma^2 - \mathcal{I}\{k < k_0\} \frac{G^2 \nu_L^2}{2\rho L} \alpha_k^2 \right] + \frac{r_k(1 - r_k) q_k}{q_{k+1}} \mathbb{E}\left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right)$$

$$+ (1 - r_k)\mathbb{E}[\langle \nabla f(y_k), w_k - y_k \rangle] - (1 - r_k)\mathcal{N}_k$$

$$= \mathbb{E}[f(w_{k+1})] + \frac{r_k(1 - r_k) q_k}{q_{k+1}} \mathbb{E}\left(\frac{\tilde{\mu}}{2} \|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle \right) + (1 - r_k)\mathbb{E}[\langle \nabla f(y_k), w_k - y_k \rangle]$$

$$- \left[(1 - r_k)\mathcal{N}_k + \frac{\nu_L^2}{2\rho^2 L} \alpha_k^2 \sigma^2 + \mathcal{I}\{k < k_0\} \frac{G^2 \nu_L^2}{2\rho L} \alpha_k^2 \right]$$

Since $\mathcal{N}_{k+1} = \left[(1 - r_k)\mathcal{N}_k + \frac{\nu_L^2}{2\rho^2 L}\alpha_k^2\sigma^2 + \mathcal{I}\{k < k_0\} \frac{G^2\nu_L^2}{2\rho L}\alpha_k^2 \right]$,

$$\mathbb{E}[\phi_{k+1}^*] \geq \mathbb{E}\left[f(w_{k+1})\right] - \mathcal{N}_{k+1} + (1 - r_k)\mathbb{E}[\langle \nabla f(y_k), w_k - y_k \rangle]$$
$$+ \frac{r_k(1 - r_k)\,q_k}{q_{k+1}}\mathbb{E}\left(\frac{\tilde{\mu}}{2}\|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle\right)$$

Similar to Lemma 13, we show that $(1 - r_k)\mathbb{E}\left[\langle \nabla f(y_k), w_k - y_k \rangle + \frac{r_k\,q_k}{q_{k+1}}\left(\frac{\tilde{\mu}}{2}\|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle\right)\right] \geq 0$.
For this, we use modified Eq. (13)

$$y_k = w_k - \frac{q_k r_k}{q_k + r_k\tilde{\mu}}(w_k - z_k)$$

$$\implies z_k - y_k = z_k - w_k + \frac{q_k r_k}{q_k + r_k\tilde{\mu}}(w_k - z_k) = \left(1 - \frac{q_k r_k}{q_k + r_k\tilde{\mu}}\right)(z_k - w_k)$$

$$= \left(\frac{q_k(1 - r_k) + r_k\tilde{\mu}}{q_k + r_k\tilde{\mu}}\right)(z_k - w_k) = \left(\frac{q_{k+1}}{q_k + r_k\tilde{\mu}}\right)(z_k - w_k) \qquad \text{(By Eq. (16))}$$

$$\implies \frac{r_k q_k}{q_{k+1}}\langle \nabla f(y_k), z_k - y_k \rangle = \left\langle \nabla f(y_k), \left(-\frac{r_k\,q_k}{q_k + r_k\tilde{\mu}}\right)(w_k - z_k) \right\rangle = \langle \nabla f(y_k), y_k - w_k \rangle$$

Using this relation to simplify,

$$(1 - r_k)\mathbb{E}\left[\langle \nabla f(y_k), w_k - y_k \rangle + \frac{r_k\,q_k}{q_{k+1}}\left(\frac{\tilde{\mu}}{2}\|y_k - z_k\|^2 + \langle \nabla f(y_k), z_k - y_k \rangle\right)\right]$$

$$= (1 - r_k)\mathbb{E}\left[\frac{r_k\,q_k}{q_{k+1}}\frac{\tilde{\mu}}{2}\|y_k - z_k\|^2 + (1 - r_k)\left[\langle \nabla f(y_k), w_k - y_k \rangle + \langle \nabla f(y_k), y_k - w_k \rangle\right]\right]$$

$$= (1 - r_k)\mathbb{E}\left[\frac{r_k\,q_k}{q_{k+1}}\frac{\tilde{\mu}}{2}\|y_k - z_k\|^2\right] \geq 0 \qquad \text{(Since } r_k \leq 1.\text{)}$$

Putting everything together,

$$\mathbb{E}[\phi_{k+1}^*] \geq \mathbb{E}\left[f(w_{k+1})\right] - \mathcal{N}_{k+1}$$

and we conclude that $\mathbb{E}[\phi_k^*] \geq \mathbb{E}\left[f(w_k)\right] - \mathcal{N}_k$ for all k by induction. $\qquad\square$

We now use the above lemma to prove Theorem 7.

