university of groningen

faculty of science and engineering

New Tight Bounds for SGD without Variance Assumption: A Computer-Aided Lyapunov Analysis

Master Project Applied Mathematics

June 2025

Student: D. Cortild

First supervisor: Prof. Dr. J. G. Peypouquet

Second supervisor: Prof. Dr. K. Camlibel

Abstract

The analysis of Stochastic Gradient Descent (SGD) often relies on making some assumption on the variance of the stochastic gradients, which is usually not satisfied or difficult to verify in practice. This work contributes to a recent line of works which attempt to provide guarantees without making any variance assumption, leveraging only the (strong) convexity and smoothness of the loss functions. In this context, we prove new theoretical bounds derived from the monotonicity of a simple Lyapunov energy, improving the current state-of-the-art and extending their validity to larger step-sizes. Our theoretical analysis is backed by a Performance Estimation Problem analysis, which allows us to claim that, empirically, the bias term in our bounds is tight within our framework.

Keywords. Stochastic Gradient Descent, Convex Optimization, Performance Estimation Problem, Lyapunov Analysis.

Contents

1 Introduction

Given $n \geq 1$ convex and smooth real-valued functions $f_1, \ldots, f_n \colon \mathbb{R}^d \to \mathbb{R}$, we consider the problem given by

$$\min_{x \in \mathbb{R}^d} f(x), \quad \text{where } f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{Finite-Sum}$$

We denote $S = \operatorname{argmin}(f)$, and assume it to be nonempty.

The Stochastic Gradient Descent (SGD) algorithm (Robbins and Monro, 1951) is one of the most popular approaches for solving Problem (Finite-Sum), and is widely used for large-scale machine learning and stochastic optimization problems. In its standard version, at each iteration, SGD performs a gradient step using one of the functions in the sum, which is chosen at random. More precisely, given the iterate x_t, SGD picks $i_t \in \{1, \ldots, n\}$ uniformly at random, and computes x_{t+1} via

$$x_{t+1} = x_t - \gamma \nabla f_{i_t}(x_t), \tag{SGD}$$

where $\gamma > 0$ is the step-size. Intuitively, SGD is able to progress towards the solutions of the problem because it follows a direction that, on average, is an unbiased estimator for ∇f. The simplicity and efficiency of SGD make it the go-to algorithm for training deep neural networks and other models on massive datasets, where computing the entire gradient ∇f is intractable or too expensive.

Variants of SGD. Variations of the standard SGD algorithm include *primal averaging* (Polyak and Juditsky, 1992) and *dual averaging* (Xiao, 2009), momentum terms (Sutskever et al., 2013), adaptive coefficients as in AdaGrad (Duchi et al., 2011), or both momentum and adaptive terms as in Adam (Kingma and Ba, 2015). Variance reduction techniques have also been considered, such as SAG (Schmidt et al., 2017) or SAGA (Defazio et al., 2014). An extension to equilibrium problems via maximally monotone operators was studied in Combettes and Pesquet (2015). These adaptations are not the focus of this work and will not be considered further here.

Complexity of SGD. Quantitative results on the behavior of SGD consist in upper bounds for a metric $\Delta(T)$, which is to be interpreted as an optimality gap at iteration T. Typical choices include the difference between the expected value of f at a point obtained after T iterations and its minimum value, or the expected square-distance from such a point to a solution. In most known results, these upper bounds are expressed as the sum of two terms, taking the form

$$\Delta(T) \leq \text{Bias}(T) + \text{Variance}(T),$$

where the *bias term* typically vanishes when $T \to +\infty$, and where the *variance term* can ideally be made arbitrarily small by taking a small enough step-size. Ideally, we would like the bias term to be equal to the rate of convergence of Gradient Descent (GD).

Assumptions for SGD. Classical studies on SGD typically assume that each function f_i is L-smooth and either convex, strongly convex or satisfies a Polyak-Lojasiewicz inequality. Some works assume relaxations of these hypotheses, making assumptions under the form of variational inequalities, such as the expected smoothness (Gower et al., 2019), the (L_0, L_1)-smoothness (Zhang et al., 2020), the expected residual (Gower et al., 2021b), or the ABC condition (Khaled and Richtárik, 2023). Other works assume that only the function f is smooth or convex. For the sake

of readability, we will assume that each of the functions f_i is convex and smooth, even though we show in the Section 6 that we simply need an assumption analog to expected smoothness.

Besides the assumptions on the problem itself, it is quite standard to make an assumption on the algorithm, through assumptions on the variance of the stochastic gradients $\nabla f_{i_t}(x_t)$. Many different variance assumptions have been studied in the context of SGD. The oldest and possibly the best-known of such assumptions is the uniform boundedness of the variance (Robbins and Monro, 1951; Nemirovski et al., 2009; Rakhlin et al., 2012; Schmidt and Roux, 2013), or a variant, namely

$$\sup_{x\in\mathbb{R}^d} \mathbb{E}\big[\|\nabla f_i(x) - \nabla f(x)\|^2\big] < +\infty \quad \text{or} \quad \sup_{x\in\mathbb{R}^d} \mathbb{E}\big[\|\nabla f_i(x)\|^2\big] < +\infty.$$

The latter strictly implies the former, and is usually made in non-smooth stochastic optimization (Bubeck, 2015, Section 6) or online optimization (Hazan, 2016), whereas the former is more common in our convex and smooth setting. These assumptions are however unrealistic in practice (Bottou et al., 2018; Nguyen et al., 2018), and are never verified for strongly convex functions. The bounded variance assumption has been relaxed into weaker assumptions, such as the Blum-Gladyshev assumption (Blum, 1954; Gladyshev, 1965; Alacaoglu et al., 2025), maximal strong growth (Schmidt and Roux, 2013), strong and weak growth (Vaswani et al., 2019), or relaxed growth (Bertsekas and Tsitsiklis, 2000; Bottou et al., 2018), to name a few. But the problem remains that those assumptions cannot be verified a priori in practice for smooth problems.

In this work, we make no boundedness assumptions on the variance, a setting already pioneered in (Bach and Moulines, 2011; Needell et al., 2016; Gower et al., 2019).

Our contribution. We establish new and improved quantitative results on the complexity of SGD in the convex and L-smooth setting, without making any variance assumption. Instead, our bounds will rely on the so-called *solution variance* constant, which is always finite and defined as

$$\sigma_*^2 = \frac{1}{m}\sum_{i=1}^{n} \|\nabla f_i(x_*)\|^2,$$

where x_* is any minimizer of f. Indeed, the constant does not depend on the choice of x_*, see (Garrigos and Gower, 2024, Lemma 4.17). Our results take the form

$$\Delta(T) \leq \text{Bias}(T) \cdot \|x_0 - x_*\|^2 + \text{Variance}(T) \cdot \sigma_*^2,$$

and are valid for the whole range of step-sizes $\gamma L \in (0, 2)$. We focus on obtaining the smallest *bias term* possible, and on top of improving it with respect to the existing results, we conduct numerical experiments illustrating the sharpness of our bound. To the best of our knowledge, this is the first work where a unified and comprehensive study of SGD, without bounded variance assumptions and in the whole range of step-sizes, has been carried out in the convex and smooth setting.

In the convex case, we establish bounds for the expected function value gap $\Delta(T) = \mathbb{E}[f(\bar{x}_T) - \min f]$, where $\bar{x}_T$ is the Cesàro average of the iterates generated by SGD up to iteration T. Our bounds,

with simplified notation for ease of presentation, are

$$\text{Bias}(T) = \begin{cases} \frac{1}{2\gamma T} & \text{if } \gamma L \in (0,1), \\ \frac{1}{(2-\varepsilon)\gamma T} & \text{if } \gamma L = 1, \varepsilon > 0, \\ \frac{1}{2\gamma(2-\gamma L)T} & \text{if } \gamma L \in (1,2), \end{cases} \quad \text{Variance}(T) = \begin{cases} \frac{\gamma}{2(1-\gamma L)} & \text{if } \gamma L \in (0,1), \\ \frac{\gamma(2+\varepsilon)}{\varepsilon(2-\varepsilon)} & \text{if } \gamma L = 1, \varepsilon > 0, \\ \frac{\exp(T)}{2-\gamma L} & \text{if } \gamma L \in (1,2). \end{cases}$$

In our setting, where no assumptions on the variance are made, the only existing bounds are available for step-sizes limited to $\gamma L \in (0, \frac{1}{2})$, and provide worse constants (Gower et al., 2021a). Our results are the best so far and the first ones valid in the whole interval $(0, 1)$. Moreover, it matches the bias term previously obtained under the bounded variance assumption (Taylor and Bach, 2019). For $\gamma L = 1$ we are able to have a bias term as close as possible from the optimal bias $\frac{1}{2\gamma L}$, but cannot have equality with a finite variance term. Numerical results suggest that it is not possible to achieve this optimal bias. For $\gamma L \in (1, 2)$, which is a completely uncharted territory for this problem, we find that the variance term grows exponentially with T, making it impractical. Again, our numerical experiments suggest that one cannot avoid this growth in the variance term. Note that those problems encountered for large step-sizes $\gamma L \in [1, 2)$ are typically avoided when making a uniformly bounded variance assumption, see e.g. Taylor and Bach (2019).

As a by-product, our new bounds for SGD with the step-size $\gamma L = 1$ makes it possible to study the stochastic proximal algorithm through the lens of SGD. To the best of our knowledge, we establish the first complexity guarantees for this method in a general convex nonsmooth setting.

In the case where each function f_i is μ-strongly convex, we provide bounds over the expected distance to the solution, namely $\Delta(T) = \mathbb{E}\left[\|x_T - x_*\|^2\right]$. Our results show that we can take $\text{Bias}(T) = \phi^{2T}$ where ϕ corresponds to the standard rate of GD, namely

$$\phi = \max\{1 - \gamma\mu; \gamma L - 1\},$$

with the exception of the optimal step-size $\gamma = \frac{2}{\mu + L}$, for which the bias term must be relaxed to avoid a blow up in the variance. These results improve on (Bach and Moulines, 2011; Needell et al., 2016; Gower et al., 2019) by allowing for the full range of step-sizes $\gamma L \in (0, 2)$, and by having a sharp bias term. Our numerical experiments suggest again that we cannot improve those results.

The strategy. For each fixed time horizon T, we propose a time-dependent random energy sequence

$$E_t = a_t \cdot \|x_t - x_*\|^2 + \rho \cdot \sum_{s=0}^{t-1} \big(f(x_s) - \min f\big) - \sum_{s=0}^{t-1} e_s \sigma_*^2 \quad \forall t = 0, \dots, T, \qquad \text{(Lyapunov)}$$

where ρ, (a_t) and (e_t) are nonnegative. The first ingredient of our Lyapunov energy is the distance to the solution $\|x_t - x_*\|^2$, a classical term which typically decreases for deterministic gradient dynamics. The second term involves the function gap $f(x_t) - \min f$, which also typically decreases for gradient descent. The standard Lyapunov for gradient descent usually contains a term $t(f(x_t) - \min f)$, which we have replaced here with the sum of the past function gaps, which is of the same order in time. This choice is critical to obtain bounds without variance assumptions (compare with the Lyapunov energies considered by Taylor and Bach (2019)). The last term is a *negative*

cumulated sum, where the e_t's play the role of a variance term. It is here to compensate the fluctuations caused by the variance in the SGD algorithm, and will allow the Lyapunov energy to decrease.

A decrease of $\mathbb{E}[E_t]$ along the iterations means that $\mathbb{E}[E_T] \leq E_0$, which translates into a bound for SGD. Our strategy is therefore to find *admissible* Lyapunov parameters ρ, (a_t) and (e_t), which are those that make the energy decrease in expectation. Among those admissible parameters, our goal is to find the ones minimizing, whenever possible, the bias term in our bound.

To this end, we rely on the approach followed in Taylor and Bach (2019) to design Lyapunov potentials in a systematic fashion. It is based on the *Performance Estimation Problem* (PEP) methodology, which originally appeared in Drori and Teboulle (2014), and was further developed in Taylor et al. (2017b). A similar line of thought was followed in (Fercoq, 2024). Automated Lyapunov analyses have also been developed in (Taylor et al., 2018a; Upadhyaya et al., 2025) by combining the PEP approach with the *Integral Quadratic Constraints* framework (Lessard et al., 2016). The PEP framework allowed us to numerically drive our theoretical proofs, and to assess the sharpness of our bounds.

This thesis is organized as follows: Section 2 presents our main results in the convex setting, while Section 3 focuses on the strongly convex setting. Section 4 describes the PEP framework in more details. We discuss the tightness of our results in Section 5, supported by numerical experiments. Section 6 discusses a possible relaxation of the assumptions under which our results remain valid. In Sections 7 and 8 we include extensions of our results to the case of non-uniform sampling, mini-batching and to the study of the stochastic proximal algorithm. Finally, in Section 9, we conclude the work. All technical results and extensions of our results are gathered in the appendices.

2 Convex Setting

In this section, we consider the smooth convex setting, which we formalize below. Note that the assumption could be relaxed, see the discussion in Section 6.

Assumption 2.1 (Convexity and smoothness)**.** *Considering Problem* (Finite-Sum)*, we assume that each function f_i is convex and L-smooth for some $L \in (0, +\infty)$.*

As mentioned in the introduction, our analysis is based on Lyapunov energies of the form (Lyapunov). The next lemma illustrates how a Lyapunov decrease leads to upper bounds for (SGD).

Lemma 2.2 (Bound from Lyapunov decrease)**.** *Let Assumption 2.1 hold. Assume that $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for every $t = 0, \dots, T-1$, that $\rho > 0$, and, without loss of generality, that $a_0 = 1$. Then*

$$\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{\|x_0 - x_*\|^2}{\rho T} + \frac{\bar{e}\sigma_*^2}{\rho},$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$ and $\bar{e} = \frac{1}{T}\sum_{t=0}^{T-1} e_t$.

Proof. Iterating $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for $t = 0, \dots, T-1$ yields that $\mathbb{E}[E_T] \leq E_0$, or, equivalently,

$$a_T \mathbb{E}\left[\|x_T - x_0\|^2\right] + \rho \sum_{t=0}^{T-1} \mathbb{E}[f(x_t) - \min f] - \sum_{t=0}^{T-1} e_t \sigma_*^2 \leq a_0 \|x_0 - x_*\|^2.$$

Bounding $a_T \mathbb{E}\left[\|x_T - x_0\|^2\right] \geq 0$ and dividing by $\rho T > 0$ yields

$$\min_{t=0,\dots,T-1} \mathbb{E}\left[f(x_t) - \min f\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} [f(x_t) - \min f] \leq \frac{a_0 \|x_0 - x_*\|^2}{\rho T} + \frac{\bar{e}\sigma_*^2}{\rho}.$$

If f is convex, we may use Jensen's inequality to write $\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[f(x_t) - \min f\right]$, which implies the wanted conclusion, after setting $a_0 = 1$. □

The next results are consequences of Lemma 2.2, after finding specific values of ρ, (a_k) and (e_k), aiming at the largest possible value for ρ. Most results will be presented with a uniform (and simpler) upper bound on e_k for ease of presentation, and the more detailed constants can be found in Appendix B, whereto all proofs are postponed. The presentation of our results is split into three parts based on the normalized step-size γL, namely Theorem 2.3 for short step-sizes $\gamma L \in (0, 1)$, Theorem 2.4 for the optimal step-size $\gamma L = 1$ and Theorem 2.7 for large step-sizes $\gamma L \in (1, 2)$.

Theorem 2.3 (Convex case, short step-sizes)**.** *Let Assumption 2.1 hold, and let (x_t) be generated by* (SGD) *with $\gamma L \in (0, 1)$. Then, for every $T \geq 1$, with $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$,*

$$\mathbb{E}[f(\bar{x}_T) - \min f] \leq \frac{L\|x_0 - x_*\|^2}{2\gamma L T + 2(1 - \gamma L)} + \frac{\gamma \sigma_*^2}{2(1 - \gamma L)} \leq \frac{\|x_0 - x_*\|^2}{2\gamma T} + \frac{\gamma \sigma_*^2}{2(1 - \gamma L)}.$$

The setting $\gamma L \in (0, 1)$, usually considered *short step-sizes* in the GD literature, is the only one for which we have competitors in the literature, of which we detail their results in the following paragraphs.

Comparison to Gower et al. (2021a). In (Gower et al., 2021a, Theorem D.6), the authors analyzed the standard SGD algorithm under assumptions comparable to ours, while also accommodating for variable step-sizes. When adapted to our constant step-size setting, their result for $\gamma L \in (0, \frac{1}{2})$ boils down to

$$\min_{k=0,\dots,T-1} \mathbb{E}[f(x_t) - \min f] \leq \frac{\|x_0 - x_*\|^2}{2\gamma(1-2\gamma L)T} + \frac{\gamma}{1-2\gamma L}\sigma_*^2.$$

One sees that this concurrent bound diverges when γ gets close to $\frac{1}{2L}$, on both the bias and variance terms. This comparison must of course take into account the fact that the authors are able to provide bounds under weaker assumptions than our convexity and smoothness assumptions. They indeed assume f to be smooth (instead of each f_i) and relax the functions f_i to be *quasar-convex*, which is a consequence of our expected cocoercivity assumption presented in Assumption 6.4, see Remark A.4 for more details.

Comparison to Garrigos and Gower (2024). Recently, in (Garrigos and Gower, 2024, Theorem 5.5), the finite-sum minimization problem for L-smooth convex functions under assumptions identical to ours was considered. For $\gamma \in (0, \frac{1}{4L}]$, they derived the bound

$$\mathbb{E}[f(\overline{x}_T) - \min f] \leq \frac{\|x_0 - x_*\|^2}{\gamma T} + 2\gamma \cdot \sigma_*^2.$$

This result is encompassed as a consequence of Theorem 2.3.

Furthermore, their analysis can easily be extended. For any $\varepsilon > 0$ and $\gamma \in (0, \frac{1}{(1+\varepsilon)L})$, it holds that

$$\mathbb{E}[f(\overline{x}_T) - \min f] \leq \frac{\|x_0 - x_*\|^2}{2\gamma T(1-(1+\varepsilon)\gamma L)} + \frac{(1+\varepsilon^{-1})\gamma}{2(1-(1+\varepsilon)\gamma L)} \cdot \sigma_*^2.$$

This generalized bound is also recovered by Theorem 2.3 for any $\varepsilon > 0$. In the interpolation setting, allowing $\varepsilon \to 0$, Theorem 2.3 remains strictly stronger.

To the best of our knowledge, making a uniform variance assumption does not allow to obtain a bound with a better bias term than the one we obtained in Theorem 2.3. For instance (Taylor and Bach, 2019, Theorem 6) obtain the same bias term as ours for short step-sizes, and Liu and Zhou (2023) achieved a bias of $\frac{1}{\gamma L}$, although their results focus on the last iterate.

Note that our result focuses on the Cesàro averages of the iterates, which is standard for SGD with no variance assumption. This is in contrast with the literature with uniformly bounded variance, for which it is possible to derive bounds on the last iterate x_T. The first result of this kind in the convex and smooth setting was established by Bach and Moulines (2011), with bounds recently improved in (Taylor and Bach, 2019, Theorem 5) and (Liu and Zhou, 2023, Theorem 3.1). They obtain a variance term that scales with T and $\log(T)$, respectively. Both were able to derive complexity bounds, with the latter offering a significantly better one over the former. It remains an open question to know whether or not it is possible to obtain bounds on the last iterate, similar to the one established in Theorem 2.3, without uniformly bounded variance. For the time being, this showcases a discrepancy between the uniformly bounded variance setting and ours.

We now turn to the case $\gamma L = 1$, which we call the *optimal* step-size, with a slight abuse of terminology, in analogy with the optimal step-size for GD, and in view of the numerical results presented in Section 5.

Theorem 2.4 (Convex case, optimal step-size)**.** *Let Assumption 2.1 hold, and let (x_t) be generated by (SGD) with $\gamma L = 1$. Then, for every $T \geq 1$ and $\varepsilon \in (0, 2)$, with $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$,*

$$\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{\|x_0 - x_*\|^2}{(2-\varepsilon)\gamma T} + \frac{\gamma(2+\varepsilon)\sigma_*^2}{\varepsilon(2-\varepsilon)}.$$

The case $\gamma L = 1$, which is standard for GD, has surprisingly not yet seen any theoretical guarantees without bounded variance assumptions. As far as we know, this is the first upper bound provided in this setting. We point out the fact that Theorem 2.4 cannot be seen as the limit of Theorem 2.3 when $\gamma L \to 1$, because it would require to set $\varepsilon = 0$, which results in an infinite variance term. Although it would seem like a flaw of our analysis that $\varepsilon = 0$ is not feasible, we present evidence against this in Section 5.1. This singularity is surprising, and suggests that having a bias term of the order of $\frac{L}{2T}$ is impossible for SGD without any variance assumptions. This is a striking difference with GD, but also with SGD with bounded variance assumption, where no singularity holds at $\gamma L = 1$, see for instance (Taylor and Bach, 2019, Theorem 6).

Remark 2.5 (Stochastic proximal algorithm)**.** *While the convergence rates for the deterministic proximal algorithm are well-studied, complexity results for its stochastic counterpart are unknown in a general setting. But a classic trick allows to rewrite the stochastic proximal algorithm as a particular instance of SGD, with a step-size which must be $\gamma L = 1$. Thanks to Theorem 2.4, we are now able to handle this optimal step-size and to derive the first complexity bound for this algorithm with no other assumption on the functions than convexity, avoiding, for instance, to impose finiteness and Lipschitzness as in Davis and Drusvyatskiy (2019). For more details, see Section 8.*

Interestingly, when $\sigma_*^2 = 0$, the variance term vanishes and allows us to take $\varepsilon \to 0$. This setting is known as the *interpolation* regime, because it is equivalent to ask that all the functions f_i share a common minimizer, see (Garrigos and Gower, 2024, Lemma 4.17). It is a feature typically shared by overparametrized models. See Section 6.3 for more details on the definition of interpolation.

Corollary 2.6 (Convex case with interpolation, optimal step-size)**.** *Let Assumption 2.1 hold and assume that $\sigma_*^2 = 0$. Let (x_t) be generated by (SGD) with $\gamma L = 1$. Then, for every $T \geq 1$,*

$$\mathbb{E}[f(\bar{x}_T) - \min f] \leq \frac{\|x_0 - x_*\|^2}{2\gamma T},$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

The case $\gamma L \in (1, 2)$, referred to as *large step-sizes*, is yet unexplored for SGD without bounded variance assumptions. We present below the first result of this type.

Theorem 2.7 (Convex case, large step-sizes)**.** *Let Assumption 2.1 hold, and let (x_t) be generated by (SGD) with $\gamma L \in (1, 2)$. Then, for every $T \geq 1$, with $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$,*

$$\mathbb{E}[f(\bar{x}_T) - \min f] \leq \frac{\delta\|x_0 - x_*\|^2}{2\gamma(2-\gamma L)T} + \frac{\gamma\bar{e}'_T\sigma_*^2}{2(2-\gamma L)^3},$$

where $\delta = 1 - (1 - \gamma L)^{2T} \in (0, 1)$*, and* $\bar{e}'_T$ *is exponential in* T *(see its definition in Appendix B).*

In this result, we obtain a bound where the variance term increases exponentially with T. This means that, for large step-sizes, the variance term will quickly dominate the bias term, preventing the algorithm from making significant progress. Of course, one might argue that this comment is not relevant because it is based on an upper bound, and that the proof of this upper bound could be improved. However, we provide empirical evidence in Section 5 which supports the tightness of our bound, and showcase that the exponential increase in the variance term cannot be avoided.

This highlights a second discrepancy between the uniformly bounded variance setting and ours. In the large step-size regime, we are unable to provide a uniform bound of the variance term with respect to the time horizon T. This contrasts with the result in (Taylor and Bach, 2019, Theorem 6), which, under the assumption of uniformly bounded variance, establishes a variance term that remains bounded with respect to T, for step-sizes $\gamma L \in (0, \frac{1+\sqrt{5}}{2})$.

3 Strongly Convex Setting

We now focus on the strongly convex setting. As in the convex setting, the below assumption may be relaxed, which we discuss further in Section 6.

Assumption 3.1 (Strong convexity and smoothness)**.** *Considering Problem* (Finite-Sum)*, we assume that each function f_i is μ-strongly convex and L-smooth for some $\mu, L \in (0, +\infty)$.*

Our analysis is still based on the Lyapunov energy introduced in (Lyapunov). The following lemma explains how a decrease in the energy yields an upper bound for the iterates of (SGD).

Lemma 3.2 (Bound from Lyapunov decrease)**.** *Let Assumption 3.1 hold. Assume that $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for all $t = 0, \dots, T-1$, $\rho \geq 0$, and that, without loss of generality, $a_T = 1$. Then*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq a_0 \|x_0 - x_*\|^2 + e^{sum}\sigma_*^2, \qquad \text{with} \qquad e^{sum} = \sum_{t=0}^{T-1} e_t.$$

Proof. Iterating $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for $t = 0, \dots, T-1$ yields $\mathbb{E}[E_T] \leq E_0$, or in other words that

$$a_T \mathbb{E}\left[\|x_T - x_*\|^2\right] + \rho \sum_{t=0}^{T-1} \mathbb{E}[f(x_t) - \min f] \leq a_0 \|x_0 - x_*\|^2 + e_T^{sum}\sigma_*^2.$$

Since $f(x_t) - \min f \geq 0$ for all $t = 0, \dots, T-1$, and $\rho \geq 0$, this yields the wanted inequality after setting $a_T = 1$. □

We now present our main results for the strongly convex case, which are obtained after carefully choosing Lyapunov parameters in Appendix C. The first theorem provides a bound with sharp bias terms for all step-sizes, except for the so-called *optimal* step-size $\gamma = \frac{2}{L+\mu}$ for which the variance explodes. The second theorem provides a bound valid for every step-size (including the optimal one), with an ε-sub-optimal bias term which allows for a finite variance term.

Theorem 3.3 (Strongly convex case, sharp bias)**.** *Let Assumption 3.1 hold, and let (x_t) be generated by* (SGD) *with $\gamma L \in (0, 2)$ such that $\gamma \neq \gamma_{opt} = \frac{2}{\mu+L}$. Then, for every $T \geq 1$,*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2 + \frac{1 - \phi^{2T}}{1 - \phi^2} e\sigma_*^2,$$

where $\phi = \max\{1 - \gamma\mu; \gamma L - 1\} \in [0, 1)$, and $e = \gamma^2 \left(1 + \frac{\gamma(L-\mu)}{|\gamma - \gamma_{opt}|(L+\mu)}\right)$.

