Towards Noise-adaptive, Problem-adaptive Stochastic Gradient Descent

Sharan Vaswani, Benjamin Dubois-Taine, Reza Babanezhad

▶ **To cite this version:**

Sharan Vaswani, Benjamin Dubois-Taine, Reza Babanezhad. Towards Noise-adaptive, Problem-adaptive Stochastic Gradient Descent. 2021. hal-03456663

HAL Id: hal-03456663
https://hal.science/hal-03456663

Preprint submitted on 30 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards Noise-adaptive, Problem-adaptive Stochastic Gradient Descent

Sharan Vaswani[⋆]
Amii, University of Alberta

Benjamin Dubois-Taine
CNRS, ENS Paris & Inria

Reza Babanezhad
SAIT AI lab, Montreal

Abstract

We design step-size schemes that 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 first prove that T iterations of SGD with Nesterov acceleration and exponentially decreasing step-sizes can achieve a near-optimal $\tilde{O}\left(\exp\left(-T/\sqrt{\kappa}\right) + \sigma^2/T\right)$ convergence rate. Under a relaxed assumption on the noise, with the same step-size scheme and knowledge of the smoothness, we prove that SGD can achieve an $\tilde{O}\left(\exp\left(-T/\kappa\right) + \sigma^2/T\right)$ rate. In order to be adaptive to the smoothness, we use a stochastic line-search (SLS) and show (via upper and lower-bounds) that SGD converges at the desired rate, but only to a neighbourhood of the solution. Next, we use SGD with an offline estimate of the smoothness, and prove convergence to the minimizer. However, its convergence is slowed down proportional to the estimation error and we prove a lower-bound justifying this slowdown. Compared to other step-size schemes, 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,

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

For supervised learning, n represents the number of training examples and f_i is the loss of example i. Throughout the main paper, we assume f to be a smooth, strongly-convex function and denote w^* to be the unique minimizer of the above problem. We consider broader function classes in Appendix B.

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 numerous ways of selecting it have been studied in the literature. 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 in the literature (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 and has an $O(\exp(-T/\kappa))$ convergence 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 variants of gradient descent use a different step-size strategy to obtain the optimal rates in their respective settings.

Towards noise adaptivity: Ideally, we want a *noise-adaptive* algorithm such that (i) it obtains the optimal

⋆ Correspondence to vaswani.sharan@gmail.com.

rates in both the deterministic and stochastic settings (ii) its convergence rate depends on the noise in the stochastic gradients and (iii) the resulting algorithm does not require knowledge of the stochasticity (e.g. an upper bound on the variance in stochastic gradients).

There have been three recent attempts to obtain such an algorithm. For smooth, strongly-convex functions, if σ^2 is the noise level in the stochastic gradients, Stich (2019) achieves an $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence rate using SGD that switches between two carefully designed step-sizes. When $\sigma = 0$, the resulting algorithm achieves the deterministic GD rate, whereas for a non-zero σ, its rate is governed by the σ^2/T term. Unfortunately, setting the algorithm parameters requires the knowledge of σ and hence, it is not noise-adaptive. On the other hand, Khaled and Richtárik (2020) and Li et al. (2020) do not require knowledge of σ, and can obtain the same rate for smooth functions satisfying the Polyak-Lojasiewicz (PL) condition (Karimi et al., 2016), a generalization of strong-convexity. For this, Li et al. (2020) use an exponentially decreasing sequence of step-sizes, while Khaled and Richtárik (2020) use a constant then decaying step-size. However, neither of these methods match the optimal $\sqrt{\kappa}$ dependence in the linear convergence term.

Contribution: Since we consider the easier (compared to PL) strongly-convex setting, it is unclear if we can achieve the above rate by using the conventional polynomially decreasing step-sizes (Robbins and Monro, 1951). Unfortunately, in Lemmas 3 and 4, we prove that no polynomially decreasing step-size of the form $O\left(\frac{1}{k^\delta}\right)$ for $\delta \in [0, 1]$ can achieve the desired $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ rate[1]. Consequently, we will use an exponentially decreasing step-size.

Contribution: In Section 3, we use SGD with an exponentially decreasing step-size and a stochastic variant of Nesterov acceleration (Vaswani et al., 2019a). Under a growth-condition similar to Li et al. (2020); Khaled and Richtárik (2020); Bottou et al. (2018), we prove that the resulting algorithm achieves the near-optimal $\tilde{O}\left(\exp(-T/\sqrt{\kappa}) + \frac{\sigma^2}{T}\right)$ convergence when minimizing smooth, strongly-convex functions. Our algorithm thus achieves the near-optimal rate in both the stochastic and deterministic settings and its rate smoothly varies between the two regimes. Furthermore, our algorithm does not require knowledge of σ^2 and hence satisfies three desiderata outlined above. To the best of our knowledge, this is the first such result.

Towards noise and problem adaptivity: Typically, SGD also requires the knowledge of problem-dependent constants (such as smoothness or strong-convexity) to set the step-size. In practice, it is difficult to estimate these problem-dependent constants, 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 parameters, 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, all the noise-adaptive methods (Li et al., 2020; Khaled and Richtárik, 2020; Stich, 2019) including the algorithm proposed in Section 3 require the knowledge of problem-dependent constants and are thus not problem-adaptive.

In order to make progress towards an SGD variant that is both noise-adaptive and problem-adaptive, we only consider algorithms that can achieve the (non-optimal) $\tilde{O}\left(\exp(-T/\kappa) + \frac{\sigma^2}{T}\right)$ convergence rate. We note that 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 4.2, we use stochastic line-search (Vaswani et al., 2019b) to estimate the smoothness constant on the fly. We prove that SGD in conjunction with exponentially decreasing step-sizes and stochastic line-search (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 step-size is set in an online fashion (using the sampled function like in SLS), no decreasing sequence of step-sizes can converge to the minimizer.

Contribution: In Section 4.3, we consider estimating the smoothness constant in an offline fashion (before running SGD). 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. Our upper-bound shows that even if we misestimate the smoothness constant by a multiplicative factor of ν, the convergence can slow down by a factor as large as $O(\exp(\nu))$. We complement this result with a lower-bound that shows that such a misestimation in the smoothness necessarily slows down the rate by a potentially exponential factor.

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

[1] Note that this result does not cover step-size sequences that switch between two values of δ, for example in (Khaled and Richtárik, 2020)

Contribution: In Section 5, we compare the performance of different step-size schemes on convex, supervised learning problems. Furthermore, we propose a novel variant of SLS that guarantees converges to the minimizer and demonstrate its practical effectiveness.

Contribution: Finally, in Appendix B, we prove matching results for SGD on strongly star-convex functions (Hinder et al., 2020), a class of structured non-convex functions. We also show the explicit dependence of our results on the mini-batch size. Finally, we prove upper-bounds for (non-strongly-) convex functions, showing 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 any polynomial/exponentially decreasing step-size sequence is unlikely to be noise-adaptive in this setting.

2 Problem setup and Background

We will assume that f and each f_i are differentiable and lower-bounded by f^* and f_i^*, respectively. Throughout the main paper, we will assume that f is μ-strongly convex, and each f_i is convex. Furthermore, we assume that each function f_i is L_i-smooth, implying that f is L-smooth with $L := \max_i L_i$. We include definitions of these properties in Appendix A.

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 in that direction. Specifically,

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

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 $\mathbb{E}_i\left[\nabla f_i(w)|w_k\right] = \nabla f(w)$. The product of scalars $\eta_k := \gamma_k \alpha_k$ defines the *step-size* for iteration k. The step-size consists of two parts - a problem-dependent scaling term γ_k that captures the (local) smoothness of the function, and a problem-independent term α_k 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 α_k sequence depends on the properties of f, for example, for convex functions, $\alpha_k = O(1/\sqrt{k})$ while for strongly-convex functions, $\alpha_k = O(1/k)$.

Throughout the paper, we will assume that T is known in advance (this requirement can be relaxed via the standard doubling trick), and consider exponentially decreasing step-sizes (Li et al., 2020) where $\alpha := \left[\frac{\beta}{T}\right]^{1/T} \leq 1$ for some parameter $\beta \geq 1$ and $\alpha_k := \alpha^k$.

Unlike SGD, 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. Specifically, the update in iteration k of ASGD is:

$$y_k = w_k + b_k\left(w_k - w_{k-1}\right), \tag{3}$$

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

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

3 Convergence of ASGD

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) – there exists a (ρ, σ) with $\rho \geq 1$ and $\sigma \geq 0$, such that for all w,

$$\mathbb{E}_i \left\|\nabla f_i(w)\right\|^2 \leq \rho \left\|\nabla f(w)\right\|^2 + \sigma^2. \tag{5}$$

Note that in the deterministic setting (when using the full-gradient in Eq. (4)), $\rho = 1$ and $\sigma = 0$. Similarly, $\sigma = 0$ when the stochastic gradients satisfy the *strong-growth condition* when using over-parameterized models (Schmidt and Roux, 2013; Ma et al., 2018; Vaswani et al., 2019a). Under the above growth condition, we prove the following theorem in Appendix C.

Theorem 1. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f and (iii) the growth condition in Eq. (5), ASGD (Eqs. (3) and (4)) with $w_0 = y_0$, $\gamma_k = \frac{1}{\rho L}$, $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and b_k computed as:*

$$b_k = \frac{\left(1 - r_{k-1}\right) r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}, \tag{6}$$

has the following convergence rate:

$$\mathbb{E}[f(w_T) - f^*] \leq 2c_3 \exp\left(-\frac{T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right)[f(w_0) - f^*] + \frac{8\sigma^2 c_4 \kappa}{\rho L e^2} \frac{(\ln(T/\beta))^2}{\alpha^2 T}$$

where $\kappa = \frac{L}{\mu}$, $c_3 = \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{2\beta}{\ln(T/\beta)}\right)$ and $c_4 = \exp\left(\frac{1}{\alpha\sqrt{\rho\kappa}} \frac{2\beta}{\ln(T/\beta)}\right)$.