Theorem 7. Under the same assumptions as Theorem 6 and (iv) $\nu = \nu_L \nu_\mu \leq \rho \kappa$, ASGD (Eqs. (4) and (5)) with $w_1 = y_1$, $\gamma_k = \frac{1}{\rho \tilde{L}} = \frac{\nu_L}{\rho L}$, $\alpha_k = \left(\frac{\beta}{T}\right)^{k/T}$, $\tilde{\mu} = \nu_\mu \mu \leq \mu$, $r_k = \sqrt{\frac{\tilde{\mu}}{\rho \tilde{L}}} \left(\frac{\beta}{T}\right)^{k/2T} = \sqrt{\frac{\nu}{\rho \kappa}} \left(\frac{\beta}{T}\right)^{k/2T}$ and $b_k = \frac{(1 - r_{k-1}) \, r_{k-1} \, \alpha}{r_k + r_{k-1}^2 \, \alpha}$ converges as,

$$\Delta_{T+1} \leq 2c_3 \exp\left(-\frac{\sqrt{\min\{\nu, 1\}} T}{\sqrt{\kappa \rho}} \frac{\alpha}{\ln(T/\beta)}\right) \Delta_1$$

$$+ \frac{2c_3 (\ln(T/\beta))^2}{e^2 \alpha^2 \mu T} \left[\frac{\sigma^2}{\rho} + G^2 \min\{\frac{k_0}{T}, 1\}\right] \max\{\frac{\nu_L}{\nu_\mu}, \nu_L^2\},$$

where $\Delta_k := \mathbb{E}[f(w_k) - f^*]$, $c_3 = \exp\left(\frac{1}{\sqrt{\rho \kappa}} \frac{2\beta}{\ln(T/\beta)}\right)$, $[x]_+ = \max\{x, 0\}$, $k_0 := \lfloor T \frac{[\ln(\nu_L)]_+}{\ln(T/\beta)} \rfloor$ and $G = \max_{j \in [k_0]} \|\nabla f(y_j)\|$.

Proof. In order to have a valid estimate sequence, we need to restrict values that ν. Since $\nu \leq \rho \kappa$, $0 \leq r_k \leq 1$ and hence $\lambda_k \in (0, 1)$, as required for a valid estimate sequence. For the estimating sequences defined in Appendix E.2, using Lemma 13, we know that the (reformulated) updates satisfy the following relation

$$\mathbb{E}[f(w_{T+1})] \leq \mathbb{E}[\phi_{T+1}^*] + \mathcal{N}_{T+1} \leq \mathbb{E}[\phi_{T+1}(w^*)] + \mathcal{N}_{T+1}$$

From Eq. (21), we know that for all w and k,

$$\phi_k(w) \leq (1 - \lambda_k) f(w) + \lambda_k \phi_0(w)$$

Using these relations,

$$\mathbb{E}[f(w_{T+1})] \leq (1 - \lambda_T) f^* + \lambda_T \phi_1(w^*) + \mathcal{N}_T \implies \mathbb{E}[f(w_{T+1}) - f^*] \qquad \leq \lambda_T [\phi_1(w^*) - f^*] + \mathcal{N}_{T+1}$$

By Eq. (20),

$$\leq \lambda_T \left[\phi_1^* + \frac{q_1}{2} \|w^* - z_1\|^2 - f^*\right] + \mathcal{N}_{T+1}$$

Choosing $\phi_1^* = f(w_1)$,

$$\leq \lambda_T \left[f(w_1) - f^* + \frac{q_1}{2} \|w^* - z_1\|^2\right] + \mathcal{N}_T$$