Theorem 3.4 (Strongly convex case, sub-optimal bias)**.** *Let Assumption 3.1 hold, and let (x_t) be generated by* (SGD) *with $\gamma L \in (0, 2)$. Then, for every $T \geq 1$ and $\varepsilon > 0$,*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2 + \frac{1 - \phi^{2T}}{1 - \phi^2} e\sigma_*^2,$$

where $\phi^2 = \varepsilon + (\max\{1 - \gamma\mu; \gamma L - 1\})^2 \in [0, 1)$, and $e = \gamma^2 \left(1 + \frac{\gamma^2(L-\mu)^2}{4\varepsilon}\right)$.

As in the convex case, when interpolation holds (see Section 6.3 for more details), namely when $\sigma_*^2 = 0$, the division by ε is no longer a problem, allowing us to take $\varepsilon \to 0$.

Corollary 3.5 (Strongly convex case with interpolation, optimal bias)**.** *Let Assumption 3.1 hold, and assume that* $\sigma_*^2 = 0$*. Let* (x_t) *be generated by* (SGD) *with* $\gamma L \in (0, 2)$*. Then, for every* $T \geq 1$*,*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2,$$

where $\phi = \max\{1 - \gamma\mu; \gamma L - 1\} \in [0, 1)$*.*

The first message to take from our results is that it is possible to have bounds for SGD whose bias term corresponds exactly to the standard rate for GD, namely a geometrical rate governed by ϕ^2. A second insight is that it is possible to obtain bounds for the full range of step-sizes $\gamma L \in (0, 2)$. This is in contrast with previous studies, of which we detail the results over the following paragraphs.

Comparison to Bach and Moulines (2011). In (Bach and Moulines, 2011, Theorem 1), the authors analyze the convergence of SGD for strongly convex objectives without variance assumptions, aligning with the setting of our Theorem 3.3. They present the following bound:

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq 2\exp(4L^2\gamma^2(T-1)) \cdot \exp(-\mu\gamma T/4) \cdot (\|x_0 - x_*\|^2 + \sigma_*^2/L^2) + 4\gamma\sigma_*^2/\mu.$$

Their proof is based on a Lyapunov energy function similar to the one employed in our analysis, though with different coefficients. The exponential form is presented for clarity in the context of their use of decreasing step-sizes. For the constant step-size scenario, their analysis yields

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \varphi^{2T} \cdot \|x_0 - x_*\|^2 + 2\gamma^2 \cdot \frac{1 - \varphi^T}{1 - \varphi} \cdot \sigma_*^2,$$

where $\varphi^2 = 1 - 2\mu\gamma + 2L^2\gamma^2 = \phi^2 + \gamma^2(L^2 - \mu^2)$. It may be verified that $\phi \leq \varphi$, where ϕ is defined as in Theorem 3.3, which connects their result to ours. In specific, our analysis explicitly establishes convergence over a broader range of step-sizes γ compared to the range considered in their work, and provides a sharper rate.

Comparison to Needell et al. (2016). In (Needell et al., 2016, Theorem 2.1), the authors obtain, under the same assumptions and for $\gamma < \frac{1}{L}$, the bound

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq (1 - 2\gamma\mu(1 - \gamma L))^T \cdot \|x_0 - x_*\|^2 + \frac{\gamma}{\mu(1 - \gamma L)} \cdot \sigma_*^2.$$

We note that the bound is weaker than the one obtained in our results, and is restricted to short step-sizes.

Comparison to Gower et al. (2019). The authors of (Gower et al., 2019, Theorem 3.1) study the quasi-strongly convex setting under the assumption of expected smoothness, which is tightly linked to our assumption (EC$_*$), as put forward in Remark 6.6.

For a constant step-size $\gamma L \in (0, 1/2]$, they derive the bound

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq (1 - \gamma\mu)^T \|x_0 - x_*\|^2 + \frac{2\gamma}{\mu} \cdot \sigma_*^2.$$

Our analysis extends the convergence result to a broader range of step-sizes. Furthermore, within this specific range of step-sizes, our derived convergence rate incorporates a squared term related to $(1 - \gamma\mu)$, differing from the linear term presented in their bound. This is an effect often seen in simpler proofs.

It is remarkable that being able to work with large step-sizes allows us to observe a singular phenomenon for the optimal step-size, which, to our knowledge, was never observed before. In the deterministic setting, the step-size $\gamma = \frac{2}{\mu+L}$ is the only one leading to the optimal geometric rates governed by $\phi^2 = \frac{(L-\mu)^2}{(L+\mu)^2}$. In our stochastic setting, it appears that achieving such a bias term is incompatible with having a finite variance. It seems reasonable to think that achieving the optimal rate for a deterministic method requires such precision that introducing a small perturbation or variance might break it. Such conjecture seems to be validated by our numerical experiments presented in Section 5.

4 Performance Estimation Problem

The *Performance Estimation Problem* (PEP) methodology was originally introduced in Drori and Teboulle (2016) and later refined in Taylor et al. (2017b). The purpose is to convert the analysis of an optimization algorithm into a computationally tractable semi-definite program which can be solved numerically. The primal and dual solutions to this problem then yield the convergence rates and proofs of them, if read appropriately.

In Section 4.1, we introduce the current state-of-the-art on the PEP methodology. Section 4.2 is devoted to how exactly we reformulate the study of (SGD) into a computationally tractable semi-definite program. In Section 4.3 we give a short insight into how the solution to this semi-definite program inspires the mathematical proofs.

We define $\mathcal{F}_{\mu,L}(\mathbb{R}^d)$ to be the set of L-smooth μ-strongly convex functions. We shall often drop the dependency in $\mathbb{R}^d$ when the space is clear from context.

4.1 State-Of-The-Art

Initial Work. The Performance Estimation Problem (PEP) methodology was first introduced in Drori and Teboulle (2016), where the authors considered a relaxation of the problem, which did not yield tight bounds. A significant advancement was made with the innovative introduction of interpolation conditions in Taylor et al. (2017b), which focused on variants of the gradient descent method. The same authors rapidly extended the framework to analyze other first-order methods (Taylor et al., 2017a), the proximal-gradient method (Taylor et al., 2018b), and gradient descent with exact line search (de Klerk et al., 2017). Since its initial study, the gradient descent method has been extensively analyzed (Kim, 2025; Abbaszadehpeivasti et al., 2023), with recent work establishing tight bounds in all settings for constant step-sizes (Rotaru et al., 2024). A link between PEP and sum-of-squares has also been established, allowing to study noisy gradient descent under inexact line search (Tan et al., 2021). Some work has also been done in studying gradient descent with arbitrary pre-fixed step-sizes (Altschuler, 2018; Daccache, 2019; Eloi, 2022), and acceleration through specific step-size schedules has been obtained for dynamic schedules (Teboulle and Vaisbourd, 2023) or periodic step-size (Grimmer, 2024). More recently, anytime acceleration through a specific step-size schedule has been obtained (Zhang et al., 2024).

Other Methods. The PEP framework has been adapted and expanded by various researchers to study a wide range of optimization algorithms. For instance, it has been applied to analyze biased stochastic gradient descent (Hu et al., 2021), the last-iterate convergence of the subgradient method (Zamani and Glineur, 2023), the alternating direction method of multipliers (ADMM) (Zamani et al., 2024), primal-dual methods (Bousselmi et al., 2024b), and methods involving linear operators (Bousselmi et al., 2024a). Beyond first-order methods, the framework has also been extended to study non-first-order methods, such as Newton-type methods (De Klerk et al., 2020) and nonlinear conjugate gradient methods (Das Gupta et al., 2024a). The methodology can also be extended beyond the Euclidean setting, such as to Bregman methods (Dragomir et al., 2022).

Beyond Convexity. In addition to convex optimization, the PEP methodology has been successfully applied to nonconvex function classes. Notable examples include hypoconvex functions (Rotaru et al., 2022, 2024), difference-of-convex functions (Abbaszadehpeivasti et al., 2024), and

Polyak-Lojasiewicz functions (Abbaszadehpeivasti et al., 2022), although the latter involves a relaxation of the original problem. Functions satisfying a lower restricted secant inequality and an upper error bound have been studied in Guille-Escuret et al. (2022).

Algorithm Design. Once a convergence rate can be systematically derived, a natural subsequent question arises: how to select the optimal parameters for the algorithm. Work such as Kim and Fessler (2021), Goujaud et al. (2022) and Zhou et al. (2022) use this methodology to develop optimal algorithms for convex, convex with quadratic upper bound, and sum-of-convex optimization. The work Guille-Escuret et al. (2022) proves that vanilla gradient descent is optimal amongst all first-order algorithms satisfying a lower restricted secant inequality and an upper error bound. In general however, this leads to a saddle-point problem, where the goal is to minimize the worst-case performance over all possible parameter choices. By leveraging duality, this problem can be reformulated as a single-level optimization problem. However, this reformulation typically results in a nonconvex problem, which is generally challenging to solve. Recently, Das Gupta et al. (2024b) addressed this issue using a branch-and-bound technique, providing a promising approach to tackling the nonconvexity. Despite this advancement, the exploration of optimal parameter selection within the PEP framework remains largely underexplored.

Beyond Optimization. The versatility of the PEP framework extends beyond optimization to other areas, such as operator splitting methods (Ryu et al., 2020) and feasibility problems (Luner and Grimmer, 2024).

Lyapunov Analysis. Furthermore, it has been employed to study Lyapunov analyses for optimization algorithms. The initial work in this direction was presented in Taylor et al. (2018a), with subsequent refinements in Taylor and Bach (2019), where Lyapunov analyses were used to study stochastic methods. This topic will be explored in detail in Section 4.2. Continuous-time Lyapunov analyses have also been investigated in Moucer et al. (2023), and an automated Lyapunov function design methodology based on PEP was proposed in Upadhyaya et al. (2025).

Integral Quadratic Constraints. It is worth noting that another framework, namely integral quadratic constraints (IQCs), has also seen significant development in the systematic analysis of optimization algorithms (Lessard et al., 2016; Scherer et al., 2023). This methodology has also been applied to analyse biased stochastic gradient descent (Hu et al., 2021). While this approach is interesting, it falls outside the scope of this work and will not be further explored.

4.2 Problem Reformulation

Most ideas of this section are inspired by Taylor and Bach (2019).

As we aim to obtain a computationally tractable problem, we fix the value of m in the present scenario. However, the final results will not depend on the fixed value of m, thus recovering the results of Sections 2 and 3.

As explained by Lemmas 2.2 and 3.2, our goal is to derive nonnegative parameters $\rho, (a_t, e_t)$ such that $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for all $t = 0, \dots, T-1$, where

$$E_t = a_t \|x_t - x_*\|^2 + \rho \cdot \sum_{s=0}^{t-1} [f(x_s) - \min f] - \sum_{s=0}^{t-1} e_s \cdot \mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] \quad \text{for } t = 0, \dots, T-1,$$

and $x_* \in \operatorname{argmin} f$. We replaced σ_*^2 by its equivalent value $\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right]$ for simplicity, to avoid the additional variable σ_*^2.

We call parameters $\rho, (a_t), (e_t)$ *Lyapunov parameters* if $\mathbb{E}[E_{t+1} - E_t] \leq 0$ holds for all $t = 0, \ldots, T-1$, for all $d \geq 1$, for all functions $f_1, \ldots, f_m \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, all $x_0 \in \mathbb{R}^d$, and all $x_* \in \operatorname{argmin} f$. Note that being Lyapunov parameters depends on the number of functions m and the time horizon T. We denote the set of all Lyapunov parameters by $\mathcal{V}_T$, where we omit the dependence on m, as our results happen to be independent of m.

To unify the presentation for the convex and the strongly convex setting, we introduce some notation.

- In the convex setting, Lemma 2.2 tells us that we are interested in maximizing ρ, or equivalently minimizing $\frac{1}{\rho}$, where we fix $a_0 = 1$. In this setting we denote the bias and normalization by
$$\text{Bias} = \frac{1}{\rho} \quad \text{and} \quad \mathcal{N}_T = \{\rho, (a_t, e_t) \in \mathcal{V}_T \colon a_0 = 1\}.$$
- In the strongly convex setting, Lemma 3.2 tells us we are interested in minimizing a_0, where we fix $a_T = 1$. In this setting we denote the bias and normalization by
$$\text{Bias} = a_0 \quad \text{and} \quad \mathcal{N}_T = \{\rho, (a_t, e_t) \in \mathcal{V}_T \colon a_T = 1\}.$$

With this notation in mind, we may formally state our problem as

$$\text{Bias}_{\text{opt}} = \text{Bias}_{\text{opt}}(T) \coloneqq \inf_{\rho, (a_t, e_t) \in \mathcal{N}_T} \{\text{Bias}\}.$$

The definition of being a Lyapunov parameter is such that it generates a decrease in energy at each step. Specifically, it holds true that

$$\text{Bias}_{\text{opt}} = \inf_{\mathcal{N}_T} \left\{ \text{Bias} \colon \mathbb{E}[E_{t+1} - E_t] \leq 0 \ \forall t = 0, \ldots, T-1, \forall x_0 \in \mathbb{R}^d, \forall f_i \in \mathcal{F}_{\mu,L}, \forall d \in \mathbb{Z}_{\geq 1} \right\}.$$

By permuting the quantifiers, this may then be reformulated as a bilevel program, namely:

$$\text{Bias}_{\text{opt}} = \inf_{\mathcal{N}_T} \{\text{Bias} \colon B_t \leq 0 \quad \forall t = 0, \ldots, T-1\},$$

where

$$\begin{aligned} B_t = B_t(\rho, (a_t, e_t)) \coloneqq \quad & \sup_{d \in \mathbb{Z}_{\geq 1}} \quad \mathbb{E}[E_{t+1} - E_t] \\ & \text{subject to} \quad x_t \text{ generated through SGD in } t \text{ steps from } x_0, \\ & \qquad\qquad \nabla f(x_*) = 0, \quad x_0, x_* \in \mathbb{R}^d, \quad f_i \in \mathcal{F}_{\mu,L}. \end{aligned}$$

Note that x_0 is a variable of the follower problem, and hence acts as a slack variable for the constraint that x_t is generated from x_0. Using this, and recalling the definition of E_t, we may write that

$$\begin{aligned} B_t = \quad & \sup_{d \in \mathbb{Z}_{\geq 1}} \quad a_{t+1}\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] - a_t\|x_t - x_*\|^2 + \rho(f(x_t) - f(x_*)) - e_t\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] \\ & \text{subject to} \quad \sum_{i=1}^m \nabla f_i(x_*) = 0, \quad x_t, x_* \in \mathbb{R}^d, \quad f_i \in \mathcal{F}_{\mu,L}. \end{aligned}$$

By introducing the notation $f_j^{(i)} = f_i(x_j)$ and $g_j^{(i)} = \nabla f_i(x_j)$ for $j \in \{t, *\}$ and $i = 1, \dots, m$, and by rewriting the expectations as finite sums, we may rewrite the problem equivalently as

$$\begin{aligned} B_t = \quad & \sup_{d \in \mathbb{Z}_{\geq 1}} \quad \frac{a_{t+1}}{m} \sum_{i=1}^m \|x_t - \gamma g_t^{(i)} - x_*\|^2 - a_t \|x_t - x_*\|^2 + \frac{\rho}{m} \sum_{i=1}^m (f_t^{(i)} - f_*^{(i)}) - \frac{e_t}{m} \sum_{i=1}^m \|g_*^{(i)}\|^2 \\ & \text{subject to} \quad \sum_{i=1}^m g_*^{(i)} = 0, \\ & \qquad \text{there exist } f_i \in \mathcal{F}_{\mu,L} \text{ such that } f_i(x_j) = f_j^{(i)} \text{ and } \nabla f_i(x_j) = g_j^{(i)} \\ & \qquad\qquad\qquad\qquad \text{for } j \in \{t, *\} \text{ and } i = 1, \dots, m, \\ & \qquad x_t, x_* \in \mathbb{R}^d, \quad g_t^{(i)}, g_*^{(i)} \in \mathbb{R}^d, \quad f_t^{(i)}, f_*^{(i)} \in \mathbb{R} \quad \text{for all } i = 1, \dots, m. \end{aligned}$$

The existence condition may seem like a complicated reformulation, but it allows us to use (Taylor et al., 2017b, Theorem 4), a cornerstone of the performance estimation framework:

Theorem 4.1. *Let $L \geq \mu \geq 0$ and let $\{(x_i, g_i, f_i)_{i \in \mathcal{I}}\} \subset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be a finite set of triplets. There exists a function $f \in \mathcal{F}_{\mu,L}$ such that*

$$f(x_i) = f_i \quad \text{and} \quad \nabla f(x_i) = g_i \quad \text{for all } i \in \mathcal{I},$$

if, and only if, for every pair of indices $(i, j) \in \mathcal{I}^2$, it holds that

$$f_i - f_j - \langle g_j, x_i - x_j \rangle \geq \frac{1}{2(1 - \frac{\mu}{L})} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2 \frac{\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right). \tag{1}$$

Using this theorem, we may rewrite our program equivalently as

$$\begin{aligned} B_t = \quad & \sup_{d \in \mathbb{Z}_{\geq 1}} \quad \frac{a_{t+1}}{m} \sum_{i=1}^m \|x_t - \gamma g_t^{(i)} - x_*\|^2 - a_t \|x_t - x_*\|^2 + \frac{\rho}{m} \sum_{i=1}^m (f_t^{(i)} - f_*^{(i)}) - \frac{e_t}{m} \sum_{i=1}^m \|g_*^{(i)}\|^2 \\ & \text{subject to} \quad \left\| \sum_{i=1}^m g_*^{(i)} \right\|^2 = 0, \\ & \qquad (x_t, g_t^{(i)}, f_t^{(i)}) \text{ and } (x_*, g_*^{(i)}, f_*^{(i)}) \text{ satisfy Inequality (1) for all } i = 1, \dots, m, \\ & \qquad (x_*, g_*^{(i)}, f_*^{(i)}) \text{ and } (x_t, g_t^{(i)}, f_t^{(i)}) \text{ satisfy Inequality (1) for all } i = 1, \dots, m, \\ & \qquad x_t, x_* \in \mathbb{R}^d, \quad g_t^{(i)}, g_*^{(i)} \in \mathbb{R}^d, \quad f_t^{(i)}, f_*^{(i)} \in \mathbb{R} \quad \text{for all } i = 1, \dots, m. \end{aligned}$$

The variable $g_*^{(m)}$ is redundant, as it is given by $g_*^{(m)} = -\sum_{i=1}^{m-1} g_*^{(i)}$. To simplify notation, we introduce

$$P_t = \left(g_t^{(1)}, \dots, g_t^{(m)}, g_*^{(1)}, \dots, g_*^{(m-1)}, x_t - x_* \right) \in \mathbb{R}^{d \times 2m}, \tag{2a}$$

$$F_t = \left(f_t^{(1)}, \dots, f_t^{(m)}, f_*^{(1)}, \dots, f_*^{(m)} \right) \in \mathbb{R}^{2m}, \tag{2b}$$

and define the vectors $\boldsymbol{p}_i \in \mathbb{R}^{2m}$ for $i = 1, \dots, 2m + 1$, such that

$$\begin{cases} P_t \boldsymbol{p}_i = g_t^{(i)} & \text{for } i = 1, \dots, m, \\ P_t \boldsymbol{p}_{m+i} = g_*^{(i)} & \text{for } i = 1, \dots, m-1, \\ P_t \boldsymbol{p}_{2m} = -\sum_{i=1}^{m-1} g_*^{(i)}, & \\ P_t \boldsymbol{p}_{2m+1} = x_t - x_*, & \end{cases}$$

and $\boldsymbol{f}_i \in \mathbb{R}^{2m}$ for $i = 1, \dots, 2m$ such that, for $i = 1, \dots, m$,

$$F_t \boldsymbol{f}_i = f_t^{(i)}, \quad F_t \boldsymbol{f}_{m+i} = f_*^{(i)}.$$

With this, we may write our program as

$$\begin{aligned} B_t = \quad & \sup_{d \in \mathbb{Z}_{\geq 1}} P_t^T \Delta_t P_t + F_t \tilde{\Delta} \\ & \text{subject to} \quad F_t \boldsymbol{f}_{m+i} - F_t \boldsymbol{f}_i + P_t^T A_{t,*}^{(i)} P_t \leq 0 \text{ for all } i = 1, \dots, m, \\ & \qquad\qquad F_t \boldsymbol{f}_i - F_t \boldsymbol{f}_{m+i} + P_t^T A_{*,t}^{(i)} P_t \leq 0 \text{ for all } i = 1, \dots, m, \\ & \qquad\qquad P_t \in \mathbb{R}^{d \times 2m}, \quad F_t \in \mathbb{R}^{2m}, \end{aligned}$$

where

$$\begin{cases} \Delta_t &= \dfrac{a_{t+1}}{m} \displaystyle\sum_{i=1}^{m} (\boldsymbol{p}_{2m+1} - \gamma \boldsymbol{p}_{m+i})(\boldsymbol{p}_{2m+1} - \gamma \boldsymbol{p}_{m+i})^T - a_t \boldsymbol{p}_{2m+1} \boldsymbol{p}_{2m+1}^T \\ & \qquad - \dfrac{e_t}{m} \displaystyle\sum_{i=1}^{m} \boldsymbol{p}_{m+i} \boldsymbol{p}_{m+i}^T, \\ \tilde{\Delta} &= \dfrac{\rho}{m} \displaystyle\sum_{i=1}^{m} \boldsymbol{f}_i - \boldsymbol{f}_{m+i}, \\ A_{t,*}^{(i)} &= \boldsymbol{p}_{m+i} \boldsymbol{p}_{2m+1}^T + \dfrac{1}{2(L-\mu)} (\boldsymbol{p}_i - \boldsymbol{p}_{m+i})(\boldsymbol{p}_i - \boldsymbol{p}_{m+i})^T + \dfrac{\mu L}{2(L-\mu)} \boldsymbol{p}_{2m+1} \boldsymbol{p}_{2m+1}^T \\ & \qquad - \dfrac{\mu}{L-\mu} (\boldsymbol{p}_i - \boldsymbol{p}_{m+i}) \boldsymbol{p}_{2m+1}^T, \\ A_{*,t}^{(i)} &= -\boldsymbol{p}_i \boldsymbol{p}_{2m+1}^T + \dfrac{1}{2(L-\mu)} (\boldsymbol{p}_i - \boldsymbol{p}_{m+i})(\boldsymbol{p}_i - \boldsymbol{p}_{m+i})^T + \dfrac{\mu L}{2(L-\mu)} \boldsymbol{p}_{2m+1} \boldsymbol{p}_{2m+1}^T \\ & \qquad - \dfrac{\mu}{L-\mu} (\boldsymbol{p}_i - \boldsymbol{p}_{m+i}) \boldsymbol{p}_{2m+1}^T. \end{cases}$$

This formulation is a non-convex quadratic program. A standard technique to render this solvable is to cast it into a semi-definite program. We do so by introducing $G_t = P_t^T P_t \succeq 0$, and obtain

$$\begin{aligned} B_t = \sup_{d \in \mathbb{Z}_{\geq 1}} \quad & \sup_{G_t \in \mathcal{S}^{2m}, F_t \in \mathbb{R}^{2m}} \quad \operatorname{Tr}(\Delta_t G_t) + F_t \tilde{\Delta} \\ & \text{subject to} \quad F_t \boldsymbol{f}_{m+i} - F_t \boldsymbol{f}_i + \operatorname{Tr}(A_{t,*}^{(i)} G_t) \leq 0 \text{ for all } i = 1, \dots, m \\ & \qquad\qquad F_t \boldsymbol{f}_i - F_t \boldsymbol{f}_{m+i} + \operatorname{Tr}(A_{*,t}^{(i)} G_t) \leq 0 \text{ for all } i = 1, \dots, m \\ & \qquad\qquad \operatorname{rank}(G_t) \leq d, \end{aligned}$$

where $\mathcal{S}^k$ denotes the set of symmetric positive definite matrices of dimension $k \times k$. This formulation is equivalent to the previous one due to the rank condition: given any matrix $G_t \in \mathcal{S}^k$ of rank lesser than d, a Cholesky factorization $G_t = P_t^T P_t$ would yield a $P_t \in \mathbb{R}^{d \times 2m}$, which is feasible to the previous formulation. Now, observe that the non-convex non-continuous rank constraint may be removed due to the exterior supremum over $d \geq 1$. Once this is done, we may as well remove

the supremum over d because it does not appear anymore in the inner supremum. This allows us to write

$$\begin{aligned} B_t = \quad & \sup_{G_t \in \mathcal{S}^{2m}, F_t \in \mathbb{R}^{2m}} \quad \operatorname{Tr}(\Delta_t G_t) + F_t \tilde{\Delta} \qquad \text{(Primal)} \\ & \text{subject to} \quad F_t \boldsymbol{f}_{m+i} - F_t \boldsymbol{f}_i + \operatorname{Tr}(A_{t,*}^{(i)} G_t) \leq 0 \text{ for all } i = 1, \dots, m, \\ & \qquad\qquad F_t \boldsymbol{f}_i - F_t \boldsymbol{f}_{m+i} + \operatorname{Tr}(A_{*,t}^{(i)} G_t) \leq 0 \text{ for all } i = 1, \dots, m. \end{aligned}$$

We have now expressed B_t as the optimal value of a semi-definite program. It is a simple exercise to compute the dual of this problem, which happens to be a feasibility problem:

$$\begin{aligned} \tilde{B}_t = \quad & \sup_{\Lambda_t, \lambda_{t,*}^{(i)}, \lambda_{*,t}} \quad 0 \qquad \text{(Dual)} \\ & \text{subject to} \quad -\Delta_t + \sum_{i=1}^{m} \left(\lambda_{t,*}^{(i)} A_{t,*}^{(i)} + \lambda_{*,t}^{(i)} A_{*,t}^{(i)} \right) = \Lambda_t, \\ & \qquad\qquad -\tilde{\Delta} + \sum_{i=1}^{m} \lambda_{t,*}^{(i)} (\boldsymbol{f}_{m+i} - \boldsymbol{f}_i) + \sum_{i=1}^{m} \lambda_{*,t}^{(i)} (\boldsymbol{f}_i - \boldsymbol{f}_{m+i}) = 0, \\ & \qquad\qquad \Lambda_t \in \mathcal{S}^{2m}, \quad \lambda_{t,*}^{(i)}, \lambda_{*,t}^{(i)} \in \mathbb{R}_{\geq 0} \quad \text{for all } i = 1, \dots, m. \end{aligned}$$

We prove in Appendix D that this is indeed the dual problem of Problem (Primal). We further prove that strong duality holds, meaning that $\tilde{B}_t = -B_t$. As a consequence,

$$B_t \leq 0 \iff \tilde{B}_t \geq 0 \iff \text{Problem (Dual) is feasible,}$$

where the last equivalence follows from the structure of Problem (Dual). In conclusion, we can rewrite the problem of minimizing Bias over the Lyapunov parameters into

$$\begin{aligned} \text{Bias}_{\text{opt}} = \quad & \inf_{\mathcal{N}_T} \quad \text{Bias} \\ & \text{subject to} \quad -\Delta_t + \sum_{i=1}^{m} \left(\lambda_{t,*}^{(i)} A_{t,*}^{(i)} + \lambda_{*,t}^{(i)} A_{*,t}^{(i)} \right) = \Lambda_t, \\ & \qquad\qquad -\tilde{\Delta} + \sum_{i=1}^{m} \lambda_{t,*}^{(i)} (\boldsymbol{f}_{m+i} - \boldsymbol{f}_i) + \sum_{i=1}^{m} \lambda_{*,t}^{(i)} (\boldsymbol{f}_i - \boldsymbol{f}_{m+i}) = 0, \\ & \qquad\qquad \Lambda_t \in \mathcal{S}^{2m}, \quad \lambda_{t,*}^{(i)}, \lambda_{*,t}^{(i)} \in \mathbb{R}_{\geq 0} \quad \text{for all } i = 1, \dots, m \text{ and } t = 0, \dots, T-1. \end{aligned}$$

This formulation is a finite-dimensional semi-definite program that can be solved numerically.