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 Stich (2019); Khaled and Richtárik (2020); Li et al. (2020). 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) prove convergence to a neighbourhood of the solution, while we show convergence to the minimizer at a rate governed by the $O(\sigma^2/T)$ term. In the fully-deterministic setting ($\rho=1$ and $\sigma=0$), Theorem 1 implies an $\tilde{O}(\exp(-T/\sqrt{\kappa}))$ convergence to the minimizer, matching the optimal rate in the deterministic setting (Nesterov, 2004). We note that ASGD does not require knowledge of σ^2 and is thus completely noise-adaptive. To smoothly interpolate between the stochastic (mini-batch size 1) and fully deterministic (mini-batch size n) setting, we generalize the growth condition (and the above result) to show an explicit dependence on the mini-batch size in Appendix B.

Finally, we note that ASGD requires the knowledge of both μ and L and is thus not problem-adaptive. In the next section, we consider strategies towards achieving problem-adaptivity.

3.1 Misspecified ASGD

Computing the exact values of μ and L are a challenging problem and in practice we estimate them with some error. In this section we assume that we μ and L are misspecified by factor ν_μ and ν_L i.e. $\tilde{\mu}=\nu_\mu\mu$ and $\tilde{L}=\frac{L}{\nu_L}$. The following theorem shows the effect of this misspecification over the convergence of ASGD.

Theorem 2. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f and (iii) the growth condition in Eq. (5), ASGD (Eqs. (3) and (4)) with $w_0=y_0$, $\gamma_k=\frac{\nu_L}{\rho L}$, $\alpha=\left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k=\alpha^k$, $r_k=\sqrt{\frac{\nu_L\nu_\mu\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and b_k computed as:*

$$b_k=\frac{(1-r_{k-1})\,r_{k-1}\,\alpha}{r_k+r_{k-1}^2\,\alpha}, \tag{7}$$

has the following convergence rate:

$$\begin{aligned}&\mathbb{E}[f(w_T)-f^*]\leq\\&2c_3\exp\left(-\frac{\sqrt{\nu}T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_0)-f^*]\\&+\frac{8c_4\kappa(\ln(T/\beta))^2}{(e\alpha)^2\rho LT}\nu\sigma^2\\&+\frac{2c_4\kappa(\ln(T/\beta))^2}{(e\alpha)^2LT}\min\{\frac{\lfloor\ln(\nu)\rfloor_+}{\ln(T/\beta)},1\}\nu G^2\end{aligned}$$

where $\nu=\nu_L\nu_\mu$, $\kappa=\frac{L}{\mu}$, $c_3=\exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{2\beta\sqrt{\nu}}{\ln(T/\beta)}\right)$ and $c_4=\exp\left(\frac{1}{\alpha\sqrt{\rho\kappa}}\frac{2\beta\sqrt{\nu}}{\ln(T/\beta)}\right)$ and $G:=\max_{j\in[k_0]}\mathbb{E}[\|\nabla f(y_k)\|^2]$ with $k_0:=T\frac{\lfloor\ln(\nu)\rfloor_+}{\ln(T/\beta)}$ and $\lfloor x\rfloor_+=\max\{\lfloor x\rfloor,0\}$.

The above theorem implies that misspecified ASGD achives $\tilde{O}(\exp(\frac{-T\sqrt{\nu}}{\sqrt{\kappa\rho}})+\frac{\sigma^2\nu}{T}+\frac{G^2\nu}{T})$ such that comparing against ASGD rate it has an extra term due to misspecification. Assessing the upper bound of sub-optimality in Theorem 2, we note that: (i) the term concerning the noise and the misspecification are independent i.e. the second term of RHS just depends on σ and the third term just depends on G; (ii) the bound is both noise adaptive and misspecification adaptive i.e. if $\nu=1$ we recover the result of Theorem 1; (iii) and interestingly the bound depends on the multiplication of misspecifications of L and μ i.e. $\nu=\nu_L\nu_\mu$ and therefore we still can have misspecification i.e. $\nu_L,\nu_\mu\neq 1$ but $\nu=1$ and benefit the speed of ASGD. The proof of Theorem 2 is presented in Appendix G.

4 Towards noise and problem adaptivity

In this section, we consider approaches for achieving both noise and problem adaptivity when minimizing smooth, strongly-convex functions. In order to make progress towards this objective, we will only consider SGD and aim to obtain the non-accelerated noise-adaptive rate matching Stich (2019); Li et al. (2020); Khaled and Richtárik (2020), but do so without knowing problem-dependent constants.

For this section, we will consider a different weaker notion of noise in the stochastic gradients. Instead of using the growth condition in Eq. (5) or the more 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 redefine 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 $\frac{1}{2L}z^2 \leq \sigma^2 \leq \frac{1}{2\mu}z^2$.

We will continue to use exponentially decreasing step-sizes. As a warm-up towards problem-adaptivity, we first assume knowledge of the smoothness constant in Section 4.1 and analyze the resulting SGD algorithm. In Section 4.2, 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 4.3, we analyze the convergence of SGD when using an offline estimate of the smoothness.

4.1 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 D.1.

Theorem 3. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\gamma_k = \frac{1}{L}$, $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$, has the following convergence rate,*

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

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

Compared to Moulines and Bach (2011) that use a polynomially decreasing step-sizes, the proposed step-size results in a better trade-off between the bias (initial distance to the minimizer) and variance (noise) terms. We use exponentially decreasing step-sizes rather than the step-size schemes used in Stich (2019); Khaled and Richtárik (2020) because both of these also require the knowledge of the strong-convexity parameter which is considerably more difficult to estimate.

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, $\mathbb{E}\left\|w_{T+1} - w^*\right\|^2$. This change allows us to 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 constant and set the step-size on the fly.

4.2 Online estimation of unknown smoothness

In this section, we assume that the smoothness constant is unknown, aim to estimate it and set the step-size in an *online* fashion. By online estimation, we mean that in iteration k of SGD, we use 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). Recent methods based on a stochastic line-search (Vaswani et al., 2019b, 2020), stochastic Polyak step-size (Loizou et al., 2021; Berrada et al., 2020) or stochastic Barzilai-Borwein-like step-size (Malitsky and Mishchenko, 2019) are techniques that can set the step-size by only using the current sampled function.

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 each iteration k of SGD, 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 the following conditions: $\gamma_k \leq \gamma_{\max}$ and

$$f_{ik}(w_k - \gamma_k \nabla f_{ik}(w_k)) \leq f_{ik}(w_k) - c\gamma_k \left\|\nabla f_{ik}(w_k)\right\|^2. \tag{8}$$

Here, $c \in (0,1)$ is a hyper-parameter to be set according to the theory. 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 (see Lemma 11 for the proof). 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 suffices ($\alpha_k = 1$ for all k), and SGD obtains 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 3 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 D.2.

Theorem 4. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$ and γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and Eq. (8) with $c = 1/2$, has the following convergence rate,*

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

with $\kappa' = \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'} \cdot \frac{2\beta}{\ln(T/\beta)}\right)$.

We observe that the first two terms are similar to those in Theorem 3. 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 $\left(\gamma_{\max} - \min\left\{\gamma_{\max}, \frac{1}{L}\right\}\right)$, 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 3. In fact, if we can "guess" a value of $\gamma_{\max} \leq \frac{1}{L}$, it would result in the neighbourhood term becoming negative, 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 for SGD with SLS and $\alpha_k = 1$ (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.

In the next section, we prove a lower-bound that shows the necessity of such a neighbourhood term.

4.2.1 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 $\gamma_k \leq 1/L_{ik}$[2] (see Lemma 11). With these choices, we prove the following lower-bound in Appendix E.1.

Theorem 5. *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 the step-size 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 shows that using SGD with SLS to set γ_k and *any convergent sequence of* α_k (including the exponentially-decreasing sequence in Theorem 4) 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. This result is not specific to SLS and would hold for other methods (Loizou et al., 2021; Berrada et al., 2020; Malitsky and Mishchenko, 2019) that set γ_k in an online fashion. Since the lower-bound holds for any convergent α_k sequence, a possible reason for convergence to the neighbourhood is the correlation between i_k (the sampled function) and the computation of γ_k. We verify this hypothesis in the next section.

4.3 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 of SGD, γ_k is set *before* sampling i_k and cannot

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

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 the estimation error in γ with $\nu = 1$ corresponding to an exact estimation of L. In practice, it is typical to be able 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 D.3.

Theorem 6. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$ and $\gamma_k = \frac{\nu}{L}$ for $\nu > 0$ has the following convergence rate,*

$$\|w_{T+1} - w^*\|^2 \leq \|w_1 - w^*\|^2 \; c_2 \exp\left(-\frac{\nu}{\kappa}\frac{T}{\ln(T/\beta)}\frac{\alpha}{}\right) + \frac{8\sigma^2}{LT}\exp\left(\frac{\nu}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)\frac{4\kappa^2\ln(T/\beta)^2}{e^2\alpha^2} + \left[\max_{j\in\left[T\frac{[\ln(\nu)]_+}{\ln(T/\beta)}\right]}\{f(w_j) - f^*\}\frac{(\nu-1)}{L}\right] \times \exp\left(\frac{\nu}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)\frac{8\kappa^2}{\nu e^2\alpha^2}\frac{[\ln(\nu)]_+\ln(T/\beta)}{T},$$

where $\kappa = \frac{L}{\mu}$, $c_2 = \exp\left(\frac{\nu}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)$ and $[x]_+ = \max\{x, 0\}$.

Observe that as $T \to \infty$, SGD converges to the minimizer w^*. The first two terms are similar to that in Theorem 3 and imply an $O\left(\exp(-T) + \frac{\sigma^2}{T}\right)$ convergence to the minimizer. The third term can be viewed as the *price of misestimation* of the unknown smoothness constant. Unlike in Theorem 4 where this price was convergence to a neighbourhood, here, the price of misestimation is slower 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 3 up to constants that depend on ν. For $\nu > 1$, the convergence rate slows down by a factor that depends on ν. The third term depends on $\left[\max_{j\in[T\,[\ln(\nu)]_+/\ln(T/\beta)]}\{f(w_j) - f^*\}\right]$ because if $\nu > 1$, SGD can diverge and move away from the solution for the initial $T\frac{[\ln(\nu)]_+}{\ln(T/\beta)}$ iterations. This can be explained as follows: for $\nu > 1$, the step-size $\gamma_k = \frac{\nu}{L}\alpha_k \geq \frac{1}{L}$ initially, and SGD diverges in this regime. However, since α_k is an exponentially decreasing sequence, after $k_0 := T\frac{\ln(\nu)}{\ln(T/\beta)}$ iterations, $\frac{\nu}{L}\alpha_k \leq \frac{1}{L}$, and the distance to the minimizer decreases after iteration k_0, eventually resulting in convergence to the solution.