Since $z_1 = w_1$, and we set $q_1 = \tilde{\mu}$,

$$\implies \mathbb{E}[f(w_{T+1}) - f^*] \leq \lambda_T \left[f(w_1) - f^* + \frac{\tilde{\mu}}{2} \|w^* - w_1\|^2\right]$$

$$+ \frac{\sigma^2 \nu_L^2}{2\rho^2 L} \sum_{j=1}^{T} \alpha_j^2 \prod_{i=j+1}^{T} (1 - r_i) + \left(\frac{G^2 \nu_L^2}{2\rho L}\right) \sum_{j=1}^{\min\{k_0, T\}-1} \alpha_j^2 \prod_{i=j+1}^{T} (1 - r_i)$$

Using the fact that $\lambda_1 = 1$ and $\lambda_{k+1} = (1 - r_k)\lambda_k$, and since $\nu \leq \rho \kappa$, we know that $r_k \leq 1$ and $\lambda_T = \prod_{k=1}^{T} (1 - r_k)$, and

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq \left[\prod_{k=1}^{T} (1 - r_k)\right] \left[f(w_1) - f^* + \frac{\tilde{\mu}}{2} \|w^* - w_1\|^2\right] + \frac{\sigma^2 \nu_L^2}{2\rho^2 L} \sum_{j=1}^{T} \alpha_j^2 \prod_{i=j+1}^{T} (1 - r_i)$$

$$+ \left(\frac{G^2 \nu_L^2}{2\rho L}\right) \sum_{j=1}^{\min\{k_0, T\}-1} \alpha_j^2 \prod_{i=j+1}^{T} (1 - r_i).$$

Now our task is to upper-bound bound the $1 - r_k$ terms. From Eq. (17), we know that

$$r_k = \sqrt{q_{k+1}\eta_k} = \sqrt{\frac{q_{k+1}\nu_L}{\rho L}}\sqrt{\alpha_k} \geq \sqrt{\frac{q_{k+1}\nu_L}{\rho L}}\alpha_k \qquad \text{(Since } \alpha_k \leq 1 \text{ for all } k\text{)}$$

$$\implies (1 - r_k) \leq \left(1 - \sqrt{\frac{q_{k+1}\nu_L}{\rho L}}\alpha_k\right)$$

Since $q_k = \tilde{\mu}$ for all k, putting everything together,

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq \left[\prod_{k=1}^{T}\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_k\right)\right]\left[f(w_1) - f^* + \frac{\tilde{\mu}}{2}\|w^* - w_1\|^2\right] + \frac{\sigma^2\nu_L^2}{2\rho^2 L}\sum_{j=1}^{T}\alpha_j^2\prod_{i=j+1}^{T}\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_i\right)$$

$$+ \left(\frac{G^2\nu_L^2}{2\rho L}\right)^{\min\{k_0,T\}-1}\sum_{j=1}^{T}\alpha_j^2\prod_{i=j+1}^{T}\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_i\right).$$

Denoting $\Delta_k = \mathbb{E}[f(w_k) - f^*]$ and $\Delta_1 = \mathbb{E}[f(w_1) - f^*]$, using the exponential step-size $\alpha_k = \alpha^{k/T} = \left(\frac{1}{T}\right)^{k/T}$ and that $f(w_1) - f^* \geq \frac{\tilde{\mu}}{2}\|w_1 - w^*\|^2$. Similar to the proof of Theorem 4, if $\nu > 1$, we can replace ν by 1 and get an upper-bound for the $\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_k\right)$ term. Hence, we define $\hat{\nu} := \min\{1, \nu\}$, and obtain the following upper-bound.

$$\Delta_{T+1} \leq 2\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\sum_{k=1}^{T}\alpha^k\right)\Delta_1 + \frac{\sigma^2\nu_L^2}{2\rho^2 L}\sum_{k=1}^{T}\alpha^{2k}\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\sum_{i=k+1}^{T}\alpha^i\right)$$

$$+ \frac{G^2\nu_L^2}{2\rho L}^{\min\{k_0,T\}-1}\sum_{k=1}^{T}\alpha^{2k}\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\sum_{i=k+1}^{T}\alpha^i\right)$$