4.3 Obtaining Mathematical Proofs

The Lagrangian attached to Problem (Primal) is given by

$$\begin{aligned} L(G_t, F_t, \Lambda_t, \lambda_{t,*}^{(i)}, \lambda_{*,t}^{(i)}) = & -\operatorname{Tr}(\Delta_t G_t) - F_t \tilde{\Delta} - \operatorname{Tr}(\Lambda_t G_t) \\ & + \sum_{i=1}^{m} \lambda_{t,*}^{(i)} \left(F_t \boldsymbol{f}_{m+i} - F_t \boldsymbol{f}_i + \operatorname{Tr}(A_{t,*}^{(i)} G_t) \right) \\ & + \sum_{i=1}^{m} \lambda_{*,t}^{(i)} \left(F_t \boldsymbol{f}_i - F_t \boldsymbol{f}_{m+i} + \operatorname{Tr}(A_{*,t}^{(i)} G_t) \right). \end{aligned}$$

If we have a primal-dual optimal solution $(\overline{G}_t, \overline{F}_t, \overline{\Lambda}_t, (\overline{\lambda}^{(i)}_{t,*}), (\overline{\lambda}^{(i)}_{*,t}))$, it holds that, for all $G_t \in \mathcal{S}^{2m}$ and $F_t \in \mathbb{R}^{2m}$,

$$L(\overline{G}_t, \overline{F}_t, \overline{\Lambda}_t, (\overline{\lambda}^{(i)}_{t,*}), (\overline{\lambda}^{(i)}_{*,t})) \leq L(G_t, F_t, \overline{\Lambda}_t, (\overline{\lambda}^{(i)}_{t,*}), (\overline{\lambda}^{(i)}_{*,t})).$$

Due to complementary slackness, it holds that,

$$L(\overline{G}_t, \overline{F}_t, \overline{\Lambda}_t, (\overline{\lambda}^{(i)}_{t,*}), (\overline{\lambda}^{(i)}_{*,t})) = -\operatorname{Tr}(\Delta_t \overline{G}_t) - \overline{F}_t \tilde{\Delta} = B_t.$$

By strong duality (see Appendix D), we know that the existence of a primal-dual solution implies that $B_t = 0$. As such, it holds that, for all $G_t \in \mathcal{S}^{2m}$ and $F_t \in \mathbb{R}^{2m}$,

$$0 \leq L(G_t, F_t, \overline{\Lambda}_t, (\overline{\lambda}^{(i)}_{t,*}), (\overline{\lambda}^{(i)}_{*,t})),$$

or in other terms that

$$\begin{aligned} \operatorname{Tr}(\Delta_t G_t) + F_t \tilde{\Delta} + \operatorname{Tr}(\overline{\Lambda}_t G_t) \leq & \sum_{i=1}^{m} \overline{\lambda}^{(i)}_{t,*} \left(F_t \boldsymbol{f}_{m+i} - F_t \boldsymbol{f}_i + \operatorname{Tr}(A^{(i)}_{t,*} G_t) \right) \\ & + \sum_{i=1}^{m} \overline{\lambda}^{(i)}_{*,t} \left(F_t \boldsymbol{f}_i - F_t \boldsymbol{f}_{m+i} + \operatorname{Tr}(A^{(i)}_{*,t} G_t) \right). \end{aligned} \tag{3}$$

As $(\overline{G}_t, \overline{F}_t, \overline{\Lambda}_t, (\overline{\lambda}^{(i)}_{t,*}), (\overline{\lambda}^{(i)}_{*,t}))$ is primal-dual optimal, in particular it is primal-dual feasible, and hence $\overline{\Lambda}_t \succeq 0$ and $\overline{\lambda}^{(i)}_{t,*}, \overline{\lambda}^{(i)}_{*,t} \geq 0$ for all $i = 1, \ldots, m$. In particular, if G_t and F_t are generated by functions $f_i \in \mathcal{F}_{\mu,L}$ through Equations (2), it holds that the right-hand side of Equation (3) is nonpositive by Theorem 4.1. By recalling that $\mathbb{E}[E_{t+1} - E_t] = \operatorname{Tr}(\Delta_t G_t) + F_t \tilde{\Delta}$, we write

$$\mathbb{E}[E_{t+1} - E_t] \leq -\operatorname{Tr}(\overline{\Lambda}_t G_t) \leq 0,$$

where the latter follows since $\overline{\Lambda}_t, G_t \succeq 0$. Specifically, one can prove that $\mathbb{E}[E_{t+1} - E_t] \leq 0$ by rewriting it in the form of (3), using the obtained dual optimal solution, and writing $\operatorname{Tr}(\overline{\Lambda}_t G_t)$ as an appropriate sum of squares based on the Cholesky factorization of $\overline{\Lambda}_t$.

5 Sharpness through Numerical Experiments

In this section, we delve into discussing the tightness of our bounds. To this end, we employ a numerical approach based on the *Performance Estimation Problem* (PEP) methodology. We direct the reader to Section 4 for an in-depth explanation of our PEP framework. Roughly speaking, the idea consists in reformulating the problem of finding admissible Lyapunov parameters ensuring the energy decrease into a semi-definite program, which may be solved numerically. For this thesis, we focused on finding admissible coefficients yielding minimal bias terms in our bounds.

This framework inspired and drove our proofs, helping us to postulate the best Lyapunov parameters that could be found. We made this clear in our proofs, which can be found in the Appendices B and C. Although most of our theorems in Sections 2 and 3 are stated using simplified values, for an easier presentation, in this section we will make use of the exact values found in our proofs. Moreover, it allowed us to verify the tightness of our bounds, at least numerically. We recognize that theoretical results overlapping with numerical observations do not constitute a proof of optimality. Nevertheless, these validations allow us to formulate conjectures that our results are sharp, some of which we partially prove. This tightness assessment will be the focus of the rest of this section.

All of our numerical experiments[1] were developed in Python 3.13 using the solvers MOSEK version 11.0.19 (MOSEK, 2025) and Clarabel version 0.9.0 (Goulart and Chen, 2024) to solve the resulting semi-definite programs, implemented using CVXPY version 1.6.5 (Diamond and Boyd, 2016), and run on Intel Xeon Platinum 8380 CPUs.

5.1 Convex Setting

The primary focus of our work has been to identify sharp bias terms, which directly translates to maximizing the value of ρ via Lemma 2.2, once we assume, without loss of generality, that $a_0 = 1$. We denote by ρ_{opt} the supremum of all such possible admissible constants ρ. In parallel, we define ρ_{theory} to be our guess for ρ_{opt} driven by our results in Theorems 2.3, 2.4 and 2.7:

$$\rho_{\text{theory}} \coloneqq \begin{cases} 2\gamma + \frac{2}{TL}(1-\gamma L) & \text{if } \gamma L \in (0,1), \\ 2\gamma & \text{if } \gamma L = 1, \\ 2\gamma(2-\gamma L)\left[1-(1-\gamma L)^{2T}\right]^{-1} & \text{if } \gamma L \in (1,2). \end{cases}$$

Figure 1 compares our analytically derived ρ_{theory} with the value of ρ_{opt} obtained numerically. The perfect match across the full range of step-sizes validates the sharpness of the bias term in our theoretical bounds for step-sizes $\gamma L \neq 1$. This empirical sharpness is validated by a formal proof, for large step-sizes (see Proposition B.8 for more details).

Proposition 5.1. *For large step-sizes $\gamma L \in (1,2)$, there exists no Lyapunov parameters with $a_0 = 1$ and $\rho > \rho_{theory}$ making the energy decrease.*

For the optimal step-size $\gamma L = 1$, we empirically see that we cannot do better than 2γ, but we were not able to achieve it in Theorem 2.4. To understand why, we turn to the study of the variance term.

[1] Code available on https://github.com/DanielCortild/PEP-for-SGD-without-Variance.

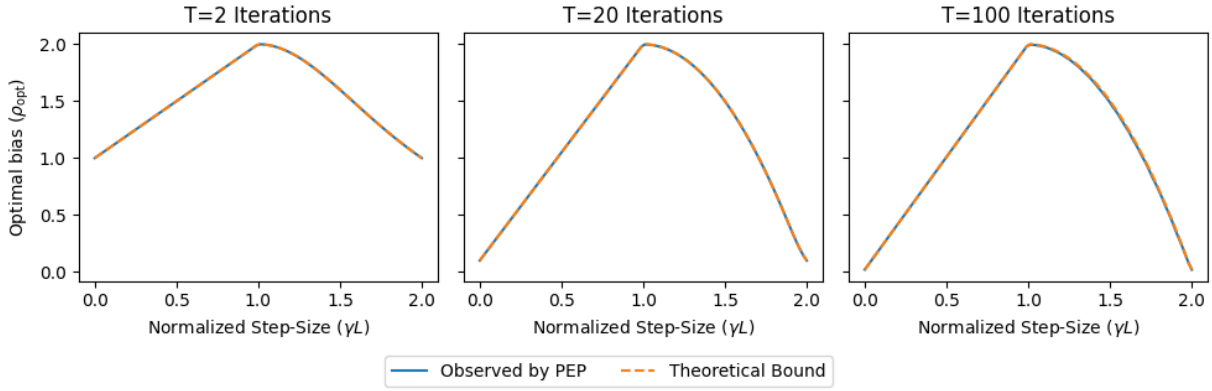

Figure 1: Theoretical and numerical bias term ($L = 1$).

As a second set of experiments, we minimized the constant $\overline{e}$ (which governs the variance term in our bound from Lemma 2.2) under the constraint that $\rho = \rho_{\text{theory}}$. We will denote the infimum of such constants as $\overline{e}_{\text{opt}}$, and will note $\overline{e}_{\text{theory}}$ the constant we obtained in our results. We compare the values of those two constants in Figure 2, for non-optimal step-sizes. We can see that, as $\gamma L \to 1$, either from the left or from the right, the variance term tends to infinity, suggesting that it is not possible to maintain $\rho = \rho_{\text{opt}}$ when $\gamma L = 1$. Moreover, the solver fails when trying to compute $\overline{e}_{\text{opt}}$ for this optimal step-size, which again points to the fact that $\rho = 2\gamma$ is not a feasible Lyapunov parameter when $\gamma L = 1$. This is why in Theorem 2.4 we proposed a bias term with any $\rho_\varepsilon < \rho_{\text{opt}}$, in view of maintaining a finite variance. In consequence we conjecture that, within such Lyapunov framework, it is not possible to obtain a bound for SGD with a bias term of the order $\frac{1}{2\gamma T}$ when $\gamma L = 1$.

A bi-product of the results displayed in Figure 2 is that the variance terms we obtained in Theorems 2.3 and 2.7 overlap with the variance returned by the solver. This suggests that we cannot avoid the issue we had with Theorem 2.7, where the variance term grows exponentially with time.

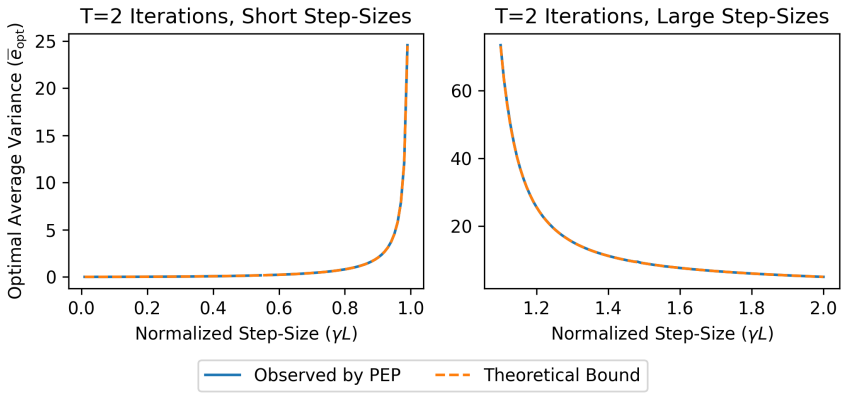

Figure 2: Theoretical and numerical average variance term ($L = 1$).

We note that Figure 2 was made for $T = 2$, as we observe numerical instability for large values of T for the variance, specifically for large step-sizes. For completeness, Figures 3 and 4 represent the variance term for larger values of T. For short step-sizes our theoretical results still match perfectly, whereas for large step-sizes we observe slight deviations, likely due to the large dimension of the underlying semidefinite problem.

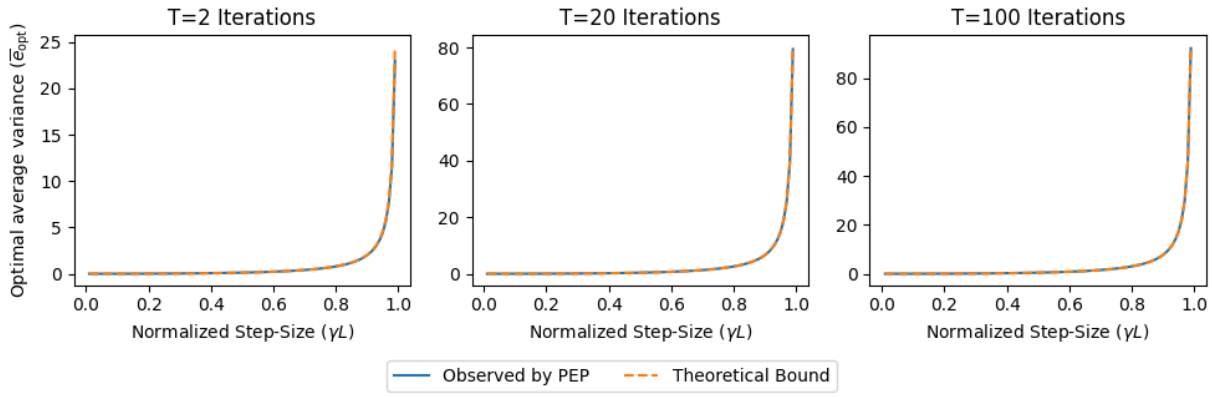

Figure 3: Theoretical and numerical average variance in the convex setting for short step-sizes.

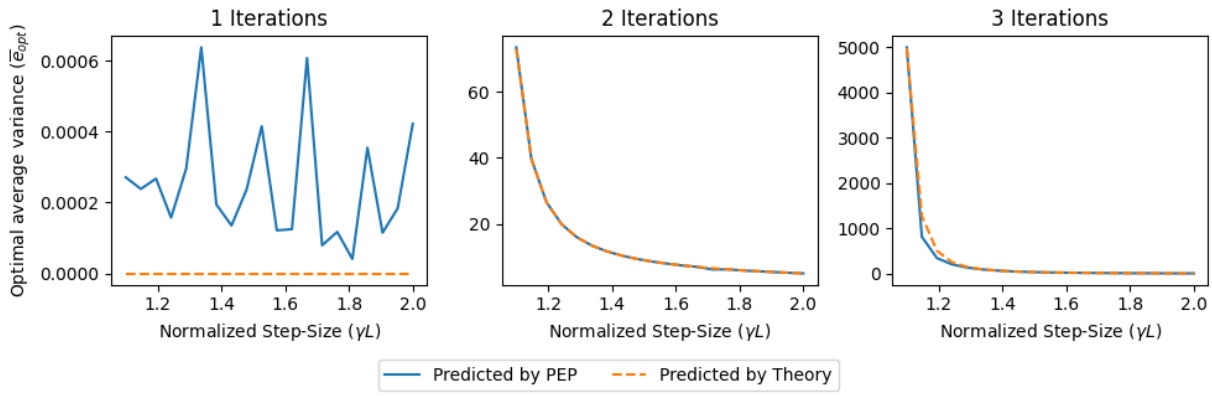

Figure 4: Theoretical and numerical average variance in the convex setting for large step-sizes.

In Figure 5 we study the case of the optimal step-size $\gamma L = 1$, with an ε-suboptimal bias, to contrast with Theorem 2.4. The plot is generated by fixing a rate of $\rho = 2\gamma - \varepsilon$, and minimizing the average variance. We note that the results are not tight due to our restriction that $a_t \equiv 1$ within the proof, which is not the case for $\varepsilon > 0$. Nevertheless, as $T \to \infty$, the PEP bound seems to converge to our (asymptotic) bound. We moreover point out that this non-tightness is not a problem, as we are already studying a suboptimal bias term, and are mainly interested in the behavior as $\varepsilon \to 0$.

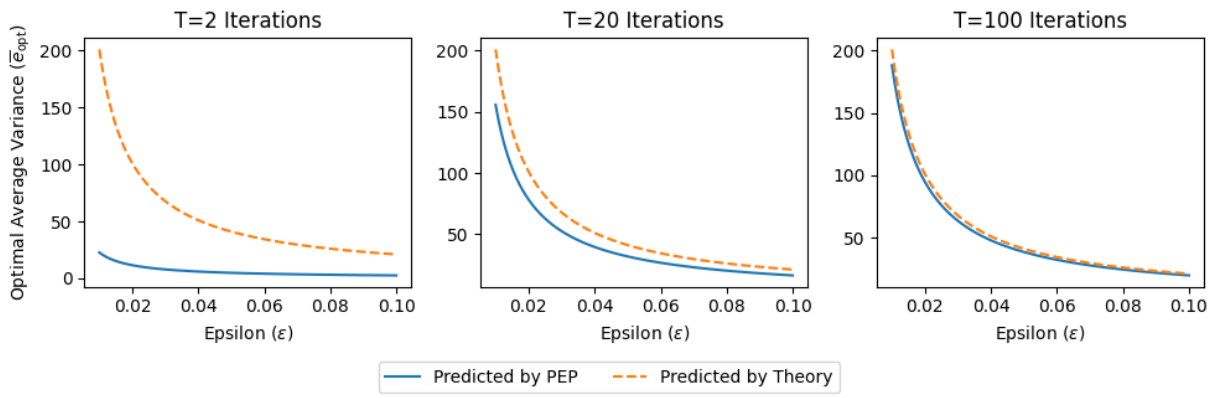

Figure 5: Numerical and theoretical variance for optimal step-size and suboptimal bias term.

Combining Figures 3, 5 and 4, we confirm that it seems impossible to obtain a finite variance for the optimal bias term at the optimal step-size $\gamma L = 1$. This confirms our discussion following Theorem 2.4.

5.2 Strongly Convex Setting

In this setting, our focus has also been to identify sharp bias terms which, through Lemma 3.2, amounts to minimizing the value of a_0, once we assume, without loss of generality, that $a_T = 1$. We denote by a_0^{opt} the infimum of all such possible admissible constants. Driven by our result in Theorem 3.3, we expect a_0^{opt} to be equal to ϕ^{2T}, with $\phi = \max\{1 - \gamma\mu, \gamma L - 1\}$. Figure 6 compares ϕ^{2T} with the value of a_0^{opt} obtained numerically, and illustrates again the tightness of our results. Note that the iteration count is smaller here than in the convex setting, due to numerical instability issues, which are made visible in Figure 7.

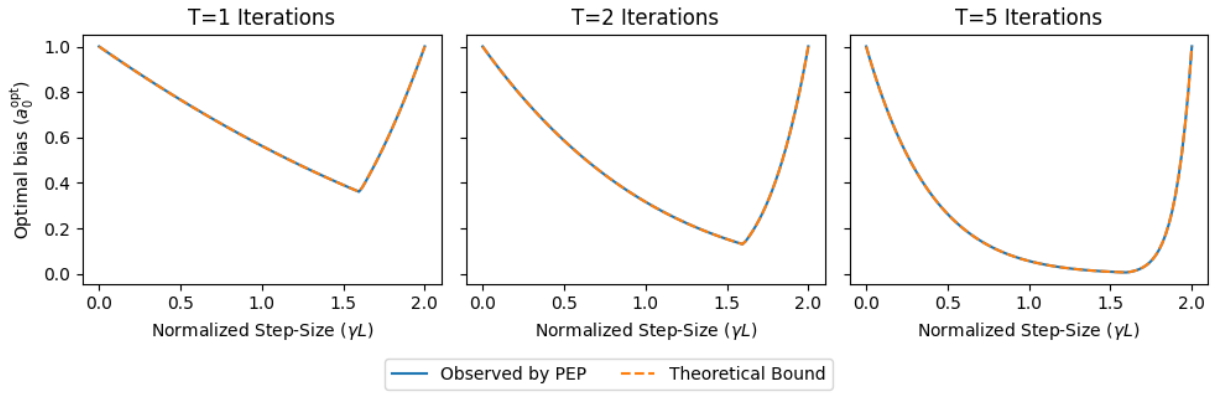

Figure 6: Theoretical and numerical bias term ($L = 1$, $\mu = 0.25$ and $\frac{2}{\mu+L} = 1.6$).

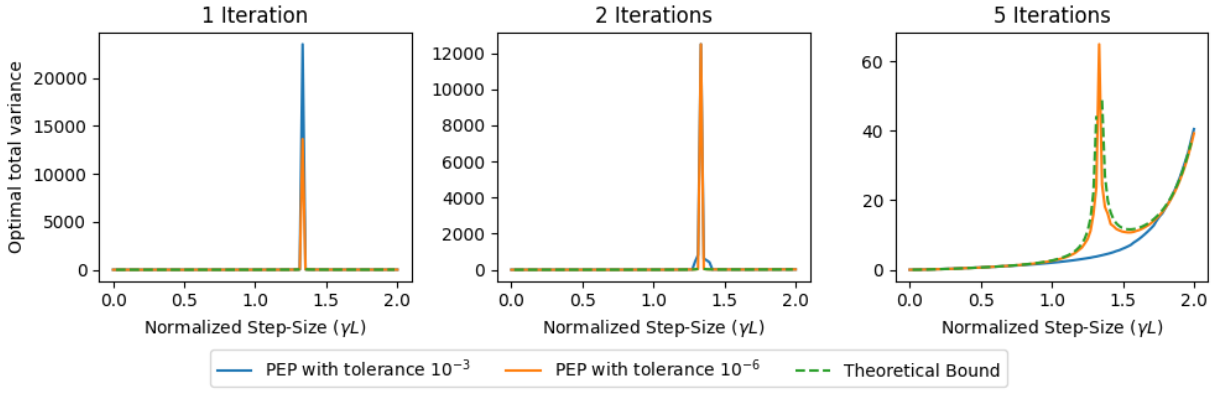

Figure 7: Example of numerical instability for $(L, \mu) = (1, 0.5)$.

For the optimal step-size $\gamma = \frac{2}{L+\mu}$, we empirically see that we cannot do better than ϕ^{2T}, but we were not able to achieve it in Theorem 3.3. Here again we will understand why by looking at the variance terms. To do so, we implement the minimization of the constant e^{sum} appearing in Lemma 3.2 under the constraint that a_0 is equal to ϕ^{2T}. Figure 8 compares the optimal value e^{sum}_{opt} to the constant we obtained in our proofs. We can see that, as the step-size gets closer to the optimal one, the variance tends to $+\infty$. This strongly suggests that it is not possible to obtain a bias term of the order $\left(\frac{L-\mu}{L+\mu}\right)^{2T}$ by means of a Lyapunov analysis such as ours.

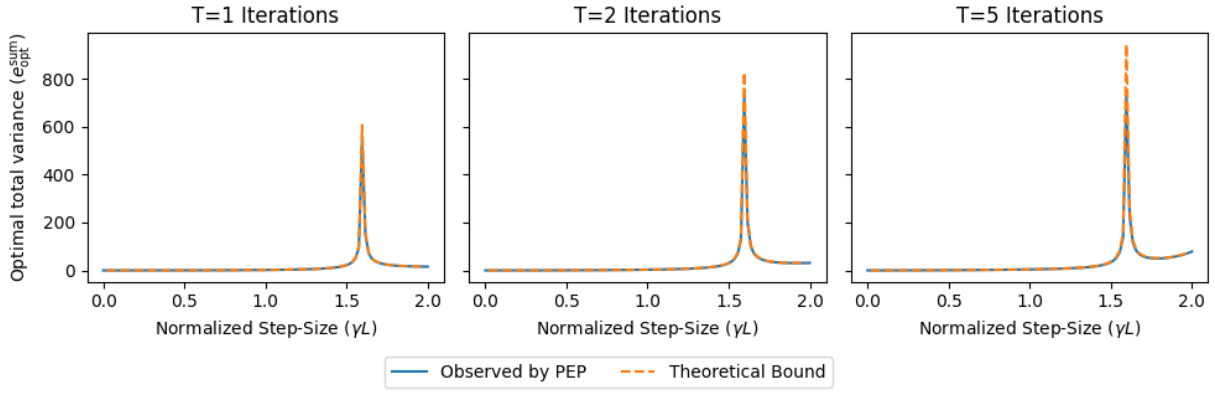

Figure 8: Theoretical and numerical total variance term ($L = 1$, $\mu = 0.25$ and $\frac{2}{\mu+L} = 1.6$).

Figure 8 studies a specific value of μ. In order to confirm the singularity at the optimal step-size, we define the relative error τ between the theoretical and numerical average variance as

$$\tau = \frac{\overline{e}_{\text{opt}} - \overline{e}_{\text{theory}}}{\overline{e}_{\text{theory}}},$$

where $\overline{e}_{\text{theory}}$ is the theoretical average variance term obtained in Theorem 3.3. Figure 9 shows this relative error for various values of T over a range of values of step-sizes γL and strong convexity parameters μ.