Finally, observe that the third term depends on $O\left(\exp(\nu)[\ln(v)]_+\right)$ meaning that if we misestimate the smoothness constant by a multiplicative factor of ν, it can slow down the convergence rate by a factor exponential in ν. In the next section, we justify this dependence by proving a corresponding lower-bound.

4.3.1 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 $\nu > 1$. For simplicity, we consider minimizing a single quadratic, thus ensuring $\sigma^2 = 0$. We prove the following lower-bound in Appendix E.2.

Theorem 7. *When using gradient descent to minimize a one-dimensional quadratic function $f(w) = \frac{1}{2}(xw - y)^2$, with $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$ and $\gamma_k = \frac{\nu}{L}$ for $\nu > 3$ we have*

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

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

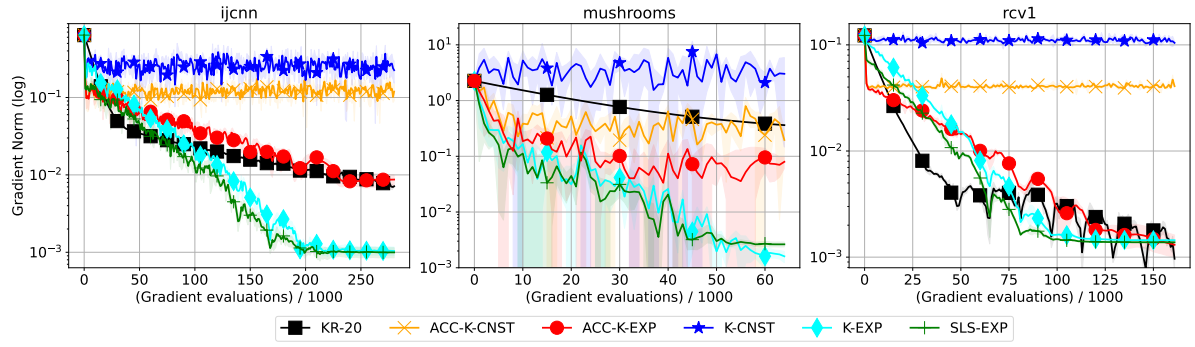
$$|w_{k'+1} - w^*| \geq 2^{k'}|w_1 - w^*|.$$

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 thus shows the (potentially exponential) slowdown in the rate of convergence by misestimating the smoothness.

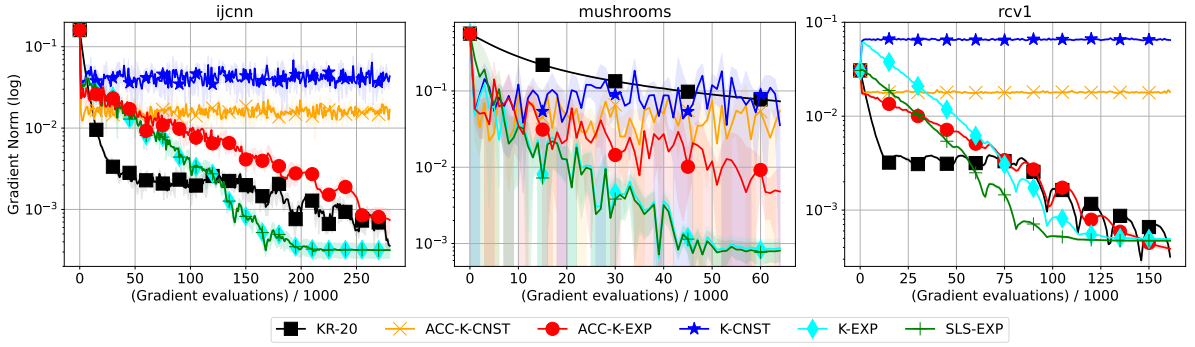
In the next section, we design a variant of SLS that ensures convergence to the minimizer while obtaining good empirical control over the misestimation.

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.



(a) Squared loss



(b) Logistic loss

Figure 1: Comparison of step-size strategies for (a) squared loss and (b) logistic loss. Observe that (i) exponentially decreasing step-sizes result in more stable performance compared to using a constant step-size (for both SGD and ASGD) and (ii) consistently outperform the noise-adaptive method in (Khaled and Richtárik, 2020) and (iii) the stochastic line-search in Eq. (9) matches the performance of the variant with known smoothness.

With 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`) and (v) ASGD with exponentially decreasing step-sizes, analyzed in Section 3 and denoted as `ACC-K-EXP`.

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. (5) and μ, the strong-convexity parameter. 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. To set ρ, we use a grid search over $\{10, 100, 1000\}$ and plot the variant that results in the smallest gradient norm.

Using a stochastic line-search (SLS) can result in convergence to the neighbourhood (Section 4.2) 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, but requires computing the gradient of two functions - one for the update and the other for the line-search. The overall procedure can be described as follows: starting with a backtracking line-search from γ_{k-1} (the conservative aspect) (with $\gamma_0 = \gamma_{\max}$) for a random or previously sampled functio n (j_k), find the largest step-size γ_k that satisfies

$$f_{j_k}(w_k - \gamma_k \nabla f_{j_k}(w_k)) \leq f_{j_k}(w_k) - c\gamma_k \left\|\nabla f_{j_k}(w_k)\right\|^2, \tag{9}$$

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 γ_k is fixed before computing $\nabla f_{ik}(w_k)$, this strategy can be analyzed using the framework in Section 4.3. Specifically, it results in $\gamma_k = \frac{\nu_k}{L}$ for some sequence of $\nu_k \geq 1$. The conservative aspect ensures that $\nu_k \leq \nu_{k-1}$. Hence, the convergence rate can be analyzed according to Theorem 6 with $\nu = \nu_1$ and the initial line-search controlling the misestimation error. We use this variant of SLS with exponentially decreasing step-sizes and denote it as `SLS-EXP` in the plots. 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 method in (Khaled and Richtárik, 2020). We also observe that (iii) the stochastic line-search in Eq. (9) (`SLS-EXP`) matches the performance of the variant with known smoothness (`K-EXP`) and (iv) ASGD does not result in improvements over SGD. This is because these methods are quite sensitive to their parameter values and we set these parameters by using loose theoretical upper-bounds on both L and μ.

6 Conclusion

In this paper, we first developed a variant of SGD with Nesterov acceleration and exponentially decreasing step-sizes, and proved that it achieves the near-optimal convergence rate in both the deterministic and stochastic regimes. We then considered two strategies for making SGD both noise-adaptive and problem-adaptive. Using upper and lower-bounds, we showed that there is always a price to pay for problem-adaptivity – estimating the smoothness constant 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 empirically demonstrated the effectiveness of a noise-adaptive, problem-adaptive method that uses exponential step-sizes coupled with a novel variant of stochastic line-search. In the future, we hope to develop a problem-adaptive variant of ASGD.

7 Acknowledgements

We would like to thank Si Yi Meng for helpful feedback on the paper. Benjamin Dubois-Taine would like to acknowledge funding by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

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

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.

Gower, R., Sebbouh, O., and Loizou, N. (2021). Sgd for structured nonconvex functions: Learning rates, minibatching 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.

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.

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*.

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*.

Ma, S., Bassily, R., and Belkin, M. (2018). The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML*.

Malitsky, Y. and Mishchenko, K. (2019). Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*.

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.

Supplementary material

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,

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

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,

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

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

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

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^*,

$$f(w^*) \geq f(w) + \frac{1}{\zeta}\langle \nabla f(w), w^* - w \rangle + \frac{\mu}{2}\left\|w - w^*\right\|^2. \tag{10}$$

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 3 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 4 and Theorem 6. 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 D.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,*

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\left\|w_1 - w^*\right\|^2}{\sum_{k=1}^T \alpha_k} + \sigma^2 \frac{\sum_{k=1}^T \alpha_k^2}{\sum_{k=1}^T \alpha_k} \tag{11}$$

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,

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

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. (11) 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 12 and 13, 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 3 and a different definition of σ for the SGD proofs in Section 4. We refine these quantities here, and show the explicit dependence on the mini-batch size.

Lemma 1. *If*

$$\mathbb{E}_i \left\|\nabla f_i(w)\right\|^2 \leq \rho \left\|\nabla f(w)\right\|^2 + \sigma^2,$$

then,

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

Proof.

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

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

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

$$\begin{aligned}
&\leq \frac{n-B}{nB}\left(E_i \left\|\nabla f_i(w)\right\|^2 - \left\|\nabla f(w)\right\|^2\right) + \left\|\nabla f(w)\right\|^2 \\
&\leq \frac{n-B}{nB}\left((\rho-1)\left\|\nabla f(w)\right\|^2 + \sigma^2\right) + \left\|\nabla f(w)\right\|^2 \qquad \text{(Using the growth condition)} \\
\implies \mathbb{E}_{\mathcal{B}} \left\|\nabla f_{\mathcal{B}}(w))\right\|^2 &\leq \left((\rho-1)\frac{n-B}{nB}+1\right)\left\|\nabla f(w)\right\|^2 + \frac{n-B}{nB}\sigma^2.
\end{aligned}$$

□

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,

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

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

$$\begin{aligned} &\leq \frac{1}{2\mu}\frac{n-B}{nB}\mathbb{E}_i \left\|\nabla f_i(w^*)\right\|^2 \\ &\leq \frac{L}{\mu}\frac{n-B}{nB}\mathbb{E}[f_i(w^*) - f_i^*] && \text{(By smoothness of } f_i\text{)} \\ \implies \sigma_B^2 &\leq \frac{L}{\mu}\frac{n-B}{nB}\sigma^2. \end{aligned}$$

□

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

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

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

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

Using the bound in the proof of Lemma 12, we have

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

Putting this together we have that

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

Putting everything together we get that

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

□

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 $f_2(w) = \frac{1}{2}(2w + 1/2)^2$ with a constant step-size $\eta = \frac{1}{L}$, the following holds: whenever $|w_k - w^*| < 1/8$, the next iterate satisfies $|w_{k+1} - w^*| > 1/8$.*

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

$$\text{If } i_k = 1: \quad w_{k+1} = w_k - \eta(w_k - 1) = w_k(1 - \frac{1}{4}) + \frac{1}{4} = \frac{3}{4}w_k + \frac{1}{4}$$
$$\text{If } i_k = 2: \quad w_{k+1} = w_k - \eta 2(2w_k + \frac{1}{2}) = w_k(1 - \frac{4}{4}) - \frac{1}{4} = -\frac{1}{4}$$

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

C Proof for ASGD

C.1 Reformulation

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

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

$$b_k = \frac{(1-r_{k-1})r_{k-1}\frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2\frac{\eta_k}{\eta_{k-1}}}, \tag{13}$$

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. (12).