Using Lemma 5, we can bound the first term as

$$2\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\sum_{k=1}^{T}\alpha^k\right)\Delta_1 \leq 2\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\left(\frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}\right)\right)\Delta_1$$

$$\leq 2\exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)\exp\left(-\frac{T\sqrt{\hat{\nu}}}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_1) - f^*] \qquad \text{(since } \hat{\nu} \leq 1\text{)}$$

$$= 2c_3\exp\left(-\frac{T\sqrt{\hat{\nu}}}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_1) - f^*]$$

where $c_3 = \exp\left(\frac{2\beta}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$. We can now bound the second term by a proof similar to Lemma 6. Indeed we have

$$\sum_{k=1}^{T}\alpha^{2k}\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\sum_{i=k+1}^{T}\alpha^i\right) = \sum_{k=1}^{T}\alpha^{2k}\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\frac{\alpha^{k+1} - \alpha^{T+1}}{1-\alpha}\right)$$

$$= \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\sum_{k=1}^{T}\alpha^{2k}\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\frac{\alpha^{k+1}}{1-\alpha}\right)$$

$$\leq \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\sum_{k=1}^{T}\alpha^{2k}\left(\frac{2(1-\alpha)\sqrt{\rho\kappa}}{e\alpha^{k+1}\sqrt{\hat{\nu}}}\right)^2 \qquad \text{(Lemma 17)}$$

$$= \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\frac{4\rho\kappa}{e^2\hat{\nu}\alpha^2}T(1-\alpha)^2$$

$$\leq \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\frac{4\rho\kappa}{e^2\hat{\nu}\alpha^2}T\ln(1/\alpha)^2$$

$$= \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2}{e^2\hat{\nu}\alpha^2T}$$

Similarly,

$$\sum_{k=1}^{\min\{k_0,T\}-1}\alpha^{2k}\exp\left(-\sqrt{\frac{\hat{\nu}}{\rho\kappa}}\sum_{i=k+1}^{T}\alpha^i\right) \leq \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2\min\{k_0,T\}}{e^2\hat{\nu}\alpha^2T^2}$$

$$= \exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2\min\left\{\frac{\ln(\nu)}{\ln(T/\beta)},1\right\}}{e^2\hat{\nu}\alpha^2T}$$

Finally,

$$\exp\left(\frac{\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}}\frac{\alpha^{T+1}}{1-\alpha}\right) \leq \exp\left(\frac{2\beta\sqrt{\hat{\nu}}}{\sqrt{\rho\kappa}\ln(T/\beta)}\right) = c_3$$

where the inequality comes from the bound in Eq. (10) in the proof of Lemma 5. Putting everything together we obtain

$$\mathbb{E}[f(w_{T+1}) - f^*] \leq 2c_3\exp\left(-\frac{\sqrt{\hat{\nu}}T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\Delta_1 + \frac{2c_3\kappa\ln(T/\beta)^2}{e^2\alpha^2T}\frac{\sigma^2\nu_L^2}{\hat{\nu}\rho L}$$

$$+ \frac{2c_3\kappa\ln(T/\beta)^2}{e^2\alpha^2T}\min\left\{\frac{[\ln(\nu_L)]_+}{\ln(T/\beta)},1\right\}\frac{G^2\nu_L^2}{L\hat{\nu}}$$

$$\leq \exp\left(-\frac{\sqrt{\hat{\nu}}T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\Delta_1 + \frac{2c_3\kappa\ln(T/\beta)^2}{e^2\alpha^2T}\frac{\sigma^2\max\{\nu_L^2,\nu_L/\nu_\mu\}}{\rho L}$$

$$+ \frac{2c_3\kappa\ln(T/\beta)^2}{e^2\alpha^2T}\min\left\{\frac{[\ln(\nu_L)]_+}{\ln(T/\beta)},1\right\}\frac{G^2\max\{\nu_L^2,\nu_L/\nu_\mu\}}{L}$$

$$= \exp\left(-\frac{\sqrt{\hat{\nu}}T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)\Delta_1 + \frac{2c_3\ln(T/\beta)^2}{e^2\alpha^2T}\frac{\sigma^2\max\{\nu_L^2,\nu_L/\nu_\mu\}}{\rho\mu}$$

$$+ \frac{2c_3\ln(T/\beta)^2}{e^2\alpha^2T}\min\left\{\frac{[\ln(\nu_L)]_+}{\ln(T/\beta)},1\right\}\frac{G^2\max\{\nu_L^2,\nu_L/\nu_\mu\}}{\mu}$$