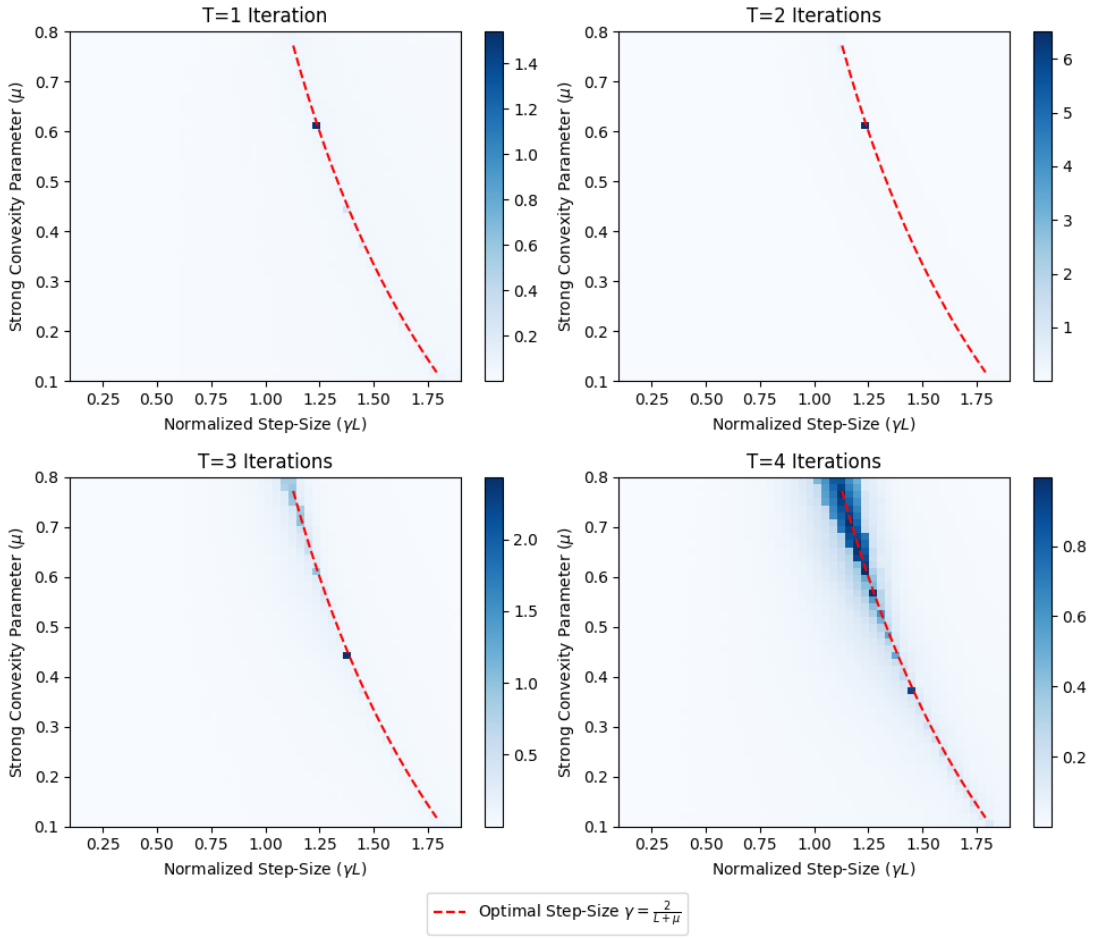

Figure 9: Relative errors τ in optimal average variance.

In Figure 10 we plot the numerical and theoretical variance term upon fixing an ε-tight bias term. Our bounds might not appear tight, but we suspect this to be due to numerical instabilities. To emphasize this, we include the same results using a different solver in Figure 11, whose results are widely different. We leave this for further investigation. In either case, we observe that the variance term approaches $+\infty$ as $\varepsilon \to 0$, thus numerically validating Theorem 3.4.

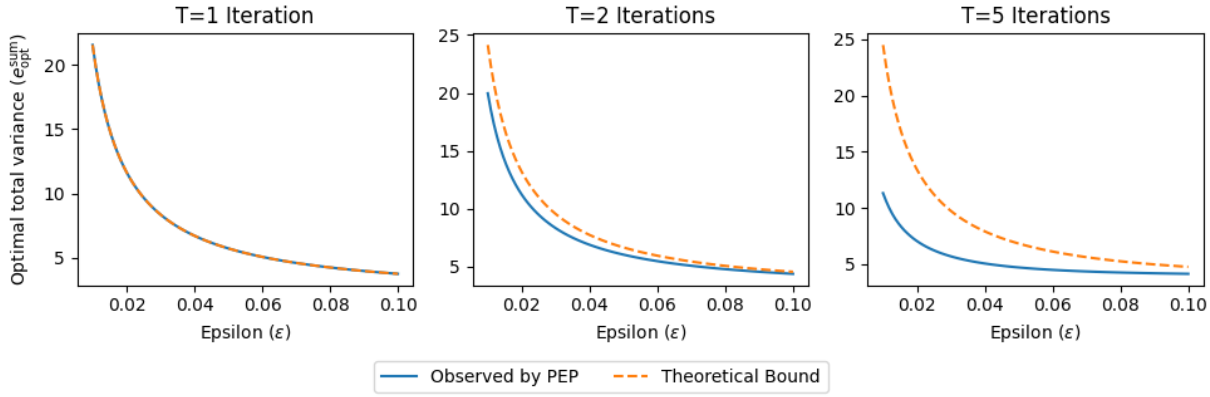

Figure 10: Theoretical and numerical variance in the strongly convex setting with $\mu = 0.25$, in the optimal step-size (using MOSEK).

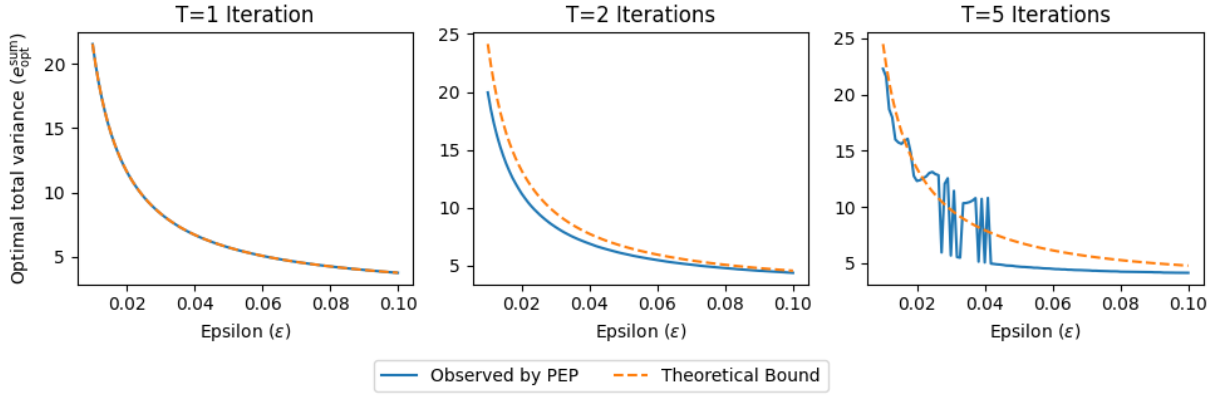

Figure 11: Theoretical and numerical variance in the strongly convex setting with $\mu = 0.25$, in the optimal step-size (using CLARABEL).

6 Relaxation of Problem Statement and Assumptions

In theory, it is possible to consider a more general setting than in Sections 2 and 3. The technical proofs postponed to the appendix are assuming the more general assumptions presented in this section. It is important to note these assumptions are a generalization of the previously provided setting, and hence not invalidating the results of Sections 2 and 3.

Instead of limiting ourselves to the minimization of a finite sum of functions in $\mathbb{R}^d$, we consider the minimization of an expectation of functions defined on a real Hilbert space.

Problem 6.1. *Let $\mathcal{H}$ be a real Hilbert space with associated inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $\{f_i\}_{i \in \mathcal{I}}$ be a family of real-valued functions $f_i \colon \mathcal{H} \to \mathbb{R}$, where $\mathcal{I}$ is a (possibly infinite) set of indices. We consider the problem of minimizing $f := \mathbb{E}[f_i]$, where the expectation is taken over the indices $i \in \mathcal{I}$, with respect to some probability distribution $\mathcal{D}$ over $\mathcal{I}$. We assume that*

1. *the problem is well-defined, in the sense that $i \mapsto f_i(x)$ is $\mathcal{D}$-measurable, and that $\mathbb{E}[f_i(x)]$ is finite for every $x \in \mathcal{H}$;*
2. *the problem is well-posed, in the sense that* $\operatorname{argmin} f \neq \emptyset$*;*
3. *the problem is differentiable, in the sense that each f_i is differentiable, and so is f, with $\nabla f(x) = \mathbb{E}[\nabla f_i(x)]$.*

Within the context of Problem 6.1, we study the performance of the SGD algorithm, which generates a sequence $(x_t)_{t \geq 0}$ according to

$$x_{t+1} = x_t - \gamma \nabla f_{i_t}(x_t), \tag{SGD$_{\mathbb{E}}$}$$

where $\gamma > 0$ is a fixed step-size, and $i_t \in \mathcal{I}$ is sampled i.i.d. from the distribution $\mathcal{D}$. Note that Problem 6.1 and (SGD$_{\mathbb{E}}$) boil down to (Finite-Sum) and (SGD) in the setting introduced in the earlier sections.

6.1 Smoothness and Convexity

The main assumptions that we presented previously were smoothness and convexity (or strong convexity) of the functions f_i. We define a setting combining both convexity and strong convexity, which allows us to unify parts of the analyses. We also define a more relaxed setting, which is all that is needed for our results to hold.

Assumption 6.2 (Smoothness and μ-strong convexity)**.** *Considering Problem 6.1, we assume that*

- *there exists $L \in (0, +\infty)$ such that each function f_i is L-smooth, in the sense that $\nabla f_i \colon \mathcal{H} \to \mathcal{H}$ is L-Lipschitz continuous;*
- *there exists $\mu \in [0, +\infty)$ such that each function f_i is μ-strongly convex. When $\mu = 0$, this simply means that f_i is convex.*

Smoothness and μ-strong convexity can be characterized by the means of a variational inequality:

Lemma 6.3 (Expected Cocoercivity)**.** *Let Assumption 6.2 hold. Then, for every* $(x, y) \in \mathcal{H}^2$,

$$\frac{1}{2}\mathbb{E}\left[\|\nabla f_i(y) - \nabla f_i(x)\|^2\right] + \frac{\mu L}{2}\|y - x\|^2 \leq (L - \mu)\left(f(y) - f(x)\right) + \langle \mu \nabla f(y) - L\nabla f(x), y - x\rangle. \tag{EC}$$

Proof. Because f_i is L-smooth and μ-convex, we can use (Taylor et al., 2017b, Theorem 4) to write, at every (x, y),

$$\frac{1}{2}\|\nabla f_i(y) - \nabla f_i(x)\|^2 + \frac{\mu L}{2}\|y - x\|^2 \leq (L - \mu)\left(f_i(y) - f_i(x)\right) + \langle \mu \nabla f_i(y) - L\nabla f_i(x), y - x\rangle.$$

The conclusion follows after taking expectation. □

As will be made clear throughout our proofs, we may relax Assumption 6.2 to only assume that the conclusion of Lemma 6.3, namely (EC), holds between any point x and any minimizer x^*. We will decline this assumption into two flavors.

Assumption 6.4 (Expected Cocoercivity$_*$)**.** *Considering Problem 6.1, we assume that there exist* $L \in (0, +\infty)$ *and* $\mu \in [0, L]$ *such that*

$$\text{for every } x \in \mathcal{H} \text{ and every } x_* \in \operatorname{argmin} f, \text{ (EC) holds for } (x, x_*). \tag{EC$_*$}$$

Assumption 6.5 (Symmetric Expected Cocoercivity$_*$)**.** *Considering Problem 6.1, we assume that there exist* $L \in (0, +\infty)$ *and* $\mu \in [0, L]$ *such that*

$$\text{for every } x \in \mathcal{H} \text{ and every } x_* \in \operatorname{argmin} f, \text{ (EC) holds for } (x, x_*) \text{ and } (x_*, x). \tag{SEC$_*$}$$

Remark 6.6 (Comparing (SEC$_*$) and (EC$_*$) with other assumptions in the literature)**.** *From what precedes, it is clear that*

$$\text{each } f_i \text{ is } \mu\text{-convex and } L\text{-smooth} \Rightarrow \text{(SEC}_*\text{)} \Rightarrow \text{(EC}_*\text{)}.$$

We also note that (EC$_*$) *implies the* μ*-strongly quasi-convexity assumption from Gower et al. (2019). If we focus on the case* $\mu = 0$*, we see that* (SEC$_*$) *holds if, and only if,* (EC$_*$) *and* (ES) *hold, where the latter is the Expected Smoothness property introduced by Gower et al. (2019), stating that, for all* $x \in \mathcal{H}$ *and* $x_* \in \operatorname{argmin} f$,

$$\frac{1}{2L}\mathbb{E}\left[\|\nabla f_i(x_*) - \nabla f_i(x)\|^2\right] \leq f(x) - f(x_*). \tag{ES}$$

6.2 Gradient Variance at the Solution

We also assume that the variance of $\nabla f_i(x)$ exists at the solutions. More precisely,

Assumption 6.7 (Solution Gradient Variance)**.** *Considering Problem 6.1, we assume that the variance at the solution exists, meaning that*

$$\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] < +\infty \tag{GV$_*$}$$

for every $x_* \in \operatorname{argmin} f$*. We will note*

$$\sigma_*^2 \coloneqq \sup_{x_* \in \operatorname{argmin} f} \mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right].$$

It is important to note that under Assumption 6.4 the variance term $\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right]$ does not depend on the choice of $x_* \in \operatorname{argmin} f$, as was already observed in (Garrigos and Gower, 2024, Lemma 4.17).

Lemma 6.8. *Let Assumption 6.4 hold. Then* $\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right]$ *is constant over* $\operatorname{argmin} f$.

Proof. Let x_*, x'_* be two minimizers of f. From Assumption 6.4 we know that (EC_*) holds true for the couple (x_*, x'_*), which means that

$$\frac{1}{2}\mathbb{E}\left[\|\nabla f_i(x'_*) - \nabla f_i(x_*)\|^2\right] \leq (L-\mu)\left(f(x'_*) - f(x_*)\right) + \langle \mu \nabla f(x'_*) - L\nabla f(x'_*), y - x\rangle.$$

Because $f(x'_*) = f(x_*) = \inf f$ and $\nabla f(x'_*) = \nabla f(x_*) = 0$, we deduce that

$$\mathbb{E}\left[\|\nabla f_i(x'_*) - \nabla f_i(x_*)\|^2\right] = 0.$$

Then, almost surely with respect to the distribution $\mathcal{D}$ over the indices $i \in \mathcal{I}$, we have $\|\nabla f_i(x'_*) - \nabla f_i(x_*)\|^2 = 0$. This means that $\nabla f_i(x'_*) = \nabla f_i(x_*)$ almost surely, from which we deduce that

$$\mathbb{E}\left[\|\nabla f_i(x'_*)\|^2\right] = \mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right],$$

meaning that in fact $\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right]$ is constant over $\operatorname{argmin} f$. □

Assuming that the variance at the solution σ_*^2 exists is trivially satisfied in the case of the minimization of a finite sum, as presented in Problem (Finite-Sum).

Lemma 6.9 (Finite-sum problems have finite solution variance)**.** *If Problem 6.1 treats a finite sum of functions, i.e.* $\mathcal{I}$ *is finite, then Assumption 6.7 holds true.*

Even for true expectation-minimization problems, where $\mathcal{I}$ is infinite, the Assumption 6.7 is very mild. In particular, this assumption is automatically verified in problems where the loss functions f_i are *nonnegative*, which is standard for problems arising in inverse problems and machine learning.

Lemma 6.10 (Sufficient condition for finite solution variance)**.** *Consider Problem 6.1 and assume that the functions* f_i *are* L*-smooth and are uniformly bounded from below functions: there exists* $r \in \mathbb{R}$ *such that* $f_i(x) \geq r$ *for every* $i \in \mathcal{I}$ *and* $x \in \mathcal{H}$*. Then Assumption 6.7 is true, with moreover the information that*

$$\sigma_*^2 \leq 2L(\min f - r).$$

Proof. Since each f_i is L-smooth, we can write (Garrigos and Gower, 2024, Lemma 2.28), for all $x \in \mathcal{H}$,

$$\|\nabla f_i(x)\|^2 \leq 2L(f_i(x) - \min f_i) \leq 2L(f_i(x) - r).$$

Setting $x = x_* \in \operatorname{argmin} f$ and taking the expectation of this inequality, we obtain

$$\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] \leq 2L(f(x_*) - r) = 2L(\min f - r) < +\infty,$$

which concludes since $\sigma_*^2 = \mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right]$. □

6.3 Interpolation

We conclude by connecting the variance at the solution with interpolation. The following facts can be found in Garrigos and Gower (2024), which we have extended here from a finite-sum problem to our expectation-minimization Problem 6.1.

Definition 6.11. *Consider Problem 6.1. We say that interpolation holds for the family of functions $(f_i)_{i\in\mathcal{I}}$ if*

$$\bigcap_{i\in\mathcal{I}} \operatorname{argmin} f_i \neq \emptyset,$$

where the intersection is to be understood for $\mathcal{D}$-almost every $i \in \mathcal{I}$.

Lemma 6.12 (Consequences for interpolation)**.** *Consider Problem 6.1. If interpolation holds, then*

1. $\mathbb{E}[\inf f_i] = \inf f$.
2. $\cap_{i\in\mathcal{I}} \operatorname{argmin} f_i = \operatorname{argmin} f$.

Proof. Because interpolation holds, we may select $x_* \in \cap_{i\in\mathcal{I}} \operatorname{argmin} f_i$. Let us start by showing that $x_* \in \operatorname{argmin} f$. For this, consider any $x \in \mathcal{H}$ and write

$$f(x_*) = \mathbb{E}[f_i(x_*)] = \mathbb{E}[\inf f_i] \leq \mathbb{E}[f_i(x)] = f(x).$$

This proves that $x_* \in \operatorname{argmin} f$. Now let us prove the first part, by observing that

$$\inf f = f(x_*) = \mathbb{E}[f_i(x_*)] = \mathbb{E}[\inf f_i].$$

Secondly, we prove the second point, for which we are only left with the reverse inclusion. Consider some $x \in \operatorname{argmin} f$. Then

$$f(x) = \inf f = \mathbb{E}[\inf f_i] \implies \mathbb{E}[f_i(x) - \inf f_i] = 0.$$

Since $f_i(x) - \inf f_i \geq 0$ for all i, it must hold that $f_i(x) - \inf f_i = 0$ for almost every $i \in \mathcal{I}$, which proves the claim. □

Lemma 6.13 (Interpolation means $\sigma_*^2 = 0$)**.** *Consider Problem 6.1.*

1. *If interpolation holds, then $\sigma_*^2 = 0$.*
2. *The above becomes an equivalence if all the f_i are convex.*

Proof. If interpolation holds, then for every $x_* \in \operatorname{argmin} f$ we have that $x_* \in \operatorname{argmin} f_i$ for almost every $i \in \mathcal{I}$. For such $i \in \mathcal{I}$, the optimality condition gives $\nabla f_i(x_*) = 0$, which after taking the norm and expectation leads to $\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] = 0$.

Reciprocally, if we assume that $\sigma_*^2 = 0$, it means that there exists $x_* \in \operatorname{argmin} f$ such that $\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] = 0$. This is an expectation of nonnegative terms, such that $\nabla f_i(x_*) = 0$ for almost every $i \in \mathcal{I}$. Exploiting our convexity assumption, we deduce that x_* is a minimizer of f_i for almost every $i \in \mathcal{I}$, which shows that interpolation holds. □

7 Extensions of SGD

All the main results in our work can be applied to variants of (SGD), such as mini-batching or nonuniform sampling. This is because these variants can be seen as instances of (SGD) applied to different but equivalent problems. Our results can be applied as a black-box, and the results follow by computing the main constants of the equivalent problem, namely L, μ, σ_*^2. This strategy was described in Gower et al. (2019), which contains all the formulas needed to compute those constants. For the sake of completeness, we include some details for the two aforementioned instances below.

7.1 Non-Uniform Sampling

Consider a finite family of functions $f_1, \dots, f_n$ which we assume to be μ_i-strongly convex and L_i-smooth, for $\mu_i \geq 0$ and $L_i > 0$. We want to study the non-uniform SGD algorithm, which computes

$$x_{t+1} = x_t - \frac{\gamma}{np_{i_t}} \nabla f_{i_t}(x_t), \tag{SGD$_p$}$$

where $i_t \in \{1, \dots, n\}$ is sampled according to $\mathbb{P}(i_t = i) = p_i$, with $p_i > 0$ and $\sum_{i=1}^n p_i = 1$. Note that we are using here a specific renormalization of the step-size, following the ideas from Gower et al. (2019). It is immediate to see that this algorithm is exactly (SGD) applied to the problem

$$\min_x \; f(x) = \mathbb{E}_{\mathcal{P}}\left[\hat{f}_i(x)\right], \text{ where } \hat{f}_i(x) \coloneqq \frac{1}{np_i} f_i(x),$$

where the expectation is taken with respect $\mathcal{P}$, the distribution whose density is given by the probabilities p_i. Note that the function f remains unchanged, meaning that the set of minimizers is the same. We can then compute (Gower et al., 2019, Propositions 3.7 & 3.8)) the following quantities:

- L, which is the maximum among the constants $\mathrm{Lip}(\nabla \hat{f}_i)$, so that we can say that each $\hat{f}_i$ is L-smooth. Clearly,
$$L = \max_{i=1,\dots,n} \frac{L_i}{np_i}.$$
- μ, which, for the same reasons as above, can be computed as
$$\mu = \min_{i=1,\dots,n} \frac{\mu_i}{np_i}.$$
- σ_*^2, which is given by
$$\mathbb{E}_{\mathcal{P}}\left[\|\nabla f_i(x_*)\|^2\right] = \sum_{i=1}^n p_i \|\nabla \hat{f}_i(x_*)\|^2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{np_i} \|\nabla \hat{f}_i(x_*)\|^2.$$

If we now impose that $\gamma L \in (0, 2)$, we can apply all our results and obtain bounds on the iterates of (SGD$_p$) depending on the above constants.

7.2 Mini-Batching

Consider a finite family of functions $f_1, \dots, f_n$ which we assume to be μ_i-strongly convex and L_i-smooth, for $\mu_i \geq 0$ and $L_i > 0$. We want to study the mini-batch SGD algorithm, which computes at each iteration a mini-batch of fixed size $1 \leq b \leq n$, that is

$$x_{t+1} = x_t - \frac{\gamma}{b} \sum_{i \in B_t} \nabla f_i(x_t), \tag{SGD$_b$}$$

where B_t is sampled i.i.d. and uniformly among the subsets of size b of $\{1, \dots, n\}$. It is a simple exercise to see that this algorithm is exactly (SGD) applied to the problem

$$\min_x \; f(x) = \mathbb{E}_{\mathcal{B}}\Big[\hat{f}_B(x)\Big], \text{ where } \hat{f}_B(x) \coloneqq \frac{1}{b} \sum_{i \in B} \nabla f_i(x),$$

and where the expectation is taken with respect to $\mathcal{B}$ which is the uniform law over

$$\text{batch}_b \coloneqq \{B \subset \{1, \dots, n\} \; : \; |B| = b\}.$$

To be able to apply our results, we need to verify that this new problem verifies our assumptions. More precisely, all we need is for it to verify (SEC$_*$) for some L and some μ. This can be calculated using the results in Gower et al. (2019).

Lemma 7.1. *Consider Problem 6.1 with $\mathcal{I}$ finite. Assume that each function f_i is μ_i-convex and L_i-smooth. Then the family of functions $(f_B)_{B \in batch_b}$ verifies (SEC$_*$) with constants*

- $L = (1 - r)\,\text{Lip}(\nabla f) + r \max_i L_i$, *with* $r = \frac{n(b-1)}{b(n-1)} \in [0, 1]$,
- $\mu = \min_i \mu_i$.

Moreover, we have that

$$\sigma_*^2 = \mathbb{E}_{\mathcal{B}}\big[\|\nabla f_B(x_*)\|^2\big] = \frac{n - b}{nb(n-1)} \sum_{i=1}^n \|\nabla f_i(x_*)\|^2.$$

Proof. Let $\mu \coloneqq \min_i \mu_i$, such that each f_i is μ-convex. Similarly, because each $\hat{f}_B$ is an average of f_i's, we know that each $\hat{f}_B$ is μ-convex. We can therefore define

$$g = f - \tfrac{\mu}{2}\|\cdot\|^2, \quad g_i = f_i - \tfrac{\mu}{2}\|\cdot\|^2 \quad \text{and} \quad \hat{g}_B = \hat{f}_B - \tfrac{\mu}{2}\|\cdot\|^2$$

all of which are convex functions. By definition, we know that each $\hat{g}_B$ is an average of functions g_i which are each $(L_i - \mu)$-smooth and convex. Invoking (Gower et al., 2019, Section G) shows that the family $(\hat{g}_B)_{B \in \text{batch}_b}$ verifies (EC) for constants $(\mathcal{L}, 0)$ where

$$\mathcal{L} = (1 - r)\,\text{Lip}(\nabla g) + r \max_i \text{Lip}(\nabla g_i), \text{ with } r = \tfrac{n(b-1)}{b(n-1)} \in [0, 1].$$

Using Lemma 7.2 which can be found below, we see that this is equivalent to the family $(\hat{f}_B)_{B \in \text{batch}_b}$ verifying (EC) for constants $(\mathcal{L} + \mu, \mu)$. Because $\text{Lip}(\nabla g) = \text{Lip}(\nabla f) - \mu$ and $\text{Lip}(\nabla g_i) = \text{Lip}(\nabla f_i) - \mu$, we conclude by setting $L = \mathcal{L} + \mu$. □

Lemma 7.2. *Consider Problem 6.1. Let $L \geq \mu \geq 0$, and let $(x, y) \in \mathcal{H}$ be fixed. Then the following are equivalent:*

1. *the family of functions $(f_i)_{i\in\mathcal{I}}$ verify (EC) with constants (L, μ);*
2. *the family of functions $(g_i)_{i\in\mathcal{I}}$ verify (EC) with constants $(L-\mu, 0)$, with $g_i = f_i - \frac{\mu}{2}\|\cdot\|^2$.*

Proof. It suffices to show that, if $\hat{f}, \hat{g}\colon \mathbb{R}^d \to \mathbb{R}$ are such that $\hat{g} = \hat{f} - \frac{\mu}{2}\|\cdot\|^2$, then

$$\begin{aligned}
&\tfrac{1}{2}\|\nabla\hat{f}(y) - \nabla\hat{f}(x)\|^2 + \tfrac{\mu L}{2}\|y-x\|^2 - (L-\mu)\big[\hat{f}(y) - \hat{f}(x)\big] - \langle \mu\nabla\hat{f}(y) - L\nabla\hat{f}(x), y - x\rangle \\
&\qquad = \tfrac{1}{2}\|\nabla\hat{g}(y) - \nabla\hat{g}(x)\|^2 - (L-\mu)\big[\hat{g}(y) - \hat{g}(x)\big] + (L-\mu)\langle\nabla\hat{g}(x), y - x\rangle. \qquad (4)
\end{aligned}$$

We start from the right-hand side, substituting $\hat{g}(z) = \hat{f}(z) - \frac{\mu}{2}\|z\|^2$ and $\nabla\hat{g}(z) = \nabla\hat{f}(z) - \mu z$ for $z = x$ and $z = y$, to obtain

$$\begin{aligned}
&\tfrac{1}{2}\|\nabla\hat{g}(y) - \nabla\hat{g}(x)\|^2 - (L-\mu)\big[\hat{g}(y) - \hat{g}(x)\big] + (L-\mu)\langle\nabla\hat{g}(x), y - x\rangle \\
&\quad = \tfrac{1}{2}\|\nabla\hat{f}(y) - \nabla\hat{f}(x)\|^2 - \mu\langle\nabla\hat{f}(y) - \nabla\hat{f}(x), y - x\rangle + \tfrac{\mu^2}{2}\|y-x\|^2 \\
&\qquad - (L-\mu)\big[\hat{f}(y) - \hat{f}(x)\big] + \tfrac{\mu}{2}(L-\mu)\big[\|y\|^2 - \|x\|^2\big] \\
&\qquad + (L-\mu)\langle\nabla\hat{f}(x), y - x\rangle - \mu(L-\mu)\langle x, y - x\rangle \\
&\quad = \tfrac{1}{2}\|\nabla\hat{f}(y) - \nabla\hat{f}(x)\|^2 - (L-\mu)\big[\hat{f}(y) - \hat{f}(x)\big] - \langle\mu\nabla\hat{f}(y) - L\nabla\hat{f}(x), y - x\rangle \\
&\qquad + \tfrac{\mu}{2}\left(\mu\|y-x\|^2 + (L-\mu)\big[\|y\|^2 - \|x\|^2\big] - 2(L-\mu)\langle x, y - x\rangle\right) \\
&\quad = \tfrac{1}{2}\|\nabla\hat{f}(y) - \nabla\hat{f}(x)\|^2 - (L-\mu)\big[\hat{f}(y) - \hat{f}(x)\big] - \langle\mu\nabla\hat{f}(y) - L\nabla\hat{f}(x), y - x\rangle \\
&\qquad + \tfrac{\mu^2}{2}\big[\|y-x\|^2 - \|y\|^2 - \|x\|^2 + 2\langle x, y\rangle\big] + \tfrac{\mu L}{2}\big[\|y\|^2 + \|x\|^2 - 2\langle x, y\rangle\big].
\end{aligned}$$

This is precisely the left-hand side of (4). □

8 Stochastic Proximal Algorithm

An important consequence of being able to study SGD for the optimal step-size $\gamma L = 1$ is that we are also able to study the Stochastic Proximal algorithm without much additional effort. In this section, we explain exactly why that is, and how our results compare to the currently known results on the Stochastic Proximal algorithm.