We first show that the update in Eq. (3)-Eq. (4) satisfying the conditions in Eq. (13) and Eq. (12) can be written in an equivalent form more amenable to the analysis.

Lemma 5. *The following update:*

$$y_k = w_k - \frac{r_k q_k}{q_k + r_k\mu}(w_k - z_k) \tag{14}$$

$$w_{k+1} = y_k - \eta_k\nabla f_{ik}(y_k) \tag{15}$$

$$z_{k+1} = w_k + \frac{1}{r_k}[w_{k+1} - w_k] \tag{16}$$

where,

$$q_{k+1} = (1-r_k)q_k + r_k\mu \tag{17}$$

$$r_k^2 = q_{k+1}\eta_k \tag{18}$$

$$z_{k+1} = \frac{1}{q_{k+1}}\left[(1-r_k)q_k z_k + r_k\mu y_k - r_k\nabla f_{ik}(y_k)\right] \tag{19}$$

is equivalent to the update in Eq. (3)-Eq. (4).

Proof.

First we check the consistency of the update (Eq. (16)) and definition (Eq. (19)) of z_k. Using Eq. (19),

$$\begin{aligned}
z_{k+1} &= \frac{1}{q_{k+1}}\left[(1-r_k)q_k z_k + r_k\mu y_k - r_k\nabla f_{ik}(y_k)\right] \\
&= -\frac{(1-r_k)}{r_k}w_k - \frac{r_k}{q_{k+1}}\nabla f_{ik}(y_k) + y_k\left[\frac{(1-r_k)(q_k+r_k\mu)}{q_{k+1}r_k} + \frac{r_k\mu}{q_{k+1}}\right] \\
&= -\frac{(1-r_k)}{r_k}w_k - \frac{r_k}{q_{k+1}}\nabla f_{ik}(y_k) + y_k\left[\frac{q_k(1-r_k) + (r_k\mu - r_k^2\mu)}{q_{k+1}r_k} + \frac{r_k^2\mu}{q_{k+1}r_k}\right] \\
&= -\frac{(1-r_k)}{r_k}w_k - \frac{r_k}{q_{k+1}}\nabla f_{ik}(y_k) + y_k\left[\frac{(q_{k+1} - r_k\mu) + (r_k\mu - r_k^2\mu) + r_k^2\mu}{q_{k+1}r_k}\right] && \text{(From Eq. (17))} \\
&= w_k - \frac{w_k}{r_k} + \frac{1}{r_k}\left[y_k - \eta_k\nabla f_{ik}(y_k)\right] && \text{(From Eq. (18))} \\
z_{k+1} &= w_k + \frac{1}{r_k}\left[w_{k+1} - w_k\right] && \text{(From Eq. (15))}
\end{aligned}$$

which recovers Eq. (16) showing that the definition of z_k and its update is consistent.

Now we check the equivalence of Eq. (12) and Eq. (17)-Eq. (18). Eliminating q_k using Eq. (17)-Eq. (18),

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

Multiplying by η_k recovers Eq. (12).

Since Eq. (4) and Eq. (15) are equivalent, we need to establish the equivalence of Eq. (3) and the updates in Eq. (14)-Eq. (16). From Eq. (16)

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

Starting from Eq. (14) and using the above relation to eliminate z_k,

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

which is in the same form as Eq. (3). 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. (17) and Eq. (18),

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

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. (18), 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. (3) and completes the proof.

□

C.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

$$\lambda_0 = 1 \quad ; \quad \lambda_{k+1} = (1 - r_k)\lambda_k, \tag{20}$$

$$\phi_k(w) = [\inf_w \phi_k(w)] + \frac{q_k}{2}\left\|w - z_k\right\|^2, \tag{21}$$

and satisfies the following update condition

$$\phi_k(w) \leq (1 - \lambda_k) f(w) + \lambda_k \phi_0(w) \tag{22}$$

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

$$\phi_{k+1}^* = (1 - r_k)\phi_k^* + r_k\left[f(y_k) - \frac{r_k}{2q_{k+1}}\left\|\nabla f(y_k)\right\|^2 + \frac{(1 - r_k)\, q_k}{q_{k+1}}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle \nabla f(y_k), z_k - y_k\rangle\right)\right] \tag{23}$$

Finally note that the definition of ϕ_k can be used to rewrite Eq. (14) as

$$y_k = w_k - \frac{r_k}{q_k + r_k \mu}\nabla \phi_k(w_k). \tag{24}$$

C.3 Proof of Theorem 1

Given the definitions in Appendix C.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 6. *Using the update in Eq. (15) with $\eta_k = \frac{1}{\rho L}\alpha_k$, we obtain the following descent condition.*

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

Proof. By smoothness, and the update in Eq. (15),

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

Taking expectation w.r.t. i_k,

$$\begin{aligned}
\mathbb{E}[f(w_{k+1})] &\leq \mathbb{E}[f(y_k)] - \eta_k\left\|\nabla f(y_k)\right\|^2 + \frac{L}{2}\eta_k^2\mathbb{E}[\left\|\nabla f_{ik}(y_k)\right\|^2] && (\eta_k \text{ is independent of the randomness in } i_k.)\\
&\leq \mathbb{E}[f(y_k)] - \eta_k\left\|\nabla f(y_k)\right\|^2 + \frac{\rho L}{2}\eta_k^2\mathbb{E}[\left\|\nabla f(y_k)\right\|^2] + \frac{L}{2}\eta_k^2\sigma^2 && (\text{By the growth condition in Eq. (5)})\\
&= \mathbb{E}[f(y_k)] - \eta_k\left\|\nabla f(y_k)\right\|^2 + \frac{\eta_k\alpha_k}{2}\mathbb{E}[\left\|\nabla f(y_k)\right\|^2] + \frac{1}{2\rho^2 L}\alpha_k^2\sigma^2\\
&\leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\left\|\nabla f(y_k)\right\|^2 + \frac{1}{2\rho^2 L}\alpha_k^2\sigma^2
\end{aligned}$$

□

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 7. *For the estimating sequences defined in Appendix C.2 and the updates in Eq. (14)-Eq. (19), for all k,*

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

where $\mathcal{N}_k := \frac{2\sigma^2}{\rho^2 L}\sum_{j=0}^{k-1}\alpha_j^2\prod_{i=j+1}^{k-1}(1-r_i)$.

Proof. We will prove the lemma by induction. For $k=0$, we define $\phi_0^* = f(w_0)$, and since $\mathcal{N}_k \geq 0$ for all k, $\mathbb{E}[\phi^*] \geq f(w_0) - \mathcal{N}_0$, 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. (23) to prove the statement for $k+1$ as follows. Taking expectations w.r.t to the randomness in $j=1$ to k,

$$\begin{aligned}
\mathbb{E}[\phi_{k+1}^*] &= (1-r_k)\mathbb{E}[\phi_k^*] + r_k\mathbb{E}\left[f(y_k) - \frac{r_k}{2q_{k+1}}\left\|\nabla f(y_k)\right\|^2 + \frac{(1-r_k)\,q_k}{q_{k+1}}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)\right]\\
&\geq (1-r_k)\mathbb{E}[f(w_k) - \mathcal{N}_k] + r_k\mathbb{E}\left[f(y_k) - \frac{r_k}{2q_{k+1}}\left\|\nabla f(y_k)\right\|^2 + \frac{(1-r_k)\,q_k}{q_{k+1}}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)\right]\\
&\qquad\qquad (\text{by the induction hypothesis})\\
&= (1-r_k)\mathbb{E}[f(w_k)] + r_k\mathbb{E}[f(y_k)] - \frac{r_k^2}{2q_{k+1}}\mathbb{E}\left\|\nabla f(y_k)\right\|^2 + \frac{r_k(1-r_k)\,q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right)\\
&\quad - (1-r_k)\mathcal{N}_k\\
&= (1-r_k)\mathbb{E}[f(w_k)] + r_k\mathbb{E}[f(y_k)] - \frac{\eta_k}{2}\mathbb{E}\left\|\nabla f(y_k)\right\|^2\\
&\quad + \frac{r_k(1-r_k)\,q_k}{q_{k+1}}\mathbb{E}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle\nabla f(y_k), z_k - y_k\rangle\right) - (1-r_k)\mathcal{N}_k && (\text{Using Eq. (18)})
\end{aligned}$$

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

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

By Lemma 6,

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

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

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

Now we show that the remaining terms $(1-r_k)\mathbb{E}\left[\langle \nabla f(y_k), w_k - y_k\rangle + \frac{r_k\,q_k}{q_{k+1}}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle \nabla f(y_k), z_k - y_k\rangle\right)\right] \geq 0$. For this, we use Eq. (14)

$$\begin{aligned}
y_k &= w_k - \frac{q_k r_k}{q_k + r_k\mu}(w_k - z_k) \\
\implies z_k - y_k &= z_k - w_k + \frac{q_k r_k}{q_k + r_k\mu}(w_k - z_k) = \left(1 - \frac{q_k r_k}{q_k + r_k\mu}\right)(z_k - w_k) \\
&= \left(\frac{q_k(1-r_k) + r_k\mu}{q_k + r_k\mu}\right)(z_k - w_k) = \left(\frac{q_{k+1}}{q_k + r_k\mu}\right)(z_k - w_k) \qquad \text{(By Eq. (17))} \\
\implies \frac{r_k q_k}{q_{k+1}}\langle \nabla f(y_k), z_k - y_k\rangle &= \left\langle \nabla f(y_k), \left(-\frac{r_k\,q_k}{q_k + r_k\mu}\right)(w_k - z_k)\right\rangle = \langle \nabla f(y_k), y_k - w_k\rangle
\end{aligned}$$

Using this relation to simplify,

$$\begin{aligned}
&(1-r_k)\mathbb{E}\left[\langle \nabla f(y_k), w_k - y_k\rangle + \frac{r_k\,q_k}{q_{k+1}}\left(\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + \langle \nabla f(y_k), z_k - y_k\rangle\right)\right] \\
&= (1-r_k)\mathbb{E}\left[\frac{r_k\,q_k}{q_{k+1}}\frac{\mu}{2}\left\|y_k - z_k\right\|^2 + (1-r_k)\left[\langle \nabla f(y_k), w_k - y_k\rangle + \langle \nabla f(y_k), y_k - w_k\rangle\right]\right] \\
&= (1-r_k)\mathbb{E}\left[\frac{r_k\,q_k}{q_{k+1}}\frac{\mu}{2}\left\|y_k - z_k\right\|^2\right] \geq 0 \qquad \text{(Since } r_k \leq 1.\text{)}
\end{aligned}$$

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. □

We now use the above lemma to prove the rate for strongly-convex functions.