For the second inequality, we consider the term $\frac{\nu_L^2}{\hat{\nu}}$. If $\hat{\nu} = \nu$, then $\frac{\nu_L^2}{\hat{\nu}} = \frac{\nu_L^2}{\nu} = \frac{\nu_L}{\nu_\mu}$. If $\hat{\nu} = 1$, then $\frac{\nu_L^2}{\hat{\nu}} = \nu_L^2$. Putting these two cases together we get $\max\{\nu^2, \nu_L/\nu_\mu\}$. The last equality comes from the fact that $\frac{\kappa}{L} = \frac{1}{\mu}$.

$\square$

F. Helper Lemmas

Lemma 16. *For all $x > 1$,*

$$\frac{1}{x-1} \leq \frac{2}{\ln(x)}$$

Proof. For $x > 1$, we have

$$\frac{1}{x-1} \leq \frac{2}{\ln(x)} \iff \ln(x) < 2x - 2$$

Define $f(x) = 2x - 2 - \ln(x)$. We have $f'(x) = 2 - \frac{1}{x}$. Thus for $x \geq 1$, we have $f'(x) > 0$ so f is increasing on $[1, \infty)$. Moreover we have $f(1) = 2 - 2 - \ln(1) = 0$ which shows that $f(x) \geq 0$ for all $x > 1$ and ends the proof. $\square$

Lemma 17. *For all $x, \gamma > 0$,*

$$\exp(-x) \leq \left(\frac{\gamma}{ex}\right)^{\gamma}$$

Proof. Let $x > 0$. Define $f(\gamma) = \left(\frac{\gamma}{ex}\right)^{\gamma} - \exp(-x)$. We have

$$f(\gamma) = \exp\left(\gamma \ln(\gamma) - \gamma \ln(ex)\right) - \exp(-x)$$

and

$$f'(\gamma) = \left(\gamma \cdot \frac{1}{\gamma} + \ln(\gamma) - \ln(ex)\right) \exp\left(\gamma \ln(\gamma) - \gamma \ln(ex)\right)$$

Thus

$$f'(\gamma) \geq 0 \iff 1 + \ln(\gamma) - \ln(ex) \geq 0 \iff \gamma \geq \exp\left(\ln(ex) - 1\right) = x$$

So f is decreasing on $(0, x]$ and increasing on $[x, \infty)$. Moreover,

$$f(x) = \left(\frac{x}{ex}\right)^{x} - \exp(-x) = \left(\frac{1}{e}\right)^{x} - \exp(-x) = 0$$

and thus $f(\gamma) \geq 0$ for all $\gamma > 0$ which proves the lemma. $\square$

Lemma 18. *For any sequence α_k*

$$\prod_{k=1}^{T}(1 - \alpha_k) + \sum_{k=1}^{T} \alpha_k \prod_{i=k+1}^{T}(1 - \alpha_i) = 1$$

Proof. We show this by induction on T. For $T = 1$,

$$(1 - \alpha_1) + \alpha_1 = 1$$

Induction step:

$$\prod_{k=1}^{T+1}(1-\alpha_k) + \sum_{k=1}^{T+1} \alpha_k \prod_{i=k+1}^{T+1}(1-\alpha_i) = (1-\alpha_{T+1}) \prod_{k=1}^{T}(1-\alpha_k) + \left(\alpha_{T+1} + \sum_{k=1}^{T} \alpha_k \prod_{i=k+1}^{T+1}(1-\alpha_i)\right)$$

$$= (1-\alpha_{T+1}) \prod_{k=1}^{T}(1-\alpha_k) + \left(\alpha_{T+1} + (1-\alpha_{T+1}) \sum_{k=1}^{T} \alpha_k \prod_{i=k+1}^{T}(1-\alpha_i)\right)$$

$$= (1 - \alpha_{T+1}) \left(\underbrace{\prod_{k=1}^{T}(1 - \alpha_k) + \sum_{k=1}^{T} \alpha_k \prod_{i=k+1}^{T}(1 - \alpha_i)}_{=1} \right) + \alpha_{T+1}$$