8.1 Problem Formulation

We consider an extension of Problem 6.1, where we drop the assumption that the loss functions f_i are differentiable or take finite values. We do however impose that the functions are convex.

Problem 8.1. *Let $\mathcal{H}$ be a real Hilbert space with associated inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Let $\{f_i\}_{i \in \mathcal{I}}$ be a family of extended real-valued functions $f_i \colon \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, where $\mathcal{I}$ is a (possibly infinite) set of indices. We are considering the problem of minimizing $f \coloneqq \mathbb{E}[f_i]$, where the expectation is taken over the indices $i \in \mathcal{I}$, with respect to some probability distribution $\mathcal{D}$ over $\mathcal{I}$. We assume that*

1. *the problem is well-defined, in the sense that $i \mapsto f_i(x)$ is $\mathcal{D}$-measurable;*
2. *the problem is well-posed, in the sense that* $\operatorname{argmin} f \neq \emptyset$*;*
3. *the problem is convex, in the sense that each f_i is convex, lower semi-continuous and proper.*

An algorithm of choice to solve Problem 8.1 is the *Stochastic Proximal algorithm*. It is defined as

$$x_{t+1} = \operatorname{prox}_{\gamma f_{i_t}}(x_t), \tag{SProx}$$

where $i_t \in \mathcal{I}$ is sampled i.i.d. from the distribution $\mathcal{D}$, and where $\operatorname{prox}_{\gamma f_i}$ is the proximal operator of f_i with step-size γ, defined by

$$\operatorname{prox}_{\gamma f_i}(x) = \operatorname*{argmin}_{y \in \mathcal{H}} f_i(y) + \frac{1}{2\gamma}\|y - x\|^2.$$

Our assumption that f_i is convex and lower semi-continuous ensures that $\operatorname{prox}_{\gamma f_i}$ is well-defined.

Literature review. Although the proximal algorithm is widely studied and well understood in the deterministic case, its complexity seems to be poorly understood in the stochastic setting. To the best of our knowledge, the only complexity results available for (SProx) can be divided in a few distinct categories:

1. Complexity results for the stochastic *projection* algorithm, where each function f_i is the *indicator* function of some nonempty closed convex set C_i. In this case, the problem becomes a feasibility problem, which consists in finding $x_* \in C \coloneqq \cap_i C_i$. Here the proximal operator of f_i becomes the projection onto C_i, and (SProx) performs random projections onto the sets C_i. Results typically show that

$$\mathbb{E}\big[\operatorname{dist}(\bar{x}_T; C_i)^2\big] \leq \frac{\operatorname{dist}(x_0; C)^2}{T},$$

see for instance (Nedić, 2010, Proposition 6). Other authors might additionally assume that the intersection between the sets C_i is regular to obtain better bounds, which is standard when studying deterministic feasibility problems Lewis et al. (2009). For instance, Necoara et al. (2019) considers the well-known linearly regular intersection assumption

$$(\exists \kappa > 0)(\forall x \in \mathcal{H}) \quad \operatorname{dist}(x, C)^2 \leq \kappa \mathbb{E}\big[\operatorname{dist}(x; C_i)^2\big],$$

and is able to exploit it to derive a stronger linear rate.

2. Complexity results under the assumption that each f_i takes *finite* values, and is G-Lipschitz continuous. This setting was explored by Bertsekas (2011), and more recently by Patrascu and Necoara (2018), Davis and Drusvyatskiy (2019) and Asi and Duchi (2019). Under this assumption, the (sub)gradients are G-bounded, and it is possible to show that

$$f(\bar{x}_T) - \min f \leq \frac{\|x_0 - x_*\|^2}{2\gamma T} + \frac{\gamma G^2}{2}. \tag{5}$$

3. Complexity results under the assumption that each f_i takes finite values and is L-*smooth*. The following bound was obtained in (Traoré et al., 2024, Theorem 4.3)

$$\mathbb{E}[f(\bar{x}_T) - \min f] \leq \frac{\|x_0 - x_*\|^2}{\gamma T} + 2\gamma\sigma_*^2.$$

Similar results can also be obtained by replacing the smoothness assumption with some weaker version, such as the (L_0, L_1)-smoothness assumption (Tovmasyan et al., 2025).

4. Complexity results under the assumption that the problem is μ-*strongly convex* with $\mu > 0$. This setting was explored in (Asi and Duchi, 2019, Proposition 5) and in (Richtárik et al., 2024, Theorem 5.3). The latter was able to obtain the following complexity bound, with the additional assumption that the functions are differentiable:

$$\mathbb{E}\big[\|x_T - x_*\|^2\big] \leq \left(\frac{1}{1+\gamma\mu}\right)^{2T} \|x_0 - x_*\|^2 + \frac{\gamma\sigma_*^2}{\mu(2+\gamma\mu)}. \tag{6}$$

We stress the fact that the gap between the settings considered in the literature is huge. Indeed, we have on the one side functions which are intrinsically taking $+\infty$ values, and on the other side functions taking finite values, with various levels of imposed regularity. As far as we know, only the results in the strongly convex case avoid making any regularity assumption on the functions f_i. In this case however we are discarding the feasibility problem, as it is not reasonable to assume that indicator functions are strongly convex, unless they are singletons, which would trivialize the problem. It is therefore clear that no unified analysis for (SProx) is available.

Studying the complexity of (SProx) is not trivial. Before trying to obtain bounds on (SProx), one must determine what metric to estimate.

In the deterministic setting, a standard result shows that the function value gap $f(x_t) - \min f$ is upper bounded by $O(\frac{1}{t})$. For SGD, it is standard as well to upper bound the expected averaged function value gap $\mathbb{E}\left[f(\bar{x}_T) - \min f\right]$. The problem with the stochastic proximal algorithm is that

we *cannot* hope to ever bound a function value gap. Indeed, nothing guarantees that the iterates x_t (or their average) remain in the domain of f. We of course know that $x_{t+1} \in \operatorname{dom}(f_{i_t})$, but there is no reason for it to belong in the intersection of all domains. This is even clearer if one considers the stochastic projection algorithm, where f is the indicator function of the intersection $\cap_i C_i$. For $f(x_t)$ to be finite, we need x_t to already be a solution. As such, we cannot expect bounds based on the function value gap.

One could think about providing a bound for the distance between the current iterate and a solution of the problem (or the distance to the set of solutions). However, this must also be discarded, as it is known in the deterministic case that such distance can decrease arbitrarily slowly (Garrigos et al., 2023). This is why, for convex problems, we usually do not have bounds on terms $\|x_t - x_*\|$, contrary to strongly convex problems.

In this section, we shall provide bounds for a different function gap, which is always well-defined, and which in some cases can be connected to well-known quantities, allowing to recover existing results.

8.2 Stochastic Proximal Algorithm as SGD

The proximal algorithm is a particular case of the gradient descent algorithm. This can be stated formally, by introducing the notion of Moreau regularization for a convex function. Given a proper convex lower semi-continuous function $f\colon \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, given $\gamma > 0$, we define its γ-Moreau regularization as the function f^γ by

$$f^\gamma(x) = \min_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma}\|y - x\|^2.$$

The Moreau regularization enjoys a few properties (Bauschke and Combettes, 2017, Proposition 12.30), namely

1. $f^\gamma\colon \mathcal{H} \to \mathbb{R}$ takes finite values, and is differentiable;
2. ∇f^γ is $\frac{1}{\gamma}$-Lipschitz continuous;
3. $\nabla f^\gamma(x) = \frac{x - \operatorname{prox}_{\gamma f}(x)}{\gamma}$.

In particular, we see that computing $\operatorname{prox}_{\gamma f}(x)$ is the same as computing one step of gradient descent for the function f^γ, satisfying $\operatorname{Lip}(\nabla f^\gamma) = \frac{1}{\gamma}$, with step-size γ:

$$\operatorname{prox}_{\gamma f}(x) = x - \gamma \nabla f^\gamma(x).$$

Therefore, it is clear that the stochastic proximal algorithm can be seen as an instance of SGD, whose step-size is exactly the optimal step-size:

$$x_{t+1} = \operatorname{prox}_{\gamma f_i}(x_t) = x_t - \gamma \nabla f_i^\gamma(x_t).$$

This connection between (SProx) and (SGD$_{\mathbb{E}}$) is not new, and was already discussed and exploited in Patrascu and Necoara (2018), Necoara et al. (2019). The novelty in our work is that we are able to exploit easily this connection, since we were able to derive complexity rates for SGD when $\gamma L = 1$.

We can see that the stochastic proximal algorithm aims at minimizing another function, which is given by

$$F^\gamma(x) \coloneqq \mathbb{E}[f_i^\gamma(x)].$$

The following result can immediately be derived from our bounds for SGD with optimal step-size in Theorem 2.4.

Theorem 8.2 (Bound for SProx - General case)**.** *Consider Problem 8.1. Let x_t be generated by (SProx) with any step-size $\gamma > 0$. Then, for all $T \geq 1$,*

$$\mathbb{E}\left[F^\gamma(\bar{x}_T) - \min F^\gamma\right] \leq \frac{\|x_0 - x_*^\gamma\|^2}{(2-\varepsilon)\gamma T} + \frac{(2+\varepsilon)\gamma\sigma_*^2(\gamma)}{\varepsilon(2-\varepsilon)},$$

where $\sigma_*^2(\gamma) = \mathbb{E}\left[\|\nabla f_i^\gamma(x_*^\gamma)\|^2\right]$, and $x_*^\gamma \in \operatorname{argmin} F^\gamma$. If we further assume that each f_i is μ-strongly convex ($\mu > 0$), then with $\mu_\gamma \coloneqq \frac{\mu}{1+\gamma\mu}$, we have

$$\mathbb{E}\left[\|x_T - x_*^\gamma\|^2\right] \leq \left(1 - \frac{\mu_\gamma}{L}\right)^{2T} \|x_0 - x_*^\gamma\|^2 + \frac{2}{\mu_\gamma L(2 - \frac{\mu_\gamma}{L})}\sigma_*^2(\gamma).$$

Proof. Apply Theorem 2.4 and Corollary 3.4 with $\gamma L = 1$. For the strongly convex case, we use the fact that if the f_i are μ-strongly convex, then f_i^* is $\frac{1}{\mu}$-smooth. This in turn implies that (Bauschke and Combettes, 2017, Proposition 13.24.iii)

$$(f_i^\gamma)^* = f^* + \frac{\gamma}{2}\|\cdot\|^2$$

is $\frac{1}{\mu} + \gamma$ smooth, allowing to conclude that f_i^γ is strongly convex with constant $\frac{1}{1/\mu+\gamma} = \mu_\gamma$. □

The previous theorem might not appear to be satisfying in the convex case, as it yields a bound on the regularized function value gap which is not clearly related to our original problem. We will see that in some cases, however, it is. The rest of this section is dedicated to investigating those cases. We will see that in the case of interpolation, and for feasibility problems, we recover the standard bounds from literature. We will moreover investigate the Lipschitz and smooth cases, where we can recover meaningful bounds, even though they are less good than the existing literature.

It remains to discuss the strongly convex case (with no further assumptions). Our bound reduces to

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \left(1 - \frac{\mu}{L}\frac{1}{1+\gamma\mu}\right)^{2T} \|x_0 - x_*\|^2 + O\left(\frac{1}{\mu}\right),$$

which is comparable with the bound $\exp(-\mu T) + O\left(\mu^{-1}\right)$ obtained in (Asi and Duchi, 2019, Proposition 5). Nevertheless, our result is clearly outperformed by (Richtárik et al., 2024, Theorem 6.4) which obtains the cleaner bound (6), even if we take into account the fact that it requires differentiability of the functions f_i.

Before continuing, we stress the fact that, while we focus on the vanilla Stochastic Proximal algorithm, all our results remain valid for standard extensions such as non-uniform sampling or mini-batching. It suffices to combine our results with the techniques from Section 7.

8.3 Recovering Previous Results

Theorem 8.2 provides a bound for (SProx) for any step-size. We now compare this theorem to the previously stated results, and observe that we obtain comparable results in a number of cases.

8.3.1 Case of Interpolation

Recall from Definition 6.11 that interpolation corresponds to the case where almost every function f_i share a common minimizer. Note that in Section 6.3 we considered differentiable functions, but the results of Lemma 6.12 trivially extend to the setting of Problem 8.1.

Proposition 8.3. *Consider Problem 8.1, and assume that interpolation holds for the family $(f_i)_{i\in\mathcal{I}}$. Then*

1. *Interpolation holds for the family $(f_i^\gamma)_{i\in\mathcal{I}}$.*
2. $\min F^\gamma = \min f$.
3. $\operatorname{argmin}\ F^\gamma = \operatorname{argmin}\ f$.

Proof. It is well-known that $\min f_i^\gamma = \min f_i$ and $\operatorname{argmin}\ f_i^\gamma = \operatorname{argmin}\ f_i$, see e.g. (Bauschke and Combettes, 2017, Proposition 12.9). Our assumption that interpolation holds for the family $(f_i)_{i\in\mathcal{I}}$ means that $\cap_i \operatorname{argmin}\ f_i \neq \emptyset$, from which we deduce that $\cap_i \operatorname{argmin}\ f_i^\gamma \neq \emptyset$, proving the first point. Then we can use Lemma 6.12 to write that

$$\operatorname{argmin}\ F^\gamma = \bigcap_{i\in\mathcal{I}} \operatorname{argmin}\ f_i^\gamma = \bigcap_{i\in\mathcal{I}} \operatorname{argmin}\ f_i = \operatorname{argmin}\ f,$$

and

$$\min F^\gamma = \mathbb{E}[\inf f_i^\gamma] = \mathbb{E}[\inf f_i] = \min f,$$

concluding the proof. □

This result shows that the function value gap $F^\gamma(x) - \min F^\gamma$ is a meaningful metric, in the sense that it provides a faithful and continuous measure of how far we are from optimality. This is particularly clear in the case of stochastic projections, which we will consider next. But first, we specialize our bound from Theorem 8.2 to this interpolation setting.

Theorem 8.4 (Bounds for SProx - Interpolation case)**.** *Consider Problem 8.1 and assume that interpolation holds. Let x_t be generated by (SProx) with step-size $\gamma > 0$. Then for every $x_* \in \operatorname{argmin}\ f$,*

$$\mathbb{E}\left[F^\gamma(\bar{x}_T) - \min f\right] \leq \frac{\|x_0 - x_*\|^2}{2\gamma T}.$$

If we further assume that each f_i is μ-strongly convex ($\mu > 0$), then with $\mu_\gamma \coloneqq \frac{\mu}{1+\gamma\mu}$ we have

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \left(1 - \frac{\mu_\gamma}{L}\right)^{2T} \|x_0 - x_*\|^2.$$

Proof. We apply Theorem 8.2 taking into account that $\operatorname{argmin}\ F^\gamma = \operatorname{argmin}\ f$ and $\sigma_*^2(\gamma) = 0$ because interpolation holds for the regularized problem as well, as seen in Proposition 8.3. In this case, we take $\varepsilon \to 0$ to obtain the result in the convex case. □

The above result is new in the convex case. We are not aware of any result providing bounds for (SProx) under the sole assumption of interpolation.

The result in the strongly convex case can be compared to (Tovmasyan et al., 2025, Theorem 6.1), where the authors obtain a worse rate, under the additional restrictions that $\gamma \leq \frac{\mu}{2}$, that the functions are differentiable, and that a bound on the gradient variance called *star-similarity* is verified. Note nevertheless that they only need the strong convexity of the overall function f, and are able to obtain a result without interpolation in (Tovmasyan et al., 2025, Theorem 6.4), which we are not.

8.3.2 Case of Indicator Functions

Here, we assume that $f_i = \delta_{C_i}$, the indicator function of some nonempty closed convex set $C_i \subset \mathcal{H}$. We recall that $\delta_{C_i}(x)$ is 0 if $x \in C_i$ and is $+\infty$ if $x \notin C_i$. In this case, (SProx) specializes into the Stochastic Projection algorithm:

$$x_{t+1} = \operatorname{proj}_{C_{i_t}}(x_t). \tag{SProj}$$

A first observation to make in this version of Problem 8.1 is that, for the problem to be well-posed, that is for argmin f to be nonempty, we need $C \coloneqq \cap_i C_i$ to be nonempty. But $C_i = \operatorname{argmin} f_i$, so this means that interpolation hold for well-posed feasibility problems. Note that one could also investigate (SProj) when the problem is not feasible, as done by Nedić (2010), but this goes beyond the scope of this work.

We may compute $f^\gamma(x) = \frac{1}{2\gamma}\operatorname{dist}(x; C_i)^2$, and therefore the regularized function value gap becomes

$$F^\gamma(x) - \min F^\gamma = \frac{1}{2\gamma}\mathbb{E}\big[\operatorname{dist}(x; C_i)^2\big], \tag{7}$$

which clearly is a good measure of how far x is from the set of solutions $C = \cap_i C_i$. We can now directly obtain a bound for (SProj), which is exactly the one obtained in (Nedić, 2010, Proposition 6).

Corollary 8.5 (Bounds for SProj)**.** *Consider Problem 8.1 where $f_i = \delta_{C_i}$, and let $C = \cap_i C_i$. Let x_t be generated by (SProj). Then,*

$$\mathbb{E}\big[\operatorname{dist}(\bar{x}_T; C_i)^2\big] \leq \frac{\operatorname{dist}(x_0; C)^2}{T}.$$

Proof. Apply Theorem 8.4 with (7). Moreover, we take x_* to be the projection of x_0 onto C. □

8.3.3 The case of functions with finite values

If we assume that the functions f_i take finite value, the previously mentioned issue of the function value gap $f(x) - \min f$ not being a good metric is no longer applicable. The next proposition shows that the former can be controlled by the regularized function gap $F^\gamma(x) - \min F^\gamma$.

In what follows, we will denote by $\partial f(x)$ the subdifferential of a function f at a point x:

$$\partial f(x) = \{g \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ f(y) - f(x) - \langle g, y - x\rangle \geq 0\}.$$

We will further note $\partial^0 f(x)$ to denote the element of minimal norm within $\partial f(x)$. The latter is well-defined because $\partial f(x)$ is nonempty (Bauschke and Combettes, 2017, Proposition 16.27), convex and closed (Bauschke and Combettes, 2017, Proposition 16.4).

Proposition 8.6. *Consider problem 8.1, and assume that each function f_i takes finite values. Then, for all $x \in \mathcal{H}$,*

$$f(x) - \min f \leq F^\gamma(x) - \min F^\gamma + \frac{\gamma}{2}\mathbb{E}\big[\|\partial^0 f_i(x)\|^2\big].$$

Proof. First, it is a standard result that $f_i^\gamma \leq f_i$. So we deduce easily that $F^\gamma \leq f$ and that $\min F^\gamma \leq \min f$. The rest of the proof is devoted to prove that $f(x) \leq F^\gamma(x) + \frac{\gamma}{2}\mathbb{E}\big[\|\partial^0 f_i(x)\|^2\big]$. We start by writing

$$f_i(x) - f_i^\gamma(x) \quad = \quad \sup_{y \in \mathcal{H}} f_i(x) - f_i(y) - \frac{1}{2\gamma}\|y - x\|^2.$$

As we assumed the functions f_i to be convex, we know that

$$f_i(x) - f_i(y) \leq -\langle \partial^0 f_i(x), y - x\rangle.$$

Combining all this, we obtain

$$\begin{aligned} f_i(x) - f_i^\gamma(x) \quad &\leq \quad \sup_{y \in \mathcal{H}} -\langle \partial^0 f_i(x), y - x\rangle - \frac{1}{2\gamma}\|y - x\|^2 \\ &= \quad \sup_{z \in \mathcal{H}} -\langle \partial^0 f_i(x), z\rangle - \frac{1}{2\gamma}\|z\|^2. \end{aligned}$$

Writing the optimality conditions, we deduce that the optimal z is $z = -\gamma\partial^0 f_i(x)$. Injecting this optimal solution leads to

$$f_i(x) - f_i^\gamma(x) \leq \frac{\gamma}{2}\|\partial^0 f_i(x)\|^2,$$

and the desired inequality follows after taking expectation. □

Corollary 8.7 (Bounds for SProx - Finite case)**.** *Consider Problem 8.1 and assume that each function f_i takes finite values. Let x_t be generated by (SProx), with step-size $\gamma > 0$. Then*

$$\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{\|x_0 - x_*^\gamma\|^2}{\gamma T} + 3\gamma\sigma_*^2(\gamma) + \frac{\gamma}{2}\mathbb{E}\big[\|\partial^0 f_i(\bar{x}_T)\|^2\big],$$

where $x_*^\gamma \in \operatorname{argmin} F^\gamma$ and $\sigma_*^2(\gamma) = \mathbb{E}\left[\|\nabla f_i^\gamma(x_*^\gamma)\|^2\right]$.

Proof. Combine Theorem 8.2 with Proposition 8.6, and take $\varepsilon = 1$ to simplify terms. □

This general bound, albeit totally new under such general assumptions, is not very useful. The main flaw is that it depends on $\mathbb{E}\big[\|\partial^0 f_i(\bar{x}_T)\|^2\big]$ which, without further assumptions, we cannot control. But under an additional assumption of Lipschitzness or smoothness, this is possible.

8.3.4 Case of Lipschitz Problems

We first translate the bound in Proposition 8.6 to the setting of Lipschitz functions.

Proposition 8.8. *Consider Problem 8.1, and assume that each function f_i is G-Lipschitz continuous. Then, for all $x \in \mathcal{H}$*

$$f(x) - \min f \leq F^\gamma(x) - \min F^\gamma + \frac{\gamma G^2}{2}.$$

Proof. This is a direct consequence of Proposition 8.6, where we further use the fact that the f_i are G-Lipschitz. This means that their subgradients are all bounded by G, and in particular that $\|\partial^0 f_i(x)\| \leq G$, which is what we need to prove the claim. □

This thus allows us to obtain bound for the stochastic proximal algorithm in the same setting.

Corollary 8.9 (Bounds for SProx - Lipschitz case)**.** *Consider Problem 8.1 and assume that each function f_i is G-Lipschitz continuous. Let x_t be generated by (SProx), with step-size $\gamma > 0$. Then*

$$\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{\|x_0 - x_*^\gamma\|^2}{\gamma T} + 4\gamma G^2,$$

where $x_*^\gamma \in \operatorname{argmin} F^\gamma$.

Proof. Combine Theorem 8.2 with Proposition 8.6, and take $\varepsilon = 1$ to simplify terms. Moreover, we use the fact that

$$\sigma_*^2(\gamma) = \mathbb{E}\big[\|\nabla f_i^\gamma(x_*^\gamma)\|^2\big] \leq \mathbb{E}\big[\|\nabla f_i(x_*^\gamma)\|^2\big] \leq G^2,$$

where we used the property that $\|\nabla f_i^\gamma(x)\| \leq \|\nabla f_i(x)\|$, a standard fact which can be, for instance, found in (Patrascu and Necoara, 2018, Lemma 3). □

We do not recover exactly the bound (5) from (Bertsekas, 2011, Theorem 5). In particular, our constants are worse by a factor 8. This should not come as a surprise, as we are relying on a result for optimal step-sizes, namely Theorem B.5, which has a singularity for $\gamma L = 1$ without uniform variance bound, and in a second time we assume that the functions have bounded gradients. It would be more efficient to exploit the bound on the gradients from the start, as done in (Bertsekas, 2011).

8.3.5 Case of Smooth Problems

We again translate the bound in Proposition 8.6 to the setting of smooth functions.

Proposition 8.10. *Consider Problem 8.1, and assume that each function f_i is L-smooth and bounded from below. If $\gamma < \frac{1}{L}$, then*

$$f(x) - \min f \leq \frac{1}{1 - \gamma L}\left(F^\gamma(x) - \min F^\gamma\right) + \frac{\gamma \Delta_*}{1 - \gamma L},$$

where $\Delta_* \coloneqq \min f - \mathbb{E}[\min f_i]$.