Theorem 1. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f and (iii) the growth condition in Eq. (5), ASGD (Eqs. (3) and (4)) with $w_0 = y_0$, $\gamma_k = \frac{1}{\rho L}$, $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$, $r_k = \sqrt{\frac{\mu}{\rho L}}\left(\frac{\beta}{T}\right)^{k/2T}$ and b_k computed as:*

$$b_k = \frac{(1 - r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}, \tag{6}$$

has the following convergence rate:

$$\mathbb{E}[f(w_T) - f^*] \leq 2c_3 \exp\left(-\frac{T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right)[f(w_0) - f^*] + \frac{8\sigma^2 c_4 \kappa}{\rho L e^2} \frac{(\ln(T/\beta))^2}{\alpha^2 T}$$

where $\kappa = \frac{L}{\mu}$, $c_3 = \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{2\beta}{\ln(T/\beta)}\right)$ and $c_4 = \exp\left(\frac{1}{\alpha\sqrt{\rho\kappa}} \frac{2\beta}{\ln(T/\beta)}\right)$.

Proof. Using the reformulation in Lemma 5 gives us $q_k = \mu$ for all k and $z_0 = w_0$. For the estimating sequences defined in Appendix C.2, using Lemma 18, we know that the (reformulated) updates satisfy the following relation,

$$\mathbb{E}[f(w_T)] \leq \mathbb{E}[\phi_T^*] + \mathcal{N}_T \leq \mathbb{E}[\phi_T(w^*)] + \mathcal{N}_T$$

From Eq. (22), 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,

$$\begin{aligned} &\mathbb{E}[f(w_T)] \leq (1 - \lambda_T) f^* + \lambda_T \phi_0(w^*) + \mathcal{N}_T \\ \implies &\mathbb{E}[f(w_T) - f^*] \leq \lambda_T \left[\phi_0(w^*) - f^*\right] + \mathcal{N}_T \end{aligned}$$

By Eq. (21),

$$\leq \lambda_T \left[\phi_0^* + \frac{q_0}{2} \|w^* - z_0\|^2 - f^*\right] + \mathcal{N}_T$$

Choosing $\phi_0^* = f(w_0)$,

$$\leq \lambda_T \left[f(w_0) - f^* + \frac{q_0}{2} \|w^* - z_0\|^2\right] + \mathcal{N}_T$$

Since $z_0 = w_0$, $q_0 = \mu$,

$$\implies \mathbb{E}[f(w_T) - f^*] \leq \lambda_T \left[f(w_0) - f^* + \frac{\mu}{2} \|w^* - w_0\|^2\right] + \frac{2\sigma^2}{\rho^2 L} \sum_{j=0}^{T-1} \alpha_j^2 \prod_{i=j+1}^{T-1} (1 - r_i)$$

Using the fact that $\lambda_0 = 1$ and $\lambda_{k+1} = (1 - r_k)\lambda_k$, we know that that $\lambda_T = \prod_{k=1}^{T}(1 - r_k)$, and

$$\mathbb{E}[f(w_T) - f^*] \leq \left[\prod_{k=1}^{T}(1 - r_k)\right] \left[f(w_0) - f^* + \frac{\mu}{2} \|w^* - w_0\|^2\right] + \frac{2\sigma^2}{\rho^2 L} \sum_{j=0}^{T-1} \alpha_j^2 \prod_{i=j+1}^{T-1} (1 - r_i).$$

Now our task is to upper-bound bound the $1 - r_k$ terms. From Eq. (18), 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,

$$\mathbb{E}[f(w_T) - f^*] \leq \left[\prod_{k=1}^{T}\left(1 - \sqrt{\frac{1}{\rho\kappa}}\alpha_k\right)\right]\left[f(w_0) - f^* + \frac{\mu}{2}\left\|w^* - w_0\right\|^2\right] + \frac{2\sigma^2}{\rho^2 L}\sum_{j=0}^{T-1}\alpha_j^2\prod_{i=j+1}^{T-1}\left(1 - \sqrt{\frac{1}{\rho\kappa}}\alpha_i\right)$$

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

$$\Delta_T \leq 2\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\sum_{k=1}^{T}\alpha^k\right)\Delta_0 + \frac{2\sigma^2}{\rho^2 L}\sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\sum_{i=k+1}^{T-1}\alpha^i\right)$$

Using Lemma 8, we can bound the first term as

$$\begin{aligned}
2\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\sum_{k=1}^{T}\alpha^k\right)\Delta_0 &\leq 2\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\left(\frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}\right)\right)\Delta_0 \\
&= 2c_3\exp\left(-\frac{T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_0) - f^*]
\end{aligned}$$

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 9. Indeed we have

$$\begin{aligned}
\sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\sum_{i=k+1}^{T-1}\alpha^i\right) &= \sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\frac{\alpha^{k+1} - \alpha^T}{1-\alpha}\right) \\
&= \exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\frac{\alpha^{k+1}}{1-\alpha}\right) \\
&\leq \exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\sum_{k=0}^{T-1}\alpha^{2k}\left(\frac{2(1-\alpha)\sqrt{\rho\kappa}}{e\alpha^{k+1}}\right)^2 \qquad \text{(Lemma 15)} \\
&= \exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa}{e^2\alpha^2}T(1-\alpha)^2 \\
&\leq \exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa}{e^2\alpha^2}T\ln(1/\alpha)^2 \\
&= \exp\left(\frac{1}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2}{e^2\alpha^2 T}
\end{aligned}$$

Finally,

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

where the inequality comes from the bound in Eq. (25) in the proof of Lemma 8. Putting everything together we obtain

$$\mathbb{E}[f(w_T) - f^*] \leq 2c_3\exp\left(-\frac{T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_0) - f^*] + \frac{8\sigma^2 c_4\kappa\ln(T/\beta)^2}{\rho L e^2\alpha^2 T}$$

where $c_4 = \exp\left(\frac{2\beta}{\alpha\sqrt{\rho\kappa}\ln(T/\beta)}\right)$. □

C.4 Lemmas for acceleration proofs

Lemma 8.

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

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

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha\beta}{T(1-\alpha)} = \frac{\beta}{T} \cdot \frac{1}{1/\alpha - 1} \leq \frac{\beta}{T} \cdot \frac{2}{\ln(1/\alpha)} = \frac{\beta}{T} \cdot \frac{2}{\frac{1}{T}\ln(T/\beta)} = \frac{2\beta}{\ln(T/\beta)} \tag{25}$$

where in the inequality we used Lemma 14 and the fact that $1/\alpha > 1$. Plugging back into A we get,

$$\begin{aligned} A &\geq \frac{\alpha}{1-\alpha} - \frac{2\beta}{\ln(T/\beta)} \\ &\geq \frac{\alpha}{\ln(1/\alpha)} - \frac{2\beta}{\ln(T/\beta)} && (1 - x \leq \ln(\tfrac{1}{x})) \\ &= \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)} \end{aligned}$$

□

Lemma 9. *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 14 and the fact that $1/\alpha > 1$. These relations imply that,

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

We then have

$$\begin{aligned}
\sum_{k=1}^{T} \alpha^{2k} \exp\left(-\frac{1}{\kappa} \sum_{i=k+1}^{T} \alpha^i\right) &\leq c_2 \sum_{k=1}^{T} \alpha^{2k} \exp\left(-\frac{1}{\kappa} \frac{\alpha^{k+1}}{1-\alpha}\right) \\
&\leq c_2 \sum_{k=1}^{T} \alpha^{2k} \left(\frac{2(1-\alpha)\kappa}{e\alpha^{k+1}}\right)^2 && \text{(Lemma 15)} \\
&= \frac{4\kappa^2 c_2}{e^2\alpha^2} T(1-\alpha)^2 \\
&\leq \frac{4\kappa^2 c_2}{e^2\alpha^2} T(\ln(1/\alpha))^2 \\
&= \frac{4\kappa^2 c_2 (\ln(T/\beta))^2}{e^2\alpha^2 T}
\end{aligned}$$

□

D Upper-bound Proofs for Section 4

D.1 Proof of Theorem 3

Theorem 3. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\gamma_k = \frac{1}{L}$, $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$, has the following convergence rate,*

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

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

Proof.

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

Taking expectation w.r.t i_k,

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

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

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

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

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

Using Lemma 8 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 9 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,

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

□

D.2 Proof of Theorem 4

Theorem 4. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$ and γ_k as the largest step-size that satisfies $\gamma_k \leq \gamma_{\max}$ and Eq. (8) with $c = 1/2$, has the following convergence rate,*

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

with $\kappa' = \max\left\{\frac{L}{\mu}, \frac{1}{\mu\gamma_{\max}}\right\}$, $c_1 = \exp\left(\frac{1}{\kappa'}\cdot\frac{2\beta}{\ln(T/\beta)}\right)$.