(Induction hypothesis)

$$= (1 - \alpha_{T+1}) + \alpha_{T+1} = 1$$

$\square$

G. Comparison to Aybat et al. (2019)

Aybat et al. (2019) use the condition

$$\mathbb{E}_i \left\| \nabla f_i(w) - \nabla f(w) \right\|^2 \le \sigma^2 + \eta^2 \left\| w - w^* \right\|^2 \tag{26}$$

In order to relate the above condition to the one in Eq. (6), we use smoothness of f,

$$\mathbb{E} \left\| \nabla f_{i_k}(x_k) - \nabla f(x_k) \right\|^2 = \mathbb{E} \left\| \nabla f_{i_k}(x_k) \right\|^2 - \left\| \nabla f(x_k) \right\|^2 \le \sigma^2 + (\rho - 1) \left\| \nabla f(x_k) \right\|^2 \le \sigma^2 + (\rho - 1)L^2 \left\| x_k - x^* \right\|^2,$$

and thus $\eta^2 = (\rho - 1)L^2$.

Similarly, we can use the strong-convexity of f, and obtain the following relation,

$$\mathbb{E} \left\| \nabla f_{i_k}(x_k) \right\|^2 = \mathbb{E} \left\| \nabla f_{i_k}(x_k) - \nabla f(x_k) \right\|^2 + \left\| \nabla f(x_k) \right\|^2$$
$$\le \sigma^2 + \eta^2 \left\| x_k - x^* \right\|^2 + \left\| \nabla f(x_k) \right\|^2$$
$$\le \sigma^2 + \eta^2 \frac{1}{\mu^2} \left\| \nabla f(x_k) \right\|^2 + \left\| \nabla f(x_k) \right\|^2,$$

and thus $\rho = \left(\frac{\eta^2}{\mu^2} + 1 \right)$.

Hence, the two assumptions are equivalent up to constants. In order to get an idea about the strength of the two results, we compare the two upper-bounds (we note that such a comparison is not ideal, and it may be possible to improve and tailor the upper-bounds to the specific growth condition).

Under the assumption in Eq. (26), in Appendix K, the authors prove the following rate,

$$\mathbb{E}[f(w_T) - f^*] \le O(1) \left(\exp \left(\frac{-T}{O(1)(\sqrt{\kappa} + \eta^2/\mu^2)} \right) (f(w_0) - f^*) + \frac{\sigma^2}{T\mu} \right).$$

If we were to use Eq. (6), since $\eta^2 = (\rho - 1)L^2$, this result gives the following upper-bound,

$$\mathbb{E}[f(w_T) - f^*] \le O(1) \left(\exp \left(\frac{-T}{O(1)(\sqrt{\kappa} + (\rho - 1)\kappa^2)} \right) (f(x_0) - f^*) + \frac{\sigma^2}{T\mu} \right)$$

Observe that for $\rho > 1$, this bound does not result in an accelerated rate.

On the other hand, if we were to use Eq. (26), since $\rho = \left(\frac{\eta^2}{\mu^2} + 1 \right)$, the result in Theorem 6 can be written as:

$$\mathbb{E}[f(w_T) - f^*] \le 2c_3 \exp \left(-\frac{T}{\sqrt{\kappa(1 + \eta^2/\mu^2)}} \frac{\alpha}{\ln(T/\beta)} \right) (f(w_0) - f^*) + \frac{8\sigma^2 c_4 \kappa}{(\eta^2/\mu^2)Le^2} \frac{(\ln(T/\beta))^2}{\alpha^2 T}$$

Observe that for any $\eta > 1$, this bound gives an accelerated rate, meaning that the accelerated result in Theorem 6 also holds for the alternative growth condition in Eq. (26).