Proof. We start from Proposition 8.6. If the functions are L-smooth, a now standard variance transfer argument shows that (see (Garrigos and Gower, 2024, Lemma 4.19))

$$\mathbb{E}\left[\|\nabla f_i(x)\|^2\right] \leq 2L(f(x) - \min f) + 2\Delta_*, \quad \Delta_* := \min f - \mathbb{E}[\min f_i].$$

Using this bound gives us

$$f(x) - F^\gamma(x) \leq \gamma L(f(x) - \min f) + \gamma\Delta_*.$$

In the end, we obtained

$$(1 - \gamma L)(f(x) - \min f) \leq F^\gamma(x) - \min F^\gamma + \gamma\Delta_*.$$

The conclusion follows after dividing by $1 - \gamma L$. □

Note that the constant Δ_* appearing in our bound has the same status as σ_*^2, in the sense that it is an interpolation constant, see Garrigos and Gower (2024) for more details. In particular, $\Delta_* = 0$ if, and only if, interpolation holds, and, under a L-smoothness assumption, it holds that $\sigma_*^2 \leq 2L\Delta_*$.

Corollary 8.11 (Bounds for SProx - Smooth case)**.** *Consider Problem 8.1 and assume that each function f_i is L-smooth. Let x_t be generated by (SProx), with step-size $\gamma \in (0, \frac{1}{L})$. Then, for all $T \geq 1$,*

$$\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{\|x_0 - x_*^\gamma\|^2}{(1 - \gamma L)\gamma T} + \gamma\frac{6L + 1}{1 - \gamma L}\Delta_*,$$

where $x_*^\gamma \in \operatorname{argmin} F^\gamma$ and $\Delta_* = \min f - \mathbb{E}[\min f_i]$.

Proof. Combine Proposition 8.10 and Theorem 8.2, and take $\varepsilon = 1$ to simplify terms. To conclude, it remains to bound $\sigma_*^2(\gamma)$ with Δ_*. To do this, it is enough to take $x_i \in \operatorname{argmin} f_i^\gamma = \operatorname{argmin} f_i$ and to use the convexity and smoothness of f_i through Lemma 6.3 to get

$$\|\nabla f_i^\gamma(x_*^\gamma)\|^2 = \|\nabla f_i^\gamma(x_*^\gamma) - \nabla f_i^\gamma(x_i)\|^2 = 2L\left(f_i^\gamma(x_*^\gamma) - f_i^\gamma(x_i)\right) = 2L\left(f_i^\gamma(x_*^\gamma) - \min f_i\right).$$

After taking expectation

$$\sigma_*^2(\gamma) = 2L\left(\min F^\gamma - \mathbb{E}[\min f_i]\right) \leq 2L\left(\min f - \mathbb{E}[\min f_i]\right) = 2L\Delta_*,$$

which yields the wanted bound. □

Here we can see that our bounds are far from being as good as the ones in Traoré et al. (2024). Not only do we have a limitation on the step-size that our competitors do not have, but our constants are worse.

This clearly shows that, while our approach proved to be successful in the interpolation regime, there remains work to be done to properly tackle regularity assumptions on the functions.

9 Conclusion and Perspectives

In this thesis, we proved new and improved upper bounds for SGD without any variance assumption, both in the convex and strongly convex setting, for a full range of step-sizes. We moreover illustrated the sharpness of these results numerically using the Performance Estimation Problem methodology. Doing so, we raised questions regarding unexpected singularities appearing for optimal step-sizes. We moreover extended the results to the case of mini-batching and non-uniform adaptations of SGD, as well as to the study of the Stochastic Proximal Algorithm.

9.1 Future Work

We leave the following questions open and lines of research:

- Are the upper bounds obtained improvable through more general forms of Lyapunov energies? We worked with an energy of the form (Lyapunov), but do not claim this is the only sensible form.
- Is it possible to obtain upper bounds on (SGD) whose bias term match the best upper bound in the deterministic gradient descent algorithm? Both in the convex and strongly convex setting, we still observe a gap, and we conjecture this is not due to non-tightness of our analysis but rather a flaw in the core analysis technique.
- The analysis carried out solely focuses on constant step-sizes. Can this type of analysis be extended to cover varying step-sizes?
- In this work, we focused on achieving sharp bias terms. However, it is not the only possible approach, and we intend in a future work to investigate bounds providing sharp *complexity* rates.

9.2 Acknowledgments

I thank Lucas Ketels, Guillaume Garrigos and Juan Peypouquet, for their help and collaboration throughout this project.

This work benefited from the support of the FMJH Program Gaspard Monge for optimization and operations research and their interactions with data science. I also thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster.

A Reduction to a System of Inequalities

All proofs from Sections 2 and 3 are based on solving a system of inequalities that satisfy certain properties that allow us to conclude a decrease in the Lyapunov energy. In this section, we present how to reach bounds on SGD from a Lyapunov decrease (Subsection A.1), which was already partially presented in Sections 2 and 3, and how to reach a Lyapunov decrease from a system of inequalities (Subsection A.2).

A.1 Bounds from a Lyapunov Decrease

Let $T \geq 1$ be fixed, let $x_* \in \operatorname{argmin} f$, and let $(x_t)_{t=0}^T$ be generated by the SGD algorithm for some fixed step-size $\gamma > 0$. Given a set of parameters $\rho, a_0, \dots, a_T, e_0, \dots, e_{T-1} \geq 0$, we define the following Lyapunov energy, for $t = 0, \dots, T-1$:

$$E_t \coloneqq a_t \|x_t - x_*\|^2 + \rho \sum_{s=0}^{t-1} (f(x_s) - \min f) - \sum_{s=0}^{t-1} e_s \sigma_*^2,$$

where, by convention, the empty sum $\sum_{s=0}^{-1}$ is equal to zero.

The first term of this Lyapunov energy is the distance to the solution $\|x_t - x_*\|^2$, a classical term which typically decreases for deterministic monotone gradient dynamics. The second term involves the function gap $f(x_t) - \min f$, which also typically decreases for gradient descent. The standard Lyapunov for gradient descent usually contains the term $t(f(x_t) - \min f)$, which we have replaced with the sum of the past function gaps (observe that both are of the same order in time). The last term for this Lyapunov energy is a *negative* cumulated sum, where the e_t's play the role of a variance term. It is meant to compensate the fluctuations caused by the uncertainty in the SGD algorithm, and will allow the Lyapunov energy to decrease.

The core argument of our analysis is that a decrease of the Lyapunov (in expectation) yields upper bounds for SGD.

A first result, which is an immediate consequence of Lemma 2.2, consists in deriving bounds on the function value gap, provided the Lyapunov parameter ρ is nonzero. This bound will typically be obtained for convex smooth problems.

Lemma A.1 (SGD bound from Lyapunov decrease - Convex)**.** *Consider Problem 6.1. Assume that $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for every $t = 0, \dots, T-1$, and that $\rho > 0$. Then*

$$\min_{t=0,\dots,T-1} \mathbb{E}\left[f(x_t) - \min f\right] \leq \frac{a_0 \|x_0 - x_*\|^2}{\rho T} + \frac{\bar{e}\sigma_*^2}{\rho},$$

where $\bar{e} = \frac{1}{T}\sum_{t=0}^{T-1} e_t$. If we further assume that f is convex, then

$$\mathbb{E}\left[f(\bar{x}_T) - \min f\right] \leq \frac{a_0 \|x_0 - x_*\|^2}{\rho T} + \frac{\bar{e}\sigma_*^2}{\rho},$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

As a second result, a Lyapunov decrease can yield a bound of the square distance of the iterate x_t to a minimizer, provided the Lyapunov parameter a_T is nonzero. This is again an immediate consequence of Lemma 3.2, and is standard for strongly convex problems.

Lemma A.2 (SGD bound from Lyapunov decrease - Strongly convex)**.** *Consider Problem 6.1, and assume that* $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ *for every* $t = 0, \dots, T-1$*, and that* $a_T > 0$*. Then*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \frac{a_0 \|x_0 - x_*\|^2}{a_T} + \frac{e_T^{sum} \sigma_*^2}{a_T},$$

where $e_T^{sum} = \sum_{t=0}^{T-1} e_t$*.*

A.2 Lyapunov Decrease from a System of Inequalities

Which choices of parameters $\rho, a_0, \dots, a_T, e_0, \dots, e_{T-1} \geq 0$ allow the Lyapunov to decrease? The next theorem provides sufficient conditions which, as we will see later, are seemingly necessary, at least numerically. This is useful, as we will be able to focus on a system of inequalities between real numbers, instead of having to deal with all possible sequences generated by SGD. Finding parameters satisfying these sufficient conditions will be the object of the next sections.

Theorem A.3 (Sufficient conditions for Lyapunov decrease)**.** *Let Assumptions 6.4 and 6.7 hold true. Assume that there exist parameters* $(\alpha_t, \beta_t)_{t=0}^{T-1}$ *such that, for every* $t = 0, \dots, T-1$*, the following conditions are verified:*

1. $\rho, a_t, e_t, \alpha_t, \beta_t \geq 0$,
2. $\rho \leq 2(L - \mu)(\alpha_t - \beta_t)$,
3. $a_{t+1} \leq \mu L(\alpha_t + \beta_t) + a_t$,
4. $a_{t+1}\gamma^2 \leq \alpha_t + \beta_t$,
5. $(a_{t+1}\gamma - \alpha_t L - \beta_t \mu)^2 \leq (\mu L(\alpha_t + \beta_t) + a_t - a_{t+1})(\alpha_t + \beta_t - a_{t+1}\gamma^2)$,
6. $a_{t+1}\gamma^2(\alpha_t + \beta_t) \leq (\alpha_t + \beta_t - a_{t+1}\gamma^2)e_t$.

Then $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ *for every* $t = 0, \dots, T-1$*.*

Proof. We know we want to show that

$$E_{t+1} - E_t = a_{t+1}\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] - a_t \mathbb{E}\left[\|x_t - x_*\|^2\right] + \rho \mathbb{E}[f(x_t) - \min f] - e_t \sigma_*^2 \leq 0.$$

Using the definition of x_{t+1} allows us to write

$$\|x_{t+1} - x_*\|^2 = \|x_t - x_*\|^2 + \gamma^2 \|\nabla f_{i_t}(x_t)\|^2 - 2\gamma \langle \nabla f_{i_t}(x_t), x_t - x_* \rangle.$$

Therefore,

$$\begin{aligned} E_{t+1} - E_t \quad = \quad & \rho\left(f(x_t) - \min f\right) - e_t \sigma_*^2 + (a_{t+1} - a_t)\|x_t - x_*\|^2 \\ & + a_{t+1}\gamma^2 \|\nabla f_i(x_t)\|^2 - 2a_{t+1}\gamma \langle \nabla f_i(x_t), x_t - x_* \rangle. \end{aligned} \tag{8}$$

To show that this random variable is nonpositive in expectation, we shall make use of an assumption on the loss functions. While it would be natural to exploit the fact that the f_i are μ-convex and smooth for all $i \in \mathcal{I}$, our numerical simulations suggest that all we need is (SEC$_*$).

In what follows, we denote by $\mathbb{E}_t \coloneqq \mathbb{E}[\cdot \mid x_0, \dots, x_t]$ the conditional expectation with respect to the iterates up to and including x_t. Property (SEC$_*$) reads

$$\frac{1}{2}\mathbb{E}_t\left[\|\nabla f_i(x_*) - \nabla f_i(x_t)\|^2\right] + \frac{\mu L}{2}\|x_* - x_t\|^2 \leq (L-\mu)\left(\min f - f(x_t)\right) - L\langle \nabla f(x_t), x_* - x_t\rangle$$

at the pair (x_t, x_*), and

$$\frac{1}{2}\mathbb{E}_t\left[\|\nabla f_i(x_t) - \nabla f_i(x_*)\|^2\right] + \frac{\mu L}{2}\|x_t - x_*\|^2 \leq (L-\mu)\left(f(x_t) - \min f\right) + \mu\langle \nabla f(x_t), x_t - x_*\rangle$$

at the pair (x_*, x_t). Multiplying these inequalities by $2\alpha_t \geq 0$ and $2\beta_t \geq 0$ respectively, and summing them gives

$$\begin{aligned} & (\alpha_t + \beta_t)\mathbb{E}_t\left[\|\nabla f_i(x_t) - \nabla f_i(x_*)\|^2\right] + \mu L(\alpha_t + \beta_t)\|x_t - x_*\|^2 \\ \leq\ & 2(L-\mu)(\beta_t - \alpha_t)\left(f(x_t) - \min f\right) + 2(\alpha_t L + \beta_t \mu)\langle \nabla f(x_t), x_t - x_*\rangle. \end{aligned}$$

Adding $\mathbb{E}_t[E_{t+1}] - E_t$ on both sides and using the expression obtained in (8), we get

$$\begin{aligned} \mathbb{E}_t[E_{t+1}] - E_t \ \leq\ & \mathbb{E}_t[E_{t+1}] - E_t \\ & + 2(L-\mu)(\beta_t - \alpha_t)(f(x_t) - \min f) + 2(\alpha_t L + \beta_t \mu)\langle \nabla f(x_t), x_t - x_*\rangle \\ & - (\alpha_t + \beta_t)\mathbb{E}_t\|\nabla f_i(x_t) - \nabla f_i(x_*)\|^2 - \mu L(\alpha_t + \beta_t)\|x_t - x_*\|^2 \\ =\ & \rho\left(f(x_t) - \min f\right) - e_t\sigma_*^2 + (a_{t+1} - a_t)\|x_t - x_*\|^2 \\ & + a_{t+1}\gamma^2\|\nabla f_i(x_t)\|^2 - 2a_{t+1}\gamma\langle \nabla f_i(x_t), x_t - x_*\rangle \\ & + 2(L-\mu)(\beta_t - \alpha_t)(f(x_t) - \min f) + 2(\alpha_t L + \beta_t \mu)\langle \nabla f(x_t), x_t - x_*\rangle \\ & - (\alpha_t + \beta_t)\mathbb{E}_t\|\nabla f_i(x_t) - \nabla f_i(x_*)\|^2 - \mu L(\alpha_t + \beta_t)\|x_t - x_*\|^2 \\ =\ & (\rho + 2(L-\mu)(\beta_t - \alpha_t))(f(x_t) - \min f) - e_t\sigma_*^2 \\ & + (a_{t+1} - a_t - \mu L(\alpha_t + \beta_t))\|x_t - x_*\|^2 + a_{t+1}\gamma^2\|\nabla f_i(x_t)\|^2 \\ & + 2(\alpha_t L + \beta_t \mu - a_{t+1}\gamma)\langle \nabla f_i(x_t), x_t - x_*\rangle \\ & - (\alpha_t + \beta_t)\mathbb{E}_t\|\nabla f_i(x_t) - \nabla f_i(x_*)\|^2. \end{aligned}$$

Taking the expectation and developing the last square yields

$$\begin{aligned} \mathbb{E}[E_{t+1} - E_t] \ \leq\ & (\rho + 2(L-\mu)(\beta_t - \alpha_t))\mathbb{E}\left[f(x_t) - \min f\right] \\ & + (a_{t+1} - a_t - \mu L(\alpha_t + \beta_t))\mathbb{E}\left[\|x_t - x_*\|^2\right] \\ & + (a_{t+1}\gamma^2 - \alpha_t - \beta_t)\mathbb{E}\left[\|\nabla f_i(x_t)\|^2\right] \\ & - (e_t + \alpha_t + \beta_t)\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] \\ & + 2(\alpha_t L + \beta_t \mu - a_{t+1}\gamma)\mathbb{E}\left[\langle \nabla f_i(x_t), x_t - x_*\rangle\right] \\ & + 2(\alpha_t + \beta_t)\mathbb{E}\left[\langle \nabla f_i(x_t), \nabla f_i(x_*)\rangle\right]. \end{aligned}$$

For the Lyapunov energy to be decreasing in expectation, it is enough that the right-hand side is nonpositive. For the first term, because $f(x_t) - \min f \geq 0$, it is enough to assume that

$\rho \leq 2(L-\mu)(\alpha_t - \beta_t)$, which corresponds to Condition 2. For the other terms, we enforce that

$$\begin{aligned} & A\mathbb{E}\left[\|x_t - x_*\|^2\right] + B\mathbb{E}\left[\|\nabla f_i(x_t)\|^2\right] + B'\mathbb{E}\left[\|\nabla f_i(x_*)\|^2\right] \\ + \quad & 2C\mathbb{E}\left[\langle \nabla f_i(x_t), x_t - x_*\rangle\right] + 2D\mathbb{E}\left[\langle \nabla f_i(x_t), \nabla f_i(x_*)\rangle\right] \geq 0, \end{aligned} \tag{9}$$

where we simplified the expression by introducing the following constants

$$\begin{cases} A &= \mu L(\alpha_t + \beta_t) + a_t - a_{t+1}, \\ B &= \alpha_t + \beta_t - a_{t+1}\gamma^2, \\ B' &= e_t + \alpha_t + \beta_t, \\ C &= a_{t+1}\gamma - \alpha_t L - \beta_t \mu, \\ D &= -(\alpha_t + \beta_t). \end{cases}$$

The quantity in (9) is simply the expectation of a quadratic polynomial in $X = x_t - x_*$, $Y = \nabla f_i(x_t)$ and $Y' = \nabla f_i(x_*)$. Using elementary linear algebra, it can be shown that for the expression in (9) to be nonnegative, it is enough to require

$$A, B, B' \geq 0, \quad C^2 \leq AB, \quad \text{and} \quad D^2 \leq BB'.$$

We postpone the proof of this claim to Lemma A.5.

A simple calculation shows that these conditions correspond exactly to the remaining wanted. In fact, $A \geq 0$ corresponds to Condition 3 and $B \geq 0$ to Condition 4. The inequality $B' \geq 0$ is trivially satisfied because $e_t, \alpha_t, \beta_t \geq 0$ by Condition 1. Moreover, $C^2 \leq AB$ may be developed as

$$(a_{t+1}\gamma - \alpha_t L - \beta_t \mu)^2 \leq (\mu L(\alpha_t + \beta_t)a_t - a_{t+1})(\alpha_t + \beta_t - a_{t+1}\gamma^2),$$

which corresponds to Condition 5. Finally, $D^2 \leq BB'$ expands to

$$(\alpha_t + \beta_t)^2 \leq (\alpha_t + \beta_t - a_{t+1}\gamma^2)(e_t + \alpha_t + \beta_t),$$

which can be simplified to

$$a_{t+1}\gamma^2(\alpha_t + \beta_t) \leq (\alpha_t + \beta_t - a_{t+1}\gamma^2)e_t,$$

which is exactly Condition 6. □

Remark A.4 (Relaxing (SEC$_*$) when $\beta_t = 0$)**.** *Looking at how we used (SEC$_*$) in the above proof, it is clear that if we impose $\beta_t = 0$, then one of the two inequalities in (SEC$_*$) is useless, and that all we need in that case is (EC$_*$). This will be, for instance, the case when proving our bounds for short step-sizes, see later in Remark B.3.*

We end this section by proving the technical lemma at the core of the previous proof.

Lemma A.5 (Nonnegative quadratic polynomial)**.** *Let $A, B, B', C, D \in \mathbb{R}$ be such that*

$$A, B, B' \geq 0, \quad C^2 \leq AB, \quad \text{and} \quad D^2 \leq BB'. \tag{10}$$

Let X, Y, and Y' be three random variables over $\mathcal{H}$ such that $\mathbb{E}[Y'] = 0$. Then

$$\mathbb{E}\big[A\|X\|^2 + B\|Y\|^2 + B'\|Y'\|^2 + 2C\langle Y, X\rangle - 2D\langle Y, Y'\rangle\big] \geq 0.$$

Proof. Let us note P to be the quantity that we want to see nonnegative. We start by exploiting the fact that $\mathbb{E}[Y'] = 0$ to write that $P = P + \theta\mathbb{E}[\langle X, Y'\rangle]$, for every $\theta \in \mathbb{R}$. Now, see that for $P \geq 0$ to be true, it is enough for

$$\hat{P}_\theta(x, y, y') = A\|x\|^2 + B\|y\|^2 + B'\|y'\|^2 + 2C\langle y, x\rangle - 2D\langle y, y'\rangle + 2\theta\langle x, y'\rangle$$

to be nonnegative for every $x, y, y' \in \mathcal{H}$ and for some $\theta \in \mathbb{R}$. Our polynomial $\hat{P}_\theta$ is equal to $\langle M_\theta z, z\rangle$ where $z = (x, y, y')$ and

$$M_\theta = \begin{pmatrix} A & C & \theta \\ C & B & -D \\ \theta & -D & B' \end{pmatrix}.$$

As such, $\hat{P}_\theta \geq 0$ if, and only if, $M_\theta \succeq 0$, which, by Sylvester's criterion, is equivalent to

$$A, B, B' \geq 0, \tag{11a}$$

$$C^2 \leq AB, \quad \theta^2 \leq AB', \quad D^2 \leq BB', \tag{11b}$$

$$A(BB' - D^2) \geq B'C^2 + B\theta^2 + 2CD\theta. \tag{11c}$$

So $P \geq 0$ if there exists $\theta \in \mathbb{R}$ such that Inequalities (11) are verified. We can already see Equations (11a)-(11b) imply Equations (10). Let us prove that in fact the existence of a $\theta \in \mathbb{R}$ satisfying Equations (11) is equivalent to Equations (10).

To prove equivalence, we assume Equations (10) and show that there exists a $\theta \in \mathbb{R}$ such that

$$AB' \geq \theta^2 \quad \text{and} \quad A(BB' - D^2) \geq B'C^2 + B\theta^2 + 2CD\theta. \tag{12}$$

The last inequality is quadratic in θ, and has solutions if, and only if, its discriminant Δ is nonnegative. A straightforward computation shows that

$$\Delta = 4(BB' - D^2)(AB - C^2),$$

which guarantees that $\Delta \geq 0$ given Equations (10). The quadratic may hence be rewritten as

$$\frac{-2CD - \sqrt{\Delta}}{2B} \leq \theta \leq \frac{-2CD + \sqrt{\Delta}}{2B}.$$

For (12) to have a solution, the above must hold simultaneously with $-\sqrt{AB'} \leq \theta \leq \sqrt{AB'}$. This is equivalent to the intersection of two intervals to be nonempty, which is the case if

$$\begin{cases} \frac{-2CD+\sqrt{\Delta}}{2B} \geq -\sqrt{AB'} & \text{if } CD \geq 0, \\ \frac{-2CD-\sqrt{\Delta}}{2B} \leq \sqrt{AB'} & \text{if } CD \leq 0. \end{cases}$$

or equivalently,

$$2|CD| \leq 2B\sqrt{AB'} + \sqrt{\Delta}.$$

As all terms are nonnegative, this inequality is equivalent to

$$4C^2D^2 \leq 4B^2AB' + 4B\sqrt{AB'\Delta} + \Delta,$$

which holds true since $D^2 \leq BB'$, $C^2 \leq AB$ and $\Delta \geq 0$. □

B Proofs in the Smooth Convex Setting

In this section, we derive parameters that satisfy the sufficient conditions of Theorem A.3 in order to derive upper bounds for SGD in the convex setting. The propositions in this section can be verified symbolically, yet we include the proofs to provide intuition on how these parameters were derived.

B.1 Bounds for Short Step-Sizes

Proposition B.1 (Lyapunov parameters. Convex case, short step-sizes)**.** *Let $\gamma L \in (0, 1)$, and consider the parameters $\rho, a_t, e_t, \alpha_t, \beta_t$ defined by*

- $\rho = 2\gamma + \frac{2(1-\gamma L)}{LT}$,
- $a_t = \frac{T-t}{T}\frac{1+\gamma L(T-1)}{1+\gamma L(T-t-1)}$,
- $e_t = \frac{a_{t+1}\gamma^2\alpha_t}{\alpha_t - a_{t+1}\gamma^2}$,
- $\alpha_t \equiv \alpha = \frac{\rho}{2L}$,
- $\beta_t \equiv \beta = 0$.

These parameters satisfy the sufficient Lyapunov conditions of Theorem A.3. Moreover, it holds that

$$e_t = \frac{\gamma^2}{1-\gamma L}\frac{T-t-1}{T}\frac{(1-\gamma L)+\gamma LT}{(1-\gamma L)+\gamma L(T-t)} \leq \frac{\gamma^2}{1-\gamma L}.$$

Proof. Notice that the sufficient conditions from Theorem A.3 are homogeneous. Because our goal is to obtain bounds whose bias term is the smallest possible, we can without loss of generality impose that $a_0 = 1$, and try to maximize ρ. Through our numerical analysis, we empirically observe that we can take $\alpha_t \equiv \alpha > 0$ and $\beta_t \equiv \beta = 0$, thus justifying these choices. We dedicate the remainder of the proof to justify our remaining choices.