Proof.

$$\left\|w_{k+1} - w^*\right\|^2 \leq \left\|w_k - w^*\right\|^2 - 2\gamma_k\alpha_k\langle\nabla f_{ik}(w_k), w_k - w^*\rangle + \gamma_k\alpha_k^2\left[\frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \qquad \text{(By Lemma 11)}$$

Setting $c = 1/2$,

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

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

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

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,

$$\begin{aligned}\mathbb{E}\left\|w_{k+1} - w^*\right\|^2 &\leq \left\|w_k - w^*\right\|^2 + 2\gamma_{\min}\alpha_k\left[-\langle\nabla f(w_k), w_k - w^*\rangle + [f(w_k) - f(w^*)]\right] - (\alpha_k - \alpha_k^2)\mathbb{E}\left[2\gamma_k[f_{ik}(w_k) - f_{ik}(w^*)]\right] \\ &+ 2\gamma_{\max}\alpha_k^2\sigma^2 \\ \mathbb{E}\left\|w_k - w^*\right\|^2 &\leq (1 - \alpha_k\gamma_{\min}\mu)\left\|w_k - w^*\right\|^2 - (\alpha_k - \alpha_k^2)\mathbb{E}\left[2\gamma_k[f_{ik}(w_k) - f_{ik}(w^*)]\right] + 2\gamma_{\max}\alpha_k^2\sigma^2\end{aligned}$$

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

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

Putting this relation back,

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

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

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

Using Lemma 8 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,

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

where $c_1 = \exp\left(\frac{1}{\kappa'}\frac{2\beta}{\ln(T/\beta)}\right)$. Using Lemma 9 to upper-bound B_t, we obtain $B_t \leq \frac{4(\kappa')^2c_1(\ln(T/\beta))^2}{e^2\alpha^2T}$, thus bounding the second term. Using Lemma 10 to upper-bound C_t, we obtain $C_t \leq c_1\frac{\kappa'\ln(T/\beta)}{e\alpha}$, thus bounding the third term. Finally, by Lemma 11 we have that $\gamma_{\min} \geq \min\left\{\gamma_{\max}, \frac{1}{L}\right\}$.

Putting everything together,

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

□

D.3 Proof of Theorem 6

Theorem 6. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f, SGD (Eq. (2)) with $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$ and $\gamma_k = \frac{\nu}{L}$ for $\nu > 0$ has the following convergence rate,*

$$\begin{aligned}
&\|w_{T+1} - w^*\|^2 \leq \|w_1 - w^*\|^2 \; c_2 \exp\left(-\frac{\nu T}{\kappa} \frac{\alpha}{\ln(T/\beta)}\right) \\
&+ \frac{8\sigma^2}{LT} \exp\left(\frac{\nu}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right) \frac{4\kappa^2 \ln(T/\beta)^2}{e^2\alpha^2} \\
&+ \left[\max_{j \in \left[T \frac{[\ln(\nu)]_+}{\ln(T/\beta)}\right]} \{f(w_j) - f^*\} \frac{(\nu - 1)}{L}\right] \\
&\times \exp\left(\frac{\nu}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right) \frac{8\kappa^2}{\nu e^2 \alpha^2} \frac{[\ln(\nu)]_+ \ln(T/\beta)}{T},
\end{aligned}$$

where $\kappa = \frac{L}{\mu}$, $c_2 = \exp\left(\frac{\nu}{\kappa} \frac{2\beta}{\ln(T/\beta)}\right)$ and $[x]_+ = \max\{x, 0\}$.

Proof. Following the steps from the proof of Theorem 3,

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

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

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

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

(By strong convexity)

Since $\gamma_k = \frac{\nu}{L}$ for some $\nu \geq 1$, we require $\alpha_k \leq \frac{1}{\nu}$ for the last term to be negative. 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 $(2L\gamma_k^2\alpha_k^2 - 2\gamma_k\alpha_k) \leq \frac{2\nu\,(\nu-1)}{L}\alpha_k^2$, meaning that for $k < k_0$,

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

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

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

Bounding the first term similar to Lemma 8,

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

Bounding the second term similar to Lemma 9,

$$\begin{aligned}\sum_{k=1}^{k_0-1}\alpha_k^2\prod_{i=k+1}^{k_0-1}(1-\frac{\mu\nu}{L}\alpha_i) &\leq \sum_{k=1}^{k_0-1}\alpha_k^2\exp\left(-\frac{\mu\nu}{L}\sum_{i=k+1}^{k_0-1}\alpha^i\right)\\ &= \sum_{k=1}^{k_0-1}\alpha_k^2\exp\left(-\frac{\nu}{\kappa}\frac{\alpha^{k+1}-\alpha^{k_0}}{1-\alpha}\right)\\ &= \exp\left(\frac{\nu\alpha^{k_0}}{\kappa(1-\alpha)}\right)\sum_{k=1}^{k_0-1}\alpha_k^2\exp\left(-\frac{\nu\alpha^{k+1}}{\kappa(1-\alpha)}\right)\\ &\leq \exp\left(\frac{\nu\alpha^{k_0}}{\kappa(1-\alpha)}\right)\sum_{k=1}^{k_0-1}\alpha_k^2\left(\frac{2(1-\alpha)\kappa}{\nu e\alpha^{k+1}}\right)^2\\ &\leq \exp\left(\frac{\nu\alpha^{k_0}}{\kappa(1-\alpha)}\right)\frac{4(1-\alpha)^2\kappa^2}{\nu^2e^2\alpha^2}k_0\\ &\leq \exp\left(\frac{\nu\alpha^{k_0}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{k_0\ln(T/\beta)^2}{T^2}\end{aligned}$$

Putting everything together, we obtain,

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

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

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

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

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

Bounding the first term similar to Lemma 8,

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

Bounding the second term similar to Lemma 9,

$$\begin{aligned}\sum_{k=k_0}^{T}\alpha_k^2\prod_{i=k+1}^{T}(1-\frac{\mu\nu}{L}\alpha_i) &\leq \sum_{k=k_0}^{T}\alpha_k^2\exp\left(-\frac{\mu\nu}{L}\sum_{i=k+1}^{T}\alpha^i\right)\\ &= \sum_{k=k_0}^{T}\alpha_k^2\exp\left(-\frac{\nu}{\kappa}\frac{\alpha^{k+1}-\alpha^{T+1}}{1-\alpha}\right)\\ &= \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right)\sum_{k=k_0}^{T}\alpha_k^2\exp\left(-\frac{\nu\alpha^{k+1}}{\kappa(1-\alpha)}\right)\\ &\leq \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right)\sum_{k=k_0}^{T}\alpha_k^2\left(\frac{2(1-\alpha)\kappa}{\nu e\alpha^{k+1}}\right)^2\\ &= \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4(1-\alpha)^2\kappa^2}{\nu^2e^2\alpha^2}(T-k_0+1)\\ &\leq \exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}\end{aligned}$$

Putting everything together,

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

Combining the above bounds, we get,

$$\begin{aligned}\Delta_{T+1} \leq & \exp\left(\frac{-\mu\nu}{L}\frac{\alpha^{k_0}-\alpha^{T+1}}{1-\alpha}\right)\left(\Delta_1\exp\left(-\frac{\mu\nu}{L}\frac{\alpha-\alpha^{k_0}}{1-\alpha}\right) + c_5\exp\left(\frac{\mu\nu\alpha^{k_0}}{L(1-\alpha)}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{k_0\ln(T/\beta)^2}{T^2}\right)\\ &+ \frac{2c^2\sigma^2}{L}\exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}\\ &= \Delta_1\exp\left(\frac{-\mu\nu}{L}\frac{\alpha-\alpha^{T+1}}{1-\alpha}\right) + c_5\exp\left(\frac{\mu\nu}{L}\frac{\alpha^{T+1}}{1-\alpha}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{k_0\ln(T/\beta)^2}{T^2}\\ &+ \frac{2\nu^2\sigma^2}{L}\exp\left(\frac{\nu\alpha^{T+1}}{\kappa(1-\alpha)}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{(T-k_0+1)\ln(T/\beta)^2}{T^2}\end{aligned}$$

Using Lemma 8 to bound the first term

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

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

For bounding the second and third terms, note that $\frac{\alpha^{T+1}}{1-\alpha} \leq \frac{2\beta}{\ln(T/\beta)}$, and $k_0 = T\frac{\ln(\nu)}{\ln(T/\beta)}$. Using these relations and the fact that $\alpha \leq 1$,

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

Putting in the value of c_5 and rearranging, we get

$$\begin{aligned}\Delta_{T+1} \leq & \Delta_1 c_2\exp\left(-\frac{\nu T}{\kappa}\frac{\alpha}{\ln(T/\beta)}\right) + \frac{2\nu^2\sigma^2}{LT}\exp\left(\frac{\nu}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)\frac{4\kappa^2\ln(T/\beta)^2}{\nu^2e^2\alpha^2}\\ &+ \left[\max_{j\in[k_0]}\{f(w_j)-f^*\}\frac{2\nu(\nu-1)}{L}\right]\exp\left(\frac{\nu}{\kappa}\frac{2\beta}{\ln(T/\beta)}\right)\frac{4\kappa^2}{\nu^2e^2\alpha^2}\frac{\ln(\nu)\ln(T/\beta)}{T}\end{aligned}$$

□

D.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,*

$$\mathbb{E}[f(\bar{w}_{T+1}) - f(w^*)] \leq \frac{2L\left\|w_1 - w^*\right\|^2}{\sum_{k=1}^T \alpha_k} + \sigma^2 \frac{\sum_{k=1}^T \alpha_k^2}{\sum_{k=1}^T \alpha_k} \tag{11}$$

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,

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

Proof. Following the proof of Theorem 3,

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

Taking expectation,

$$\begin{aligned}
\mathbb{E}\left\|w_{k+1} - w^*\right\|^2 &\leq \left\|w_k - w^*\right\|^2 - \frac{\alpha_k}{L}[f(w_k) - f(w^*)] + \frac{\alpha_k^2}{2L}\left[f(w_k) - f(w^*)\right] + \frac{\alpha_k^2}{2L}\sigma^2 \\
&\leq \left\|w_k - w^*\right\|^2 - \frac{\alpha_k}{2L}[f(w_k) - f(w^*)] + \frac{\alpha_k^2}{2L}\sigma^2 \qquad \text{(Since } f(w_k) - f(w^*) \geq 0 \text{ and } \alpha_k \leq 1)
\end{aligned}$$