We start by focusing on Condition 5 from Theorem A.3, which reads

$$(a_{t+1}\gamma - L\alpha)^2 \leq (\alpha - a_{t+1}\gamma^2)(a_t - a_{t+1}).$$

Rearranging this inequality yields

$$\begin{aligned}
&\iff \quad (a_{t+1}\gamma - L\alpha)^2 \leq (1-\gamma L)\alpha(a_t - a_{t+1}) + (\gamma L\alpha - a_{t+1}\gamma^2)(a_t - a_{t+1}) \\
&\iff \quad (a_{t+1}\gamma - L\alpha)^2 - \gamma(L\alpha - a_{t+1}\gamma)(a_t - a_{t+1}) \leq (1-\gamma L)\alpha(a_t - a_{t+1}) \\
&\iff \quad (L\alpha - a_{t+1}\gamma)\left(L\alpha - a_{t+1}\gamma - \gamma(a_t - a_{t+1})\right) \leq (1-\gamma L)\alpha(a_t - a_{t+1}) \\
&\iff \quad (L\alpha - a_{t+1}\gamma)\left(L\alpha - a_t\gamma\right) \leq (1-\gamma L)\alpha(a_t - a_{t+1}).
\end{aligned}$$

Condition 3 boils down to $a_t \geq a_{t+1}$. By additionally imposing $a_t > a_{t+1}$, we may divide by $a_t - a_{t+1}$ to obtain

$$\begin{aligned}
&\iff \quad \frac{(L\alpha - a_{t+1}\gamma)\left(L\alpha - a_t\gamma\right)}{a_t - a_{t+1}} \leq (1-\gamma L)\alpha \\
&\iff \quad \frac{(L\alpha - a_{t+1}\gamma)\left(L\alpha - a_t\gamma\right)}{(L\alpha - a_{t+1}\gamma) - (L\alpha - a_t\gamma)} \leq \frac{(1-\gamma L)\alpha}{\gamma}.
\end{aligned}$$

As (a_t) is decreasing, with $a_0 = 1$, it holds true that $a_t \leq 1$ for all $t = 0, \dots, T-1$. Specifically, if $L\alpha > \gamma$ (which we shall assume), then both $L\alpha - a_{t+1}\gamma > 0$ and $L\alpha - a_t\gamma > 0$, thus allowing us to write

$$\begin{aligned} &\iff && \frac{(L\alpha - a_{t+1}\gamma) - (L\alpha - a_t\gamma)}{(L\alpha - a_{t+1}\gamma)(L\alpha - a_t\gamma)} \geq \frac{\gamma}{(1-\gamma L)\alpha} \\ &\iff && \frac{1}{L\alpha - a_t\gamma} - \frac{1}{L\alpha - a_{t+1}\gamma} \geq \frac{\gamma}{(1-\gamma L)\alpha} \\ &\iff && \frac{1}{L\alpha - a_{t+1}\gamma} \leq \frac{1}{L\alpha - a_t\gamma} - \frac{\gamma}{(1-\gamma L)\alpha}. \end{aligned}$$

Introducing the temporary variables $u_t \coloneqq \frac{1}{L\alpha - \gamma a_t}$ and $c = \frac{\gamma}{(1-\gamma L)\alpha}$, we see that we obtained an arithmetic inequality $u_{t+1} \leq u_t - c$, such that $u_t \leq u_0 - ct$. We know that $u_0 = \frac{1}{L\alpha - \gamma}$, so we can write

$$u_t \leq \frac{1}{L\alpha - \gamma} - \frac{t\gamma}{(1-\gamma L)\alpha} = \frac{(1-\gamma L)\alpha - t\gamma(L\alpha - \gamma)}{(1-\gamma L)\alpha(L\alpha - \gamma)}.$$

Returning to the definition of u_t we deduce the following bound for a_t:

$$\begin{aligned} a_t &\leq \frac{1}{\gamma}\left(L\alpha - \frac{(1-\gamma L)\alpha(L\alpha - \gamma)}{(1-\gamma L)\alpha - t\gamma(L\alpha - \gamma)}\right) \\ &= \frac{1}{\gamma}\frac{L\alpha^2(1-\gamma L) - L\alpha t\gamma(L\alpha - \gamma) - (1-\gamma L)\alpha(L\alpha - \gamma)}{(1-\gamma L)\alpha - t\gamma(L\alpha - \gamma)} \\ &= \frac{1}{\gamma}\frac{-L\alpha t\gamma(L\alpha - \gamma) + \alpha\gamma(1-\gamma L)}{(1-\gamma L)\alpha - t\gamma(L\alpha - \gamma)} \\ &= \alpha\frac{(1-\gamma L) - tL(L\alpha - \gamma)}{(1-\gamma L)\alpha - t\gamma(L\alpha - \gamma)}. \end{aligned}$$

We note that our Condition 1 requires that $a_t \geq 0$ for all $t = 0, \dots, T$, which is equivalent to having $a_T \geq 0$ as we assumed (a_t) to be decreasing. Specifically, we must have

$$(1-\gamma L) - TL(L\alpha - \gamma) \geq 0 \quad \iff \quad \alpha \leq \frac{1 - \gamma L + \gamma TL}{TL^2}.$$

As we want ρ as large as possible, and due to Condition 2 which reads $\rho \leq 2L(\alpha - \beta)$ with $\beta = 0$, we want α to be as large as possible. Therefore we fix

$$\alpha = \frac{1 - \gamma L + T\gamma L}{TL^2} \quad \text{and} \quad \rho = 2L\alpha = \frac{2(1-\gamma L) + 2T\gamma L}{TL}.$$

This choice respects the assumption $L\alpha > \gamma$ we made earlier, since $1 - \gamma L > 0$.

To get an expression for a_t, we simply set the inequalities to equalities, and replace α by its chosen value. We then obtain

$$L\alpha - \gamma = \frac{(1-\gamma L) + \gamma LT - \gamma LT}{TL} = \frac{1-\gamma L}{TL} \Rightarrow u_0 = \frac{TL}{1-\gamma L}.$$

So

$$\begin{aligned} u_t &= u_0 - tc = \frac{TL}{1-\gamma L} - \frac{t\gamma}{\alpha(1-\gamma L)} = \frac{\alpha TL - t\gamma}{\alpha(1-\gamma L)} \\ &= \frac{\frac{1}{L}(1-\gamma L) + \gamma T - t\gamma}{\alpha(1-\gamma L)} = \gamma\frac{T - t - 1 + \frac{1}{\gamma L}}{\alpha(1-\gamma L)}. \end{aligned}$$

We now use the fact that $\gamma a_t = L\alpha - u_t^{-1}$ to write

$$\begin{aligned}
a_t &= \frac{L\alpha}{\gamma} - \frac{1}{\gamma u_t} = \frac{\alpha L}{\gamma} - \frac{1}{L\gamma}\frac{(1-\gamma L)\frac{\alpha L}{\gamma}}{T-t-1+\frac{1}{\gamma L}} \\
&= \frac{\alpha L}{\gamma}\left(1 - \frac{1}{\gamma L}\frac{(1-\gamma L)}{T-t-1+\frac{1}{\gamma L}}\right) = \frac{\alpha L}{\gamma}\left(1 - \frac{(1-\gamma L)}{\gamma L(T-t-1)+1}\right) \\
&= \frac{\alpha L}{\gamma}\frac{\gamma L(T-t)}{\gamma L(T-t-1)+1} = \frac{(1-\gamma L)+\gamma LT}{T\gamma L}\frac{\gamma L(T-t)}{\gamma L(T-t-1)+1} \\
&= \frac{(T-t)}{T}\frac{1+\gamma L(T-1)}{\gamma L(T-t-1)+1}.
\end{aligned}$$

To conclude the proof, we need to check all the sufficient Lyapunov conditions of Theorem A.3. Conditions 1, 2, 3 and 5 are already satisfied, by construction. It remains to check Conditions 4 and 6.

Condition 4 is equivalent to

$$\begin{aligned}
a_{t+1}\gamma^2 \leq \alpha \quad &\iff \quad \gamma^2\frac{(T-t-1)}{T}\frac{1+\gamma L(T-1)}{1+\gamma L(T-t-2)} \leq \frac{(1-\gamma L)+\gamma LT}{TL^2} \\
&\iff \quad \frac{\gamma^2L^2(T-t-1)}{1+\gamma L(T-t-2)}(1+\gamma L(T-1)) \leq 1+\gamma L(T-1) \\
&\iff \quad \gamma^2L^2(T-t-1) \leq 1+\gamma L(T-t-2) = 1+\gamma L(T-t-1)-\gamma L \\
&\iff \quad 0 \leq 1-\gamma L+\gamma L(T-t-1)(1-\gamma L),
\end{aligned}$$

which is true for every $t = 0, \dots, T-1$.

Condition 6 is a lower bound on e_t, of which the equality case coincides with the chosen value of e_t. More specifically, this gives

$$e_t = \frac{a_{t+1}\gamma^2\alpha}{\alpha - a_{t+1}\gamma^2} = \frac{\gamma^2}{\frac{1}{a_{t+1}} - \frac{\gamma^2}{\alpha}},$$

where

$$\begin{aligned}
\frac{1}{a_{t+1}} - \frac{\gamma^2}{\alpha} &= \frac{T}{T-t-1}\frac{(1-\gamma L)+\gamma L(T-t-1)}{(1-\gamma L)+\gamma LT} - \frac{T\gamma^2L^2}{(1-\gamma L)+\gamma LT} \\
&= \frac{T}{T-t-1}\frac{(1-\gamma L)+\gamma L(T-t-1)}{(1-\gamma L)+\gamma LT} - \frac{T\gamma^2L^2}{(1-\gamma L)+\gamma LT} \\
&= T\frac{(1-\gamma L)+\gamma L(T-t-1)-\gamma^2L^2(T-t-1)}{(T-t-1)((1-\gamma L)+\gamma LT)} \\
&= T\frac{(1-\gamma L)((1-\gamma L)+\gamma L(T-t))}{(T-t-1)((1-\gamma L)+\gamma LT)}.
\end{aligned}$$

We thus get

$$e_t = \frac{\gamma^2}{1-\gamma L}\frac{T-t-1}{T}\frac{(1-\gamma L)+\gamma LT}{(1-\gamma L)+\gamma L(T-t)}.$$

We readily verify that e_t is nonincreasing with respect to t. Indeed, its derivative with respect to t has the same sign as

$$(T-1)((1-\gamma L)+\gamma LT)\gamma L - ((1-\gamma L)+\gamma LT)^2,$$

which is always nonpositive. Therefore $e_t \leq e_0$, where

$$e_0 = \frac{\gamma^2}{1-\gamma L}\frac{T-1}{T} \leq \frac{\gamma^2}{1-\gamma L},$$

thus concluding the proof. □

Theorem B.2 (Bounds for SGD. Convex case, short step-sizes)**.** *Let Assumptions 6.4 and 6.7 hold, with $\mu = 0$. Let x_t be generated by SGD, with $\gamma L \in (0,1)$. Then for every $T \geq 1$*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(x_t) - \inf f] \leq \frac{\|x_0 - x_*\|^2}{\rho T} + \frac{\gamma^2\sigma_*^2}{\rho(1-\gamma L)}\bar{e}'_T,$$

where $\rho = 2\gamma + \frac{2(1-\gamma L)}{LT}$ and

$$\bar{e}'_T = \frac{1}{T}\sum_{t=0}^{T-1} e'_t, \quad \text{with } e'_t = \frac{T-t-1}{T}\frac{(1-\gamma L)+\gamma LT}{(1-\gamma L)+\gamma L(T-t)}.$$

Moreover, we have the simpler bounds

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(x_t) - \inf f] \leq \frac{L\|x_0 - x_*\|^2}{2\gamma LT + 2(1-\gamma L)} + \frac{\gamma\sigma_*^2}{2(1-\gamma L)} \leq \frac{\|x_0 - x_*\|^2}{2\gamma T} + \frac{\gamma\sigma_*^2}{2(1-\gamma L)}.$$

If moreover f is convex, the above bounds hold true for $\mathbb{E}\left[f(\bar{x}_T) - \min f\right]$, where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

Proof. We simply combine Proposition B.1 with Theorem A.3 and Lemma A.1. We moreover exploit the fact that $\beta_t = 0$ in Proposition B.1, which allows us to replace Assumption 6.5 with Assumption 6.4 in Theorem A.3, see Remark A.4 for more details. In the statement, we introduced the notation e'_t such that $e_t = \frac{\gamma^2}{1-\gamma L}e'_t$. □

Remark B.3. *We note that the parameters selected in Proposition B.1 include $\beta_t \equiv 0$, which, in line with Remark A.4, justify that we only require Assumption 6.4 instead of Assumption 6.5.*

B.2 Bounds for the Optimal Step-Size

Considering the result of Theorem B.2 and letting $\gamma L \to 1$, we note that the bias term tends to $\frac{1}{2\gamma T}$ while the variance term diverges since its denominator is a multiple of $1-\gamma L$. This suggests there might be a complication to obtain tight bounds in the case $\gamma L = 1$. As seen in Section 5, we have empirical results showcasing that it is not possible to obtain a bound for SGD where the bias term is $\frac{1}{2\gamma T}$ for $\gamma L = 1$ whilst preserving a finite variance term. Therefore, in this section, we will prove bounds where the bias term is given by $\frac{1}{\rho T}$, with $\rho < 2\gamma$.

Proposition B.4 (Lyapunov parameters. Convex case, optimal step-size)**.** *Let $\gamma L = 1$ and $\varepsilon \in (0, 2)$, and consider the parameters $\rho, a_t, e_t, \alpha_t, \beta_t$ defined by*

- $\rho = (2 - \varepsilon)\gamma$,
- $a_t \equiv a = 1$,
- $\alpha_t = \gamma^2$,
- $\beta_t = \frac{\gamma^2 \varepsilon}{2}$,
- $e_t = \gamma^2 \frac{2+\varepsilon}{\varepsilon}$.

These parameters satisfy the sufficient Lyapunov conditions of Theorem A.3.

Proof. First, let us impose that $a_0 = 1$ without loss of generality. Second, we set $\rho = (2 - \varepsilon)\gamma$ for $\varepsilon > 0$ since numerical experiments suggest that $\rho = 2\gamma$ is impossible. Third, taking inspiration from our numerical findings, we impose that the sequence a_t is constantly equal to 1, that is $a_t \equiv a = 1$.

If we want to satisfy the sufficient Lyapunov conditions of Theorem A.3, we need in particular to verify Condition 5, which reduces to

$$\gamma = L\alpha_t,$$

because $a_t - a_{t+1} = 0$. This means that $\alpha_t \equiv \alpha = \gamma^2$, since $\gamma L = 1$. Condition 2 implies that β_t must satisfy

$$\beta_t \leq \alpha - \frac{\rho}{2L} = \frac{\gamma^2 \varepsilon}{2}.$$

Condition 4 is trivially satisfied since it requires

$$\gamma^2 \leq \alpha_t + \beta_t = \gamma^2 + \beta_t,$$

that is $\beta_t \geq 0$. It remains to verify Condition 6, which requires that

$$e_t \geq \frac{\gamma^2(\alpha + \beta_t)}{\alpha + \beta_t - \gamma^2}.$$

First, observe that this is the only constraint on e_t, which we want to see as small as possible. So we shall set the inequality to be an equality. Second, the expression is decreasing with respect to β_t (its derivative has the same sign as $-\gamma^2$). Therefore, e_t will be minimal if β_t is maximal, meaning we shall fix β_t to constantly be equal to its earlier derived upper bound, that is $\beta_t \equiv \beta = \frac{\gamma^2 \varepsilon}{2}$. Finally, we compute

$$e_t = \gamma^2 \frac{\gamma^2 + \frac{\gamma^2 \varepsilon}{2}}{\gamma^2 + \frac{\gamma^2 \varepsilon}{2} - \gamma^2} = \gamma^2 \frac{2 + \varepsilon}{\varepsilon}.$$

Note in particular that e_t is constant here. □

Theorem B.5 (Bounds for SGD. Convex case, optimal step-size)**.** *Let Assumptions 6.5 and 6.7 hold, with $\mu = 0$. Let x_t be generated by SGD, with $\gamma L = 1$. Then, for every $T \geq 1$ and $\varepsilon \in (0, 2)$,*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(x_t) - \inf f] \leq \frac{\|x_0 - x_*\|^2}{(2 - \varepsilon)\gamma T} + \frac{(2 + \varepsilon)\gamma \sigma_*^2}{\varepsilon(2 - \varepsilon)}.$$

If moreover f is convex, the above bound holds true for $\mathbb{E}\left[f(\bar{x}_T) - \min f\right]$, where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

Proof. We simply combine Proposition B.4 with Theorem A.3 and Lemma A.1. □

B.3 Bounds for Large Step-Sizes

Proposition B.6 (Lyapunov parameters. Convex case, large step-sizes)**.** *Let $1 < \gamma L < 2$, and consider the parameters $\rho, a_t, e_t, \alpha_t, \beta_t$ defined by*

- $\rho = \frac{2\gamma(2-\gamma L)}{1-\theta^T}$ *where* $\theta = (1-\gamma L)^2$,
- $a_t = \frac{1-\theta^{T-t}}{1-\theta^T}$,
- $\beta_t = \frac{\gamma(\gamma L-1)(1-\theta^{T-t-1})}{L(1-\theta^T)}$,
- $\alpha_t = \gamma\frac{(2-\gamma L)+(\gamma L-1)(1-\theta^{T-t-1})}{L(1-\theta^T)}$,
- $e_t = \frac{\gamma^2}{2-\gamma L}\frac{1-\theta^{T-t-1}}{1-\theta^T}\left(\frac{\gamma L}{\theta^{T-t-1}} - 2(\gamma L-1)\right)$.

These parameters satisfy the sufficient Lyapunov conditions of Theorem A.3. Moreover, the averaged variance $\bar{e}_T \coloneqq \frac{1}{T}\sum_{t=0}^{T-1} e_t$ has the following asymptotic equivalence as $T \to +\infty$:

$$\bar{e}_T \sim \frac{\gamma^2}{(2-\gamma L)^2}\frac{1}{T(1-\gamma L)^{2T-2}} \to +\infty.$$

Proof. To determine the parameters, we will rely on insights from our numerical experiments. Specifically, we were able to guess that

$$\beta_t = \frac{a_{t+1}\gamma(\gamma L-1)}{L},$$

and that Condition 2 is satisfied with equality, meaning that

$$\rho = 2L(\alpha_t + \beta_t) \iff \alpha_t = \frac{\rho}{2L} + \beta_t,$$

Considering Condition 5, we have

$$(a_{t+1}\gamma - L\alpha_t)^2 \leq (\alpha_t + \beta_t - \gamma^2 a_{t+1})(a_t - a_{t+1}).$$

The latter will give us an expression of a_t in terms of a_{t+1}, from which we will deduce a general expression for a_t by induction. By injecting the values of α_t and β_t, we obtain

$$\frac{1}{L^2}\left(\frac{\rho L}{2} - a_{t+1}\gamma L(2-\gamma L)\right)^2 \leq \frac{1}{L^2}\left(\frac{\rho L}{2} - a_{t+1}\gamma L(2-\gamma L)\right)(a_t - a_{t+1}).$$

This is satisfied when $\rho > 2\gamma L(2-\gamma L)$ and

$$\frac{\rho L}{2} - a_{t+1}\gamma L(2-\gamma L) \leq a_t - a_{t+1} \iff a_t \geq \frac{\rho L}{2} + a_{t+1}(\gamma L-1)^2.$$

We observe an arithmetic-geometric relation given by, where we set $\theta = (\gamma L - 1)^2$,

$$a_t \geq \frac{\rho L}{2} + a_{t+1}\theta \iff a_{t+1} \leq \theta^{-1} a_t - \frac{\rho L}{2\theta}.$$

A simple induction argument leads to

$$a_t \leq \theta^{-t} a_0 - \frac{\rho L}{2\theta} \frac{\theta^{-t} - 1}{\theta^{-1} - 1}.$$

As done previously, without loss of generality, we impose $a_0 = 1$. We impose the sequence to be nonincreasing and nonnegative. The latter leads to

$$a_T \geq 0 \iff \frac{\rho L}{2\theta} \frac{\theta^{-T} - 1}{\theta^{-1} - 1} \leq \theta^{-T} \iff \frac{\rho L}{2} \leq \theta^{-T} \frac{1 - \theta}{\theta^{-T} - 1} = \frac{1 - \theta}{1 - \theta^T}.$$

As we aim to maximize ρ, we impose

$$\frac{\rho L}{2} = \frac{1 - \theta}{1 - \theta^T} \iff \rho = \frac{2(1 - \theta)}{L(1 - \theta^T)} = \frac{2\gamma(2 - \gamma L)}{1 - \theta^T}.$$

In particular, we obtain

$$a_t \leq \theta^{-t} - \frac{\rho L}{2} \frac{\theta^{-t} - 1}{1 - \theta} = \theta^{-t} - \frac{1 - \theta}{1 - \theta^T} \frac{\theta^{-t} - 1}{1 - \theta} = \theta^{-t} - \frac{\theta^{-t} - 1}{1 - \theta^T} = \frac{1 - \theta^{T-t}}{1 - \theta^T}.$$

To enforce that (a_t) is nonincreasing, we set the above to be an equality, and since $a_T \geq 0$, it is also nonnegative, as wanted.

We now verify the remaining conditions from Theorem A.3. The nonnegativity of a_t implies the nonnegativity of $\alpha_t, \beta_t,$. Conditions 1, 3, 5 are verified by construction. It remains to check that Conditions 4 and 6 hold true.

In the following, we introduce $k \coloneqq T - t - 1$ to simplify notation. To check that Condition 4 is satisfied, we compute $\alpha_t + \beta_t$:

$$\alpha_t + \beta_t = \frac{\rho}{2L} + 2\beta_t = \frac{\gamma(2 - \gamma L) + 2\gamma(\gamma L - 1)(1 - \theta^k)}{L(1 - \theta^T)}.$$

Therefore, Condition 4, which states $\alpha_t + \beta_t - \gamma^2 a_{t+1} \geq 0$, becomes

$$\begin{aligned}
0 &\leq \frac{\gamma(2 - \gamma L) + 2\gamma(\gamma L - 1)(1 - \theta^k)}{L(1 - \theta^T)} - \gamma^2 \frac{1 - \theta^k}{1 - \theta^T} \\
&= \frac{\gamma(2 - \gamma L) + 2\gamma(\gamma L - 1)(1 - \theta^k) - \gamma^2 L(1 - \theta^k)}{L(1 - \theta^T)} \\
&= \frac{\gamma(2 - \gamma L) - (1 - \theta^k)\gamma(2 - \gamma L)}{L(1 - \theta^T)} \\
&= \frac{\gamma(2 - \gamma L)\theta^k}{L(1 - \theta^T)},
\end{aligned}$$

which is true because $\gamma L \in (1,2)$. It remains to study Condition 6, which is satisfied by setting

$$\begin{aligned}
e_t &= \gamma^2 \frac{a_{t+1}(\alpha_t + \beta_t)}{\alpha_t + \beta_t - \gamma^2 a_{t+1}} \\
&= \gamma^2 \frac{1-\theta^k}{1-\theta^T} \frac{\gamma(2-\gamma L) + 2\gamma(\gamma L - 1)(1-\theta^k)}{L(1-\theta^T)} \frac{L(1-\theta^T)}{\gamma(2-\gamma L)\theta^k} \\
&= \gamma^2 \frac{1-\theta^k}{1-\theta^T} \frac{(2-\gamma L) + 2(\gamma L - 1)(1-\theta^k)}{(2-\gamma L)\theta^k} \\
&= \gamma^2 \frac{1-\theta^k}{1-\theta^T} \frac{\gamma L - 2(\gamma L - 1)\theta^k}{(2-\gamma L)\theta^k} \\
&= \frac{\gamma^2}{2-\gamma L} \frac{1-\theta^k}{1-\theta^T} \left(\frac{\gamma L}{\theta^k} - 2(\gamma L - 1) \right).
\end{aligned}$$

To conclude the proof, it remains to analyze the average of the e_t's. Denote $e_t'' = (1-\theta^k)\left(\frac{\gamma L}{\theta^k} - 2(\gamma L - 1)\right)$, such that

$$\begin{aligned}
\sum_{t=0}^{T-1} e_t'' &= \gamma L \sum_{t=0}^{T-1} \frac{1-\theta^k}{\theta^k} - 2(\gamma L - 1) \sum_{t=0}^{T-1} (1-\theta^k) \\
&= \gamma L \sum_{t=0}^{T-1} (\theta^{-k} - 1) - 2(\gamma L - 1) \sum_{t=0}^{T-1} (1-\theta^k) \\
&= \gamma L \left(\sum_{t=0}^{T-1} \theta^{-k} \right) - T\gamma L - 2(\gamma L - 1)T + 2(\gamma L - 1) \left(\sum_{t=0}^{T-1} \theta^k \right) \\
&= \gamma L \left(\sum_{t=0}^{T-1} \theta^{-t} \right) - T(3\gamma L - 2) + 2(\gamma L - 1) \left(\sum_{t=0}^{T-1} \theta^t \right) \\
&= \gamma L \frac{1-\theta^{-T}}{1-\theta^{-1}} - T(3\gamma L - 2) + 2(\gamma L - 1) \frac{1-\theta^T}{1-\theta} \\
&= \frac{\theta}{2-\gamma L}(\theta^{-T} - 1) - T(3\gamma L - 2) + \frac{2(\gamma L - 1)}{\gamma L(2-\gamma L)}(1-\theta^T),
\end{aligned}$$

which is asymptotically equivalent to $\frac{\theta}{2-\gamma L}\theta^{-T}$. Therefore we can write

$$\bar{e}_T = \frac{\gamma^2}{2-\gamma L} \frac{1}{1-\theta^T} \frac{1}{2-\gamma L} \bar{e}_T',$$

where $e_t' = (2-\gamma L) e_t''$ and

$$\bar{e}_T' = \frac{1}{T} \sum_{t=0}^{T-1} e_t' = \frac{\theta}{T}(\theta^{-T} - 1) - (2-\gamma L)(3\gamma L - 2) + \frac{2(\gamma L - 1)}{\gamma L T}(1-\theta^T). \tag{13}$$

In particular, one sees that $\bar{e}_T' \sim \frac{1}{T\theta^{T-1}}$. □

Theorem B.7 (Bounds for SGD. Convex case, large step-sizes)**.** *Let Assumptions 6.5 and 6.7 hold, with $\mu = 0$. Let x_t be generated by SGD, with $\gamma L \in (1, 2)$. Then, for every $T \geq 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[f(x_t) - \inf f] \leq \frac{\|x_0 - x_*\|^2 \delta}{2\gamma(2-\gamma L)T} + \frac{\gamma \sigma_*^2 \bar{e}'_T}{2(2-\gamma L)^3},$$

where $\delta = 1 - (1-\gamma L)^{2T} \in (0,1)$, and $\bar{e}'_T$ is defined in (13) and grows exponentially with T. If moreover f is convex, the above bound holds true for $\mathbb{E}\left[f(\bar{x}_T) - \min f\right]$, where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

Proof. We simply combine Proposition B.6 with Theorem A.3 and Lemma A.1. □

We may moreover show that the bias term in Theorem B.7 is tight, in the sense that no Lyapunov proof using energies of our form may improve upon the bias term.