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

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

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

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

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 8, 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 $\alpha = \left[\frac{\beta}{T}\right]^{1/T}$,

$$\begin{aligned}\frac{\sum_{k=1}^T \alpha^{2k}}{\sum_{k=1}^T \alpha^k} &\leq \frac{\alpha^2}{1-\alpha^2}\frac{1-\alpha}{\alpha-\alpha^{T+1}} \\ &= \frac{\alpha}{1+\alpha}\frac{1}{1-\alpha^T} \leq \frac{1}{1-\alpha^T} = \frac{T}{T-\beta}\end{aligned}$$

Putting everything together,

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

□

D.5 Additional lemmas for upper-bound proofs

Lemma 10. *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 9, 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,

$$\begin{aligned}\sum_{k=1}^T \alpha^k \exp\left(-\frac{1}{\kappa}\sum_{i=t+1}^T \alpha^i\right) &\leq c_2\sum_{k=1}^T \alpha^k\frac{(1-\alpha)\kappa}{e\alpha^{k+1}} && \text{(Lemma 15)} \\ &\leq c_2(1-\alpha)\frac{\kappa T}{e\alpha} \\ &\leq c_2\ln(1/\alpha)\frac{\kappa T}{e\alpha} \\ &= c_2\frac{\kappa\ln(T/\beta)}{e\alpha}\end{aligned}$$

□

Lemma 11. *If f_i is L_i-smooth, stochastic lines-searches ensures that*

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

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,

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

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

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

The line-search condition is looser than smoothness if

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

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

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

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

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

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

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

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

E Lower-bound proofs for Section 4

E.1 Proof of Theorem 5

Theorem 5. *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 $f_2(w) = \frac{1}{2}\left(2w + 1/2\right)^2$, setting the step-size 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,*

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

Proof. For SLS with a general $c \geq 1/2$ on quadratics, we know that $\gamma_k = \frac{2(1-c)}{L_{i_k}}$ (see Lemma 11 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,

$$\begin{aligned}
f_1(w) &= \frac{1}{2}(w-1)^2 \Rightarrow L_1 = 1 \\
f_2(w) &= \frac{1}{2}(2w + \frac{1}{2})^2 \Rightarrow L_2 = 4 \\
f(w) &= \frac{1}{4}(w-1)^2 + \frac{1}{4}(2w + \frac{1}{2})^2 = \frac{5}{4}w^2 + \frac{1}{4} + \frac{1}{16} \Rightarrow w^* = 0
\end{aligned}$$

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

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

□

E.2 Proof of Theorem 7

Theorem 7. *When using gradient descent to minimize a one-dimensional quadratic function* $f(w) = \frac{1}{2}(xw - y)^2$, *with* $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$ *and* $\gamma_k = \frac{\nu}{L}$ *for* $\nu > 3$ *we have*

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

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

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

□

E.3 Lemmas for convex setting

Lemma 12. *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. □

Lemma 13. *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. □

F Helper Lemmas

Lemma 14. *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. □

Lemma 15. *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. □

Lemma 16. *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:

$$\begin{aligned}
\prod_{k=1}^{T+1}(1-\alpha_k)+\sum_{k=1}^{T+1}\alpha_k\prod_{i=k+1}^{T+1}(1-\alpha_i) &= (1-\alpha_{T+1})\prod_{k=1}^{T}(1-\alpha_k)+\left(\alpha_{T+1}+\sum_{k=1}^{T}\alpha_k\prod_{i=k+1}^{T+1}(1-\alpha_i)\right)\\
&= (1-\alpha_{T+1})\prod_{k=1}^{T}(1-\alpha_k)+\left(\alpha_{T+1}+(1-\alpha_{T+1})\sum_{k=1}^{T}\alpha_k\prod_{i=k+1}^{T}(1-\alpha_i)\right)\\
&= (1-\alpha_{T+1})\left(\underbrace{\prod_{k=1}^{T}(1-\alpha_k)+\sum_{k=1}^{T}\alpha_k\prod_{i=k+1}^{T}(1-\alpha_i)}_{=1}\right)+\alpha_{T+1}\\
&\qquad\qquad\text{(Induction hypothesis)}\\
&= (1-\alpha_{T+1})+\alpha_{T+1}=1
\end{aligned}$$

□

G Proof for misspecified ASGD

New proofs not included in current Arxiv version

Misspecified ASGD uses the same two sequences $\{w_k, y_k\}$ and an additional extrapolation parameter b_k. However, we do not have information about the smoothness L or the strong-convexity constant μ. Instead we estimate these problem-dependent parameters in an offline fashion (no correlation between i_k and the estimation), and obtain $\tilde{L}$ and $\tilde{\mu}$. W.l.o.g, we will assume that $\frac{1}{\tilde{L}} = \frac{\nu_L}{L}$ and $\tilde{\mu} = \nu_\mu \mu$ for some $\nu_L, \nu_\mu > 0$. Specifically, the extrapolation parameter is now computed as follows:

$$r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k \tilde{\mu} \eta_k. \tag{26}$$

$$b_k = \frac{(1 - r_{k-1}) r_{k-1} \frac{\eta_k}{\eta_{k-1}}}{r_k + r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}}}, \tag{27}$$

where $\eta_k = \gamma_k \alpha_k = \frac{\nu_L}{\rho L} \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\nu_L \nu_\mu \mu}{\rho L}} \left(\frac{\beta}{T}\right)^{k/2T}$ satisfy the above equations.

The above equations can be rewritten as:

$$r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\eta_k}{\eta_{k-1}} + r_k \mu \, \nu_\mu \eta_k. \tag{28}$$

Defining $\tilde{\eta_k} := \nu_\mu \eta_k$ and noting that the ratio $\frac{\tilde{\eta_k}}{\tilde{\eta}_{k-1}} = \frac{\eta_k}{\eta_{k-1}}$,

$$r_k^2 = (1 - r_k) r_{k-1}^2 \frac{\tilde{\eta_k}}{\tilde{\eta}_{k-1}} + r_k \mu \, \tilde{\eta_k}. \tag{29}$$

$$b_k = \frac{(1 - r_{k-1}) r_{k-1} \frac{\tilde{\eta_k}}{\tilde{\eta}_{k-1}}}{r_k + r_{k-1}^2 \frac{\tilde{\eta_k}}{\tilde{\eta}_{k-1}}}, \tag{30}$$

where $\tilde{\eta_k} := \frac{\nu_L \nu_\mu}{\rho L} \left(\frac{\beta}{T}\right)^{k/T}$, $r_k = \sqrt{\frac{\nu_L \nu_\mu \mu}{\rho L}} \left(\frac{\beta}{T}\right)^{k/2T}$ satisfy the above equations.

For notational convenience we will redefine $\gamma_k := \nu \gamma_k$ where $\nu = \nu_L \nu_\mu$ and hence redefine $\eta_k := \nu \eta_k$. With these redefinitions, it is easy to see that the updates can be reformulated as a 3 variable sequence as in Appendix C.1 with a different choice of η_k. Similarly, we can use the same definition of the estimating sequences as in Appendix C.2.

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

Lemma 17. *Using the update in Eq. (15) with $\eta_k = \frac{\nu}{\rho L} \alpha_k$, and defining $k_0 := T \frac{[\ln(\nu)]_+}{\ln(T/\beta)}$ and $G := \max_{j \in [k_0]} \mathbb{E}[\|\nabla f(y_k)\|^2]$, we obtain the following descent lemma.*

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] - \frac{\eta_k}{2} \|\nabla f(y_k)\|^2 + \frac{\nu^2}{2\rho^2 L} \alpha_k^2 \sigma^2 \qquad \text{(For } k \geq k_0\text{)}$$

$$\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(y_k)] + \frac{G^2 \nu^2}{2\rho L} \alpha_k^2 + \frac{\nu^2}{2\rho^2 L} \alpha_k^2 \sigma^2 \qquad \text{(For } k < k_0\text{)}$$

Proof. By smoothness, and the update in Eq. (15),

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

Taking expectation w.r.t. i_k,

$$\begin{aligned}
\mathbb{E}[f(w_{k+1})] &\leq \mathbb{E}[f(y_k)] - \eta_k \left\|\nabla f(y_k)\right\|^2 + \frac{L}{2}\eta_k^2 \mathbb{E}[\left\|\nabla f_{ik}(y_k)\right\|^2] \qquad (\eta_k \text{ is independent of the randomness in } i_k.) \\
&\leq \mathbb{E}[f(y_k)] - \eta_k \left\|\nabla f(y_k)\right\|^2 + \frac{\rho L}{2}\eta_k^2 \mathbb{E}[\left\|\nabla f(y_k)\right\|^2] + \frac{L}{2}\eta_k^2\sigma^2 \qquad (\text{By the growth condition in Eq. (5)}) \\
&= \mathbb{E}[f(y_k)] - \eta_k \left\|\nabla f(y_k)\right\|^2 + \frac{\eta_k \nu \alpha_k}{2} \mathbb{E}[\left\|\nabla f(y_k)\right\|^2] + \frac{\nu^2}{2\rho^2 L}\alpha_k^2\sigma^2 \\
&= \mathbb{E}[f(y_k)] - \frac{\eta_k}{2} \left\|\nabla f(y_k)\right\|^2 - \frac{\eta_k}{2}\left(1 - \nu\alpha_k\right) \left\|\nabla f(y_k)\right\|^2 + \frac{\nu^2}{2\rho^2 L}\alpha_k^2\sigma^2
\end{aligned}$$

For $k \geq k_0$, $1 - \nu\alpha_k \geq 0$, and

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

whereas for $k < k_0$, $1 - \nu\alpha_k > -\nu\alpha_k$, and since $\eta_k = \frac{\nu\alpha_k}{\rho L}$,

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

Defining $G := \max_{j \in [k_0]} \mathbb{E}[\left\|\nabla f(y_k)\right\|^2]$,

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

□

With this change, the proof of Lemma 18 can be modified, and the new result can be stated as follows:

Lemma 18. *For the estimating sequences defined in Appendix C.2 and the updates in Eq. (14)-Eq. (19),*

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

where $\mathcal{N}_k := \frac{\sigma^2\nu^2}{2\rho^2 L} \sum_{j=0}^{k-1} \alpha_j^2 \prod_{i=j+1}^{k-1}(1 - r_i) + \left(\frac{G^2\nu^2}{2\rho L}\right) \sum_{j=0}^{\min\{k_0,k\}-1} \alpha_j^2 \prod_{i=j+1}^{k-1}(1 - r_i)$

We now use the above lemma to prove the rate for strongly-convex functions.

Theorem 2. *Assuming (i) convexity and L_i-smoothness of each f_i, (ii) μ strong-convexity of f and (iii) the growth condition in Eq. (5), ASGD (Eqs. (3) and (4)) with $w_0 = y_0$, $\gamma_k = \frac{\nu_L}{\rho L}$, $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$, $\alpha_k = \alpha^k$, $r_k = \sqrt{\frac{\nu_L \nu_\mu \mu}{\rho L}} \left(\frac{\beta}{T}\right)^{k/2T}$ and b_k computed as:*

$$b_k = \frac{(1 - r_{k-1})\, r_{k-1}\, \alpha}{r_k + r_{k-1}^2\, \alpha}, \tag{7}$$

has the following convergence rate:

$$\begin{aligned}
&\mathbb{E}[f(w_T) - f^*] \leq \\
&2c_3 \exp\left(-\frac{\sqrt{\nu}T}{\sqrt{\kappa\rho}} \frac{\alpha}{\ln(T/\beta)}\right) [f(w_0) - f^*] \\
&+ \frac{8c_4\kappa(\ln(T/\beta))^2}{(e\alpha)^2 \rho L T} \nu\sigma^2 \\
&+ \frac{2c_4\kappa(\ln(T/\beta))^2}{(e\alpha)^2 L T} \min\{\frac{\lfloor \ln(\nu) \rfloor_+}{\ln(T/\beta)}, 1\} \nu G^2
\end{aligned}$$

where $\nu = \nu_L \nu_\mu$, $\kappa = \frac{L}{\mu}$, $c_3 = \exp\left(\frac{1}{\sqrt{\rho\kappa}} \frac{2\beta\sqrt{\nu}}{\ln(T/\beta)}\right)$ and $c_4 = \exp\left(\frac{1}{\alpha\sqrt{\rho\kappa}} \frac{2\beta\sqrt{\nu}}{\ln(T/\beta)}\right)$ and $G := \max_{j \in [k_0]} \mathbb{E}[\|\nabla f(y_k)\|^2]$ with $k_0 := T \frac{\lfloor \ln(\nu) \rfloor_+}{\ln(T/\beta)}$ and $\lfloor x \rfloor_+ = \max\{\lfloor x \rfloor, 0\}$.

Proof. Using the reformulation in Lemma 5 gives us $q_k = \mu$ for all k and $z_0 = w_0$. For the estimating sequences defined in Appendix C.2, using Lemma 18, we know that the (reformulated) updates satisfy the following relation

$$\mathbb{E}[f(w_T)] \leq \mathbb{E}[\phi_T^*] + \mathcal{N}_T \leq \mathbb{E}[\phi_T(w^*)] + \mathcal{N}_T$$

From Eq. (22), 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,

$$\begin{aligned}
&\mathbb{E}[f(w_T)] \leq (1 - \lambda_T) f^* + \lambda_T \phi_0(w^*) + \mathcal{N}_T \\
\implies &\mathbb{E}[f(w_T) - f^*] \leq \lambda_T \left[\phi_0(w^*) - f^*\right] + \mathcal{N}_T
\end{aligned}$$

By Eq. (21),

$$\leq \lambda_T \left[\phi_0^* + \frac{q_0}{2} \|w^* - z_0\|^2 - f^*\right] + \mathcal{N}_T$$

Choosing $\phi_0^* = f(w_0)$,

$$\leq \lambda_T \left[f(w_0) - f^* + \frac{q_0}{2} \|w^* - z_0\|^2\right] + \mathcal{N}_T$$

Since $z_0 = w_0$, $q_0 = \mu$,

$$\implies \mathbb{E}[f(w_T) - f^*] \leq \lambda_T \left[f(w_0) - f^* + \frac{\mu}{2} \|w^* - w_0\|^2\right] + \frac{\sigma^2\nu^2}{2\rho^2 L} \sum_{j=0}^{T-1} \alpha_j^2 \prod_{i=j+1}^{T-1} (1 - r_i) + \left(\frac{G^2\nu^2}{2\rho L}\right) \sum_{j=0}^{\min\{k_0, T\}-1} \alpha_j^2 \prod_{i=j+1}^{T-1} (1 - r_i)$$

Using the fact that $\lambda_0 = 1$ and $\lambda_{k+1} = (1 - r_k)\lambda_k$, we know that that $\lambda_T = \prod_{k=1}^{T} (1 - r_k)$, and

$$\begin{aligned}
\mathbb{E}[f(w_T) - f^*] \leq &\left[\prod_{k=1}^{T} (1 - r_k)\right] \left[f(w_0) - f^* + \frac{\mu}{2} \|w^* - w_0\|^2\right] + \frac{2\sigma^2\nu^2}{\rho^2 L} \sum_{j=0}^{T-1} \alpha_j^2 \prod_{i=j+1}^{T-1} (1 - r_i) \\
&+ \left(\frac{G^2\nu^2}{2\rho L}\right) \sum_{j=0}^{\min\{k_0, T\}-1} \alpha_j^2 \prod_{i=j+1}^{T-1} (1 - r_i).
\end{aligned}$$

Now our task is to upper-bound bound the $1 - r_k$ terms. From Eq. (18), we know that

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

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

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

$$\mathbb{E}[f(w_T) - f^*] \leq \left[\prod_{k=1}^{T}\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_k\right)\right]\left[f(w_0) - f^* + \frac{\mu}{2}\left\|w^* - w_0\right\|^2\right] + \frac{2\sigma^2\nu^2}{\rho^2 L}\sum_{j=0}^{T-1}\alpha_j^2\prod_{i=j+1}^{T-1}\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_i\right)$$
$$+ \left(\frac{G^2\nu^2}{2\rho L}\right)\sum_{j=0}^{\min\{k_0,T\}-1}\alpha_j^2\prod_{i=j+1}^{T-1}\left(1 - \sqrt{\frac{\nu}{\rho\kappa}}\alpha_i\right).$$

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

$$\Delta_T \leq 2\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\sum_{k=1}^{T}\alpha^k\right)\Delta_0 + \frac{2\sigma^2\nu^2}{\rho^2 L}\sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\sum_{i=k+1}^{T-1}\alpha^i\right)$$
$$+ \frac{G^2\nu^2}{2\rho L}\sum_{k=0}^{\min\{k_0,T\}-1}\alpha^{2k}\exp\left(-\sqrt{\frac{1}{\rho\kappa}}\sum_{i=k+1}^{T-1}\alpha^i\right)$$

Using Lemma 8, we can bound the first term as

$$2\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\sum_{k=1}^{T}\alpha^k\right)\Delta_0 \leq 2\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\left(\frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}\right)\right)\Delta_0$$
$$= 2c_3\exp\left(-\frac{T\sqrt{\nu}}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_0) - f^*]$$

where $c_3 = \exp\left(\frac{2\beta\sqrt{\nu}}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$. We can now bound the second term by a proof similar to Lemma 9. Indeed we have

$$\sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\sum_{i=k+1}^{T-1}\alpha^i\right) = \sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\frac{\alpha^{k+1} - \alpha^T}{1-\alpha}\right)$$
$$= \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\sum_{k=0}^{T-1}\alpha^{2k}\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\frac{\alpha^{k+1}}{1-\alpha}\right)$$
$$\leq \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\sum_{k=0}^{T-1}\alpha^{2k}\left(\frac{2(1-\alpha)\sqrt{\rho\kappa}}{e\alpha^{k+1}\sqrt{\nu}}\right)^2 \qquad \text{(Lemma 15)}$$
$$= \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa}{e^2\nu\alpha^2}T(1-\alpha)^2$$
$$\leq \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa}{e^2\nu\alpha^2}T\ln(1/\alpha)^2$$
$$= \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2}{e^2\nu\alpha^2 T}$$

Similarly,

$$\sum_{k=0}^{\min\{k_0,T\}-1}\alpha^{2k}\exp\left(-\sqrt{\frac{\nu}{\rho\kappa}}\sum_{i=k+1}^{T-1}\alpha^i\right) \leq \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2\min\{k_0,T\}}{e^2\nu\alpha^2T^2}$$
$$= \exp\left(\frac{\sqrt{\nu}}{\sqrt{\rho\kappa}}\frac{\alpha^T}{1-\alpha}\right)\frac{4\rho\kappa\ln(T/\beta)^2\min\left\{\frac{\ln(\nu)}{\ln(T/\beta)}, 1\right\}}{e^2\nu\alpha^2T}$$

Finally,

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

where the inequality comes from the bound in Eq. (25) in the proof of Lemma 8. Putting everything together we obtain

$$\mathbb{E}[f(w_T) - f^*] \leq 2c_3 \exp\left(-\frac{\sqrt{\nu}T}{\sqrt{\kappa\rho}}\frac{\alpha}{\ln(T/\beta)}\right)[f(w_0) - f^*] + \frac{8\rho c_4\kappa\ln(T/\beta)^2}{e^2\alpha^2 T}\frac{\sigma^2\nu}{\rho^2 L} + \frac{2\rho c_4\kappa\ln(T/\beta)^2}{e^2\alpha^2 T}\min\left\{\frac{[\ln(\nu)]_+}{\ln(T/\beta)}, 1\right\}\frac{G^2\nu}{\rho L}$$

where $c_3 = \exp\left(\frac{2\beta\sqrt{\nu}}{\sqrt{\rho\kappa}\ln(T/\beta)}\right)$ and $c_4 = \exp\left(\frac{2\beta\sqrt{\nu}}{\alpha\sqrt{\rho\kappa}\ln(T/\beta)}\right)$. □