Proposition B.8. *Assume that $\mathbb{E}[E_{t+1}] \leq \mathbb{E}[E_t]$ for all $t = 0, \ldots, T-1$. If $\gamma L > 1$ and $a_0 = 1$, then necessarily*

$$\rho \leq \frac{2\gamma(2-\gamma L)}{1-(1-\gamma L)^{2T}}.$$

Proof. We consider $d = 1$ and $m = 2$, and let $f : \mathbb{R} \to \mathbb{R}$ be defined through

$$f(x) = \frac{1}{2}\left(f_+(x) + f_-(x)\right) = \frac{1}{2}\left(\frac{L}{2}(x+\delta)^2 + \frac{L}{2}(x-\delta)^2\right).$$

One can easily compute that

$$f(x) = \frac{L\left(x^2+\delta^2\right)}{2}, \quad \min f = \frac{L\delta^2}{2}, \quad f(x) - \min f = \frac{Lx^2}{2}, \quad x_* = 0$$

$$\nabla f(x) = Lx, \quad \nabla f_\pm(x) = L(x \pm \delta), \quad \nabla f_\pm(x_*) = \pm L\delta,$$

We know that $\mathbb{E}[E_{t+1} - E_t] \leq 0$ for all $t = 0, \ldots, T-1$ and for all $x_0 \in \mathbb{R}$. We select x_0 sufficiently large such that x_t is sufficiently large (to be defined) for all $t = 0, \ldots, T-1$. Expanding the Lyapunov decrease and rearranging yields

$$\begin{aligned}
&a_{t+1}\mathbb{E}_t\left[|x_{t+1}|^2\right] - a_t|x_t|^2 + \rho\left(f(x_t) - \min f\right) \leq e_t\sigma_*^2 \\
\iff &a_{t+1}\left((1-\gamma L)^2|x_t|^2 + (\gamma L)^2\delta^2\right) - a_t|x_t|^2 + \rho L\left(\frac{|x_t|^2+\delta^2}{2} - \frac{\delta^2}{2}\right) \leq 2e_t\delta^2 \\
\iff &\left(a_{t+1}(1-\gamma L)^2 - a_t + \frac{\rho L}{2}\right)|x_t|^2 \leq \left(2e_t - a_{t+1}(\gamma L)^2\right)\delta^2.
\end{aligned}$$

As x_t is sufficiently large, and the right-hand side is constant, $a_{t+1}(1-\gamma L)^2 - a_t + \frac{\rho L}{2}$ must be nonnegative, or equivalently

$$a_{t+1} \leq \frac{1}{(1-\gamma L)^2}a_t - \frac{L}{2(1-\gamma L)^2}\rho.$$

To simplify the notations, we note $\tau = (1-\gamma L)^2$, such that $a_{t+1} \leq \tau^{-1} a_t - \tau^{-1}\frac{\rho L}{2}$. By induction, we deduce that $a_T \leq \tau^{-T} a_0 - \frac{\rho L}{2}\sum_{t=1}^{T}\tau^{-t}$. Because $a_T \geq 0$, we obtain that

$$1 = a_0 \geq \frac{\rho L}{2}\sum_{t=1}^{T}\tau^{T-t} = \frac{\rho L}{2}\sum_{t=0}^{T-1}\tau^{T-t-1} \iff \rho \leq \frac{2a_0}{L\sum_{t=0}^{T-1}(1-\gamma L)^{2t}} = \frac{2\gamma(2-\gamma L)}{1-(1-\gamma L)^{2T}},$$

concluding the proof. □

C Proofs in the Smooth Strongly Convex Setting

As in the previous section, we derive parameters that satisfy the sufficient conditions of Theorem A.3 in order to derive upper bounds for SGD, this time in the strongly convex setting. We again note that all propositions could simply be verified symbolically, yet we include the proofs to provide intuition on how these parameters were derived.

C.1 Bounds for any Step-Size

Because it seems impossible to obtain a bound with tight bias term for optimal step-sizes (see Section 5.2), we provide here a more relaxed bound, where the bias term is ε away from being tight, and subsequently the variance term remains bounded for every step-size $\gamma L \in (0, 2)$. Moreover, due to technical reasons, we plit the proofs into the cases $L > \mu$ and $L = \mu$.

Proposition C.1 (Lyapunov parameters. Strongly convex case, any step-sizes, $L > \mu$)**.** *Let $\gamma L \in (0, 2)$, with $L > \mu$. Let $\varepsilon \geq 0$ and $\gamma_{opt} = \frac{2}{\mu+L}$. Assume that either $\varepsilon > 0$ or $\gamma \neq \gamma_{opt}$. Consider the parameters $\rho, a_t, e_t, \alpha_t, \beta_t$ defined by*

- $\rho = 0$,
- $a_t = \phi^{2(T-t)}$, *where* $\phi^2 = \phi_{opt}^2 + \varepsilon$, *with*

$$\phi_{opt} = \max\{1 - \gamma\mu; \gamma L - 1\} \quad \text{and} \quad \varepsilon \in [0, 1 - \phi_{opt}^2),$$

- $\alpha_t = \beta_t = \alpha a_{t+1}$, *where* $\alpha = \omega + \omega_\varepsilon$ *and*

$$\omega = \frac{\gamma\phi_{opt}}{L - \mu} \quad \text{and} \quad \omega_\varepsilon = \frac{\varepsilon + \sqrt{\varepsilon^2 + \varepsilon\gamma(L-\mu)(L+\mu)\delta}}{(L-\mu)^2} \quad \text{with} \quad \delta = |\gamma - \gamma_{opt}|,$$

- $e_t = e a_{t+1}$, *where*

$$e = \gamma^2 \left(1 + \frac{\gamma^2(L-\mu)}{2\omega_\varepsilon(L-\mu) + (L+\mu)\gamma\delta}\right).$$

Then the sufficient Lyapunov conditions of Theorem A.3 are verified. Moreover, we have

$$e_T^{sum} := \sum_{t=0}^{T-1} e_t = e\frac{1 - \phi^{2T}}{1 - \phi^2} \leq \frac{e}{1 - \phi^2},$$

and

$$e \leq \gamma^2 \left(1 + \frac{\gamma^2(L-\mu)^2}{4\varepsilon + (L+\mu)(L-\mu)\gamma\delta}\right).$$

Proof. We start the proof by exploiting some insights from our numerical analysis. First, we observe that we can impose $\rho = 0$ and $\alpha_t = \beta_t$ without changing anything to the empirical best bias term. We moreover infer that $a_t/a_{t+1} \in (0, 1)$ is constant, so let us denote this ratio by ϕ^2. Since ϕ_{opt}^2 is the optimal rate for the deterministic gradient descent in the strongly convex smooth case, we anticipate that $\phi^2 \geq \phi_{\text{opt}}^2$. Without loss of generality, we impose $a_T = 1$, from which we

deduce that $a_t = \phi^{2(T-t)}$. We were also able to deduce that $e_t = ea_{t+1}$ and that $\alpha_t = \alpha a_{t+1}$ for some constants $e, \alpha > 0$, yet to be determined.

It remains to study the conditions of Theorem A.3 to find feasible values for α and e.

Condition 6 yields

$$a_{t+1}\gamma^2(2\alpha a_{t+1}) \leq (2\alpha a_{t+1} - a_{t+1}\gamma^2)ea_{t+1},$$

which holds true if

$$2\gamma^2\alpha \leq (2\alpha - \gamma^2)e.$$

As we want to minimize e, we set

$$e = \frac{2\gamma^2\alpha}{2\alpha - \gamma^2},$$

and impose that the denominator $2\alpha - \gamma^2$ is positive. Now it remains to find α. The above expression is decreasing with α, so we aim to find the largest α possible such that all conditions are satisfied.

Turning to Condition 5 will determine α. With our current notations and restrictions on the parameters, it can be written as

$$(a_{t+1}\gamma - \alpha a_{t+1}(L+\mu))^2 \leq (2\mu L\alpha a_{t+1} + a_t - a_{t+1})(2\alpha a_{t+1} - a_{t+1}\gamma^2),$$

which is true if

$$(\gamma - \alpha(L+\mu))^2 - (2\mu L\alpha - (1-\phi^2))(2\alpha - \gamma^2) \leq 0.$$

This is a polynomial of degree at most 2 in α, which must be nonpositive. Expanding the terms to find the coefficients of this polynomial $P(\alpha)$, we obtain

$$\begin{aligned} P(\alpha) &= \gamma^2 + \alpha^2(L+\mu)^2 - 2\alpha(L+\mu)\gamma - 4\mu L\alpha^2 + 2\mu L\alpha\gamma^2 + 2\alpha(1-\phi^2) - \gamma^2(1-\phi^2) \\ &= \alpha^2\left[(L+\mu)^2 - 4\mu L\right] - \alpha\left[2\gamma(L+\mu) - 2\mu L\gamma^2 - 2(1-\phi^2)\right] + \left[\gamma^2 - \gamma^2(1-\phi^2)\right] \\ &= a\alpha^2 - 2b\alpha + c, \end{aligned}$$

where

$$\begin{cases} a &= (L+\mu)^2 - 4\mu L = (L-\mu)^2, \\ b &= \gamma(L+\mu) - \mu L\gamma^2 - (1-\phi^2), \\ c &= \gamma^2\phi^2 = \gamma^2(\phi_{opt}^2 + \varepsilon). \end{cases} \tag{14}$$

It is a simple exercise to verify that $\phi_{opt} = 1 - \mu\gamma$ if $\gamma \leq \gamma_{opt}$ and $\phi_{opt} = \gamma L - 1$ if $\gamma \geq \gamma_{opt}$, which leads us to compute the value of b through a case distinction:

- if $\gamma \leq \gamma_{opt}$ then $1 - \phi_{opt}^2 = \gamma\mu(2 - \gamma\mu)$ and

$$\begin{aligned} b &= \gamma(L+\mu) - \mu L\gamma^2 - \gamma\mu(2-\gamma\mu) + \varepsilon \\ &= \gamma\left(L + \mu - \mu L\gamma - 2\mu + \gamma\mu^2\right) + \varepsilon \\ &= \gamma(1-\gamma\mu)(L-\mu) + \varepsilon \\ &= \gamma\phi_{opt}(L-\mu) + \varepsilon. \end{aligned}$$

- if $\gamma \geq \gamma_{opt}$ then $1 - \phi_{opt}^2 = \gamma L(2 - \gamma L)$ and

$$\begin{aligned} b &= \gamma(L+\mu) - \mu L\gamma^2 - \gamma L(2-\gamma L) + \varepsilon \\ &= \gamma\left(L + \mu - \mu L\gamma - 2L + \gamma L^2\right) + \varepsilon \\ &= \gamma(\gamma L - 1)(L-\mu) + \varepsilon \\ &= \gamma\phi_{opt}(L-\mu) + \varepsilon. \end{aligned}$$

In both cases we obtain $b = \gamma\phi_{opt}(L-\mu) + \varepsilon \geq 0$.

Since $L > \mu$, we have that $a > 0$, so the largest feasible α is the largest root of P, if existent. The discriminant Δ of P is given by

$$\begin{aligned} \Delta &= (2b)^2 - 4ac \\ &= 4\left(\gamma\phi_{opt}(L-\mu) + \varepsilon\right)^2 - 4(L-\mu)^2\gamma^2(\phi_{opt}^2 + \varepsilon) \\ &= 4\gamma^2\phi_{opt}^2(L-\mu)^2 + 4\varepsilon^2 + 8\varepsilon\gamma\phi_{opt}(L-\mu) - 4(L-\mu)^2\gamma^2\phi_{opt}^2 - 4(L-\mu)^2\gamma^2\varepsilon \\ &= 4\left[\varepsilon^2 + \varepsilon\left(2\gamma\phi_{opt}(L-\mu) - (L-\mu)^2\gamma^2\right)\right] \\ &= 4\left[\varepsilon^2 + \varepsilon\gamma(L-\mu)\left(2\phi_{opt} - (L-\mu)\gamma\right)\right]. \end{aligned}$$

A simple calculation shows that

$$2\phi_{opt} - (L-\mu)\gamma = \begin{cases} 2 - \gamma(L+\mu) & \text{if } \gamma \leq \gamma_{opt} \\ -2 + \gamma(L+\mu) & \text{if } \gamma \geq \gamma_{opt} \end{cases} = |2 - \gamma(L+\mu)| = (L+\mu)|\gamma - \gamma_{opt}|,$$

such that, with $\delta = |\gamma - \gamma_{opt}|$,

$$\Delta = 4\left[\varepsilon^2 + \varepsilon\gamma(L-\mu)(L+\mu)\delta\right] \geq 0.$$

As such, P has real roots. We thus define

$$\begin{aligned} \alpha &= \frac{2b + \sqrt{\Delta}}{2a} \\ &= \frac{\gamma\phi_{opt}(L-\mu) + \varepsilon + \sqrt{\varepsilon^2 + \varepsilon\gamma(L-\mu)(L+\mu)\delta}}{(L-\mu)^2} \\ &= \omega + \omega_\varepsilon, \end{aligned}$$

where

$$\omega = \frac{\gamma\phi_{opt}}{L-\mu} \quad \text{and} \quad \omega_\varepsilon = \frac{\varepsilon + \sqrt{\varepsilon^2 + \varepsilon\gamma(L-\mu)(L+\mu)\delta}}{(L-\mu)^2}.$$

To conclude the proof, we are now going to verify that this choice of α is feasible, in the sense that it satisfies all the conditions from Theorem A.3. The nonnegativity Condition 1 is satisfied by construction. Condition 2 is satisfied as we set $\rho = 0$ and $\alpha_t = \beta_t$. Condition 3 is equivalent to

$$1 \leq 2\alpha\mu L + \frac{a_t}{a_{t+1}} = 2\alpha\mu L + \phi^2.$$

We have $\alpha = \omega + \omega_\varepsilon \geq \omega$ and $\phi^2 \geq \phi_{opt}^2$, and the above is thus true if

$$\begin{aligned} & 1 \leq \mu L 2\omega + \phi_{opt}^2 \\ \iff \quad & 1 \leq \frac{2\mu L \gamma \phi_{opt}}{L-\mu} + \phi_{opt}^2 \\ \iff \quad & 0 \leq 2\mu L \gamma \phi_{opt} - (1-\phi_{opt}^2)(L-\mu). \end{aligned}$$

To assess whether this inequality is true, we are going to consider two cases:

- if $\gamma \leq \gamma_{opt}$, then $\phi_{opt} = 1 - \gamma\mu$, so the inequality becomes

$$\begin{aligned} 0 &\leq 2\mu L\gamma(1-\gamma\mu) - \gamma\mu(2-\gamma\mu)(L-\mu) \\ &= \mu\gamma(-\mu\gamma L - \gamma\mu^2 + 2\mu) \\ &= \mu^2\gamma(2-\gamma(L+\mu)) \\ &= \mu^2\gamma(\mu+L)\delta. \end{aligned}$$

- if $\gamma \geq \gamma_{opt}$, then $\phi_{opt} = \gamma L - 1$, so the inequality becomes

$$\begin{aligned} 0 &\leq 2\mu L\gamma(\gamma L - 1) - \gamma L(2-\gamma L)(L-\mu) \\ &= L\gamma(\mu\gamma L + \gamma L^2 - 2L) \\ &= -L^2\gamma(2-\gamma(L+\mu)) \\ &= L^2\gamma(\mu+L)\delta. \end{aligned}$$

In both cases the inequality is verified, therefore Condition 3 holds true. Let us now turn to Condition 4, which, in our context, is equivalent to

$$0 \leq 2\alpha - \gamma^2,$$

which holds true as

$$\begin{aligned} 2\alpha - \gamma^2 &= 2\omega_\varepsilon + 2\omega - \gamma^2 = 2\omega_\varepsilon + \frac{\gamma}{L-\mu}\left(2\phi_{opt} - \gamma(L-\mu)\right) \\ &= 2\omega_\varepsilon + \frac{\gamma}{L-\mu}(L+\mu)\delta. \end{aligned}$$

Condition 5 is true by the choice of α. Finally, Condition 6 holds true provided $2\alpha - \gamma^2 > 0$, which is verified through the above computation provided $\varepsilon > 0$ or $\delta \neq 0$. In this case,

$$\begin{aligned} e &= \gamma^2 \frac{2\omega + 2\omega_\varepsilon}{2\omega_\varepsilon + \frac{L+\mu}{L-\mu}\gamma\delta} \\ &= \gamma^2 \frac{2\gamma\phi_{opt} + 2(L-\mu)\omega_\varepsilon}{2\omega_\varepsilon(L-\mu) + (L+\mu)\gamma\delta} \\ &= \gamma^2 \frac{\gamma((L+\mu)\delta + \gamma(L-\mu)) + 2(L-\mu)\omega_\varepsilon}{2\omega_\varepsilon(L-\mu) + (L+\mu)\gamma\delta} \\ &= \gamma^2 \left(1 + \frac{\gamma^2(L-\mu)}{2\omega_\varepsilon(L-\mu) + (L+\mu)\gamma\delta}\right). \end{aligned}$$

The bound on e_T^{sum} is a direct consequence of the fact that a_t is a geometric sequence. The upper bound for e comes from the lower bound $\omega_\varepsilon \geq \frac{2\varepsilon}{(L-\mu)^2}$. □

We note the above proposition only holds true for $\mu < L$, thus not covering the case $L = \mu$. We cover this case in a separate proposition.

Proposition C.2 (Lyapunov parameters. Strongly convex case, any step-sizes, $L = \mu$)**.** *Let $\gamma L \in (0, 2)$, with $L = \mu$. Let $\varepsilon > 0$. Consider the parameters $\rho, a_t, e_t, \alpha_t, \beta_t$ defined by*

- $\rho = 0$,
- $a_t = \phi^{2(T-t)}$, *where* $\phi^2 = \phi_{opt}^2 + \varepsilon$, *with*

$$\phi_{opt} = \max\{1 - \gamma\mu; \gamma L - 1\} \quad \text{and} \quad \varepsilon \in [0, 1 - \phi_{opt}^2),$$

- $\alpha_t = \beta_t$ *and* $\alpha_t = \alpha a_{t+1}$ *where*

$$\alpha \geq \max\left\{\frac{\gamma^2}{2}\left(1 + \frac{\phi_{opt}^2}{\varepsilon}\right), \frac{1 - \phi_{opt}^2 - \varepsilon}{2L^2}\right\},$$

- $e_t = e a_{t+1}$ *where*

$$e = \frac{2\gamma^2\alpha}{2\alpha - \gamma^2}.$$

Then the sufficient Lyapunov conditions of Theorem A.3. Moreover, we have

$$\phi \xrightarrow{\varepsilon \to 0} \phi_{opt}, \quad \alpha \xrightarrow{\varepsilon \to 0} +\infty, \quad e \xrightarrow{\varepsilon \to 0} \gamma^2.$$

Proof. The arguments in this proof are mostly recycled from Proposition C.1. In particular, we shall use the same values for ρ and a_t. Regarding the value of α, we impose $a\alpha^2 - 2b\alpha + c \leq 0$ where a, b, c are defined in Equation (14), which we recall for convenience:

$$\begin{cases} a &= (L - \mu)^2, \\ b &= \gamma(L + \mu) - \mu L\gamma^2 - (1 - \phi^2), \\ c &= \gamma^2(\phi_{opt}^2 + \varepsilon). \end{cases}$$

Because $L = \mu$, it holds that $a = 0$, and that

$$b = 2\gamma L - L^2\gamma^2 - (1 - \phi_{opt}^2) + \varepsilon = 2\gamma L - L^2\gamma^2 - \gamma L(2 - \gamma L) + \varepsilon = \varepsilon.$$

Therefore, α must satisfy $c \leq 2b\alpha$, that is

$$\alpha \geq \frac{c}{2b} = \frac{\gamma^2(\phi_{opt}^2 + \varepsilon)}{2\varepsilon} = \frac{\gamma^2}{2}\left(1 + \frac{\phi_{opt}^2}{\varepsilon}\right). \tag{15}$$

It then remains to verify the conditions of Theorem A.3. Conditions 1 and 2 are readily satisfied. Condition 3 can be rewritten as

$$\alpha \geq \frac{1 - \phi_{opt}^2 - \varepsilon}{2L^2},$$

thus justified the bound on α. Condition 4 requires that $\alpha \geq \frac{\gamma^2}{2}$, which is already true because of Equation (15). Condition 5 is satisfied under Equation (15). Finally, Condition 6 is satisfied with

$$e = \frac{2\gamma^2\alpha}{2\alpha - \gamma^2},$$

where the denominator is strictly positive because of (15). □

Theorem C.3 (Bound for SGD. Strongly convex case, general step-sizes)**.** *Let Assumptions 6.5 and 6.7 hold, with $\mu > 0$. Let $\varepsilon \geq 0$ and $\gamma_{opt} = \frac{2}{\mu+L}$. Let x_t be generated by SGD, with $\gamma L \in (0, 2)$. Assume that either $\varepsilon > 0$ or $\gamma \neq \gamma_{opt}$. Then, for every $T \geq 1$,*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2 + \frac{1 - \phi^{2T}}{1 - \phi^2} e\sigma_*^2,$$

where $\phi^2 = \varepsilon + (\max\{1 - \gamma\mu; \gamma L - 1\})^2 \in [0, 1)$, and e is defined in Proposition C.1 and verifies

$$e \leq \gamma^2 \left(1 + \frac{\gamma^2(L-\mu)^2}{4\varepsilon + (L+\mu)(L-\mu)\gamma|\gamma - \gamma_{opt}|}\right).$$

Proof. We separate the proof into two cases, namely the case $L > \mu$ and the case $L = \mu$.

- Consider $L > \mu$. Proposition C.1, combined with Theorem A.3, provides parameters ensuring that the Lyapunov energy decreases with, in particular, $a_T = 1$, $a_0 = \phi^{2T}$. We deduce the wanted bound from Lemma A.2.
- Consider $L = \mu$. Let us consider first $\varepsilon > 0$. Combining Lemma A.2, Theorem A.3 and Proposition C.2 we obtain

  $$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2 + \frac{1 - \phi^{2T}}{1 - \phi^2} e_\varepsilon \sigma_*^2,$$

  where e_ε converges to γ^2 and $\phi \to \phi_{\text{opt}}$ when $\varepsilon \to 0$, thus yielding the wanted bound when $\varepsilon = 0$. If we wish to prove the claim for $\varepsilon > 0$, it is enough to realize that $\phi_{opt} \leq \phi$, together with the fact that $\frac{1-\phi^{2T}}{1-\phi^2}$ is nondecreasing with ϕ.

□

The above propositions allow us to formulate bounds for SGD. As in the convex case, we consider optimal and non-optimal step-sizes as separate cases.

Theorem C.4 (Tight bound for SGD. Strongly convex case, non-optimal step-sizes)**.** *Let Assumptions 6.5 and 6.7 hold, with $\mu > 0$. Let x_t be generated by SGD, with $\gamma L \in (0, 2)$ such that $\gamma \neq \gamma_{opt} = \frac{2}{\mu+L}$. Then, for every $T \geq 1$,*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2 + \frac{1 - \phi^{2T}}{1 - \phi^2} e\sigma_*^2,$$

where $\phi = \max\{1 - \gamma\mu; \gamma L - 1\} \in [0, 1)$, and

$$e = \frac{2\gamma^2\phi}{|2 - \gamma(L+\mu)|} = \gamma^2 \left(1 + \frac{\gamma(L-\mu)}{|\gamma - \gamma_{opt}|(L+\mu)}\right).$$

Proof. This is an immediate consequence of Theorem C.3 with $\varepsilon = 0$. □

Finally, we provide a bound which remains valid for every steps-size, with the counterpart that the bias term gets sub-optimal.

Theorem C.5 (Subtight bound for SGD. Strongly convex case, general step-sizes)**.** *Let Assumptions 6.5 and 6.7 hold, with $\mu > 0$. Let $\varepsilon > 0$, and let x_t be generated by SGD, with $\gamma L \in (0, 2)$. Then, for every $T \geq 1$,*

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \phi^{2T} \cdot \|x_0 - x_*\|^2 + \frac{1 - \phi^{2T}}{1 - \phi^2} e\sigma_*^2,$$

where $\phi^2 = \varepsilon + (\max\{1 - \gamma\mu; \gamma L - 1\})^2 \in [0, 1)$, and e was defined in Proposition C.1 and verifies

$$e \leq \gamma^2 \left(1 + \frac{\gamma^2 (L - \mu)^2}{4\varepsilon}\right).$$

Proof. This is an immediate consequence of Theorem C.3, where we used $\varepsilon > 0$ to lower bound the denominator of e. □

D Postponed Proof of Strong Duality

In this section, we prove the strong duality between Problems (Primal) and (Dual), as used in Section 4.

Theorem D.1. *Problem* (Dual) *is the dual of Problem* (Primal)*. Moreover, strong duality holds.*

Proof. To see that Problem (Dual) is in fact the dual of Problem (Primal), we consider the Lagrangian, which is given by

$$L(G_t, F_t, \Lambda_t, \lambda^{(i)}_{t,*}, \lambda^{(i)}_{*,t}) = \operatorname{Tr}\left(\left(-\Delta_t + \sum_{i=1}^{m} \lambda^i_{t,*} A^{(i)}_{t,*} + \sum_{i=1}^{m} \lambda^i_{*,t} A^{(i)}_{*,t} - \Lambda_t\right) G_t\right) + \left(\tilde{\Delta} + \sum_{i=1}^{m} \lambda^i_{*,t}(\boldsymbol{f}_{m+i} - \boldsymbol{f}_i) + \sum_{i=1}^{m} \lambda^i_{*,t}(\boldsymbol{f}_i - \boldsymbol{f}_{m+i})\right) F_t,$$

from which we readily recover Problem (Dual) as the dual. In order to show that strong duality holds, we employ a standard Slater argument (Boyd and Vandenberghe, 2023) and construct a feasible point (G_t, F_t) to Problem (Primal) such that $G_t \succ 0$.

To do so, we follow the same ideas as in (Taylor et al., 2017b, Theorem 6). Specifically, leveraging the discussion in Section 4.2, we shall construct m functions $f_i \in \mathcal{F}_{\mu,L}$ and points $x_t, x_* \in \mathbb{R}^d$ such that the matrix P_t given by Equation (2a) is upper triangular with positive entries on its diagonal. As such, assuming $d \geq 2m$, it will hold that $G_t = P_t^T P_t \succ 0$. We assume without loss of generality that $x_* = 0$, and set $d = 2m$.

We shall start by proving the result for $\mu = 1$ and $L = 6$. This is without loss of generality, as will be shown later.

We introduce the following functions f_i and their gradients ∇f_i for $i = 1, \dots, m$,

$$f_i \colon \mathbb{R}^d \to \mathbb{R}, x \mapsto \frac{1}{2} x^T Q_i x + b_i x \quad \text{and} \quad \nabla f_i \colon \mathbb{R}^d \to \mathbb{R}^d, x \mapsto Q_i x + b_i.$$

We first define f_i for $i = 1, \dots, m-1$. We let Q_i be a block diagonal matrix of the form

$$Q_i = \begin{pmatrix} Q_i^1 & 0 \\ 0 & Q \end{pmatrix} - \boldsymbol{e}_{m+i} \boldsymbol{e}_{m+i}^T,$$

where $Q_i^1 \in \mathbb{R}^{(m-1)\times(m-1)}$ is a diagonal matrix with entries 2 everywhere expect an entry of 5 on the ith component, and $Q \in \mathbb{R}^{(m+1)\times(m+1)}$ is a tridiagonal matrix whose diagonal elements are 4 and whose off diagonal elements are -1. We note that , $\sigma(Q_i) \subset (1, 6)$ for all $i = 1, \dots, m-1$, by the Greshgorian Theorem (Quarteroni et al., 2007).

We define $x_t = x = [y, z]$ with $y \in \mathbb{R}^{m-1}$ and $z \in \mathbb{R}^{m+1}$. Then, for $i = 1, \dots, m-1$,

$$\nabla f_i(x_t) = g_t^{(i)} = 2 \sum_{j=1, j\neq i}^{m-1} \boldsymbol{e}_j x_j + 5 \boldsymbol{e}_i x_i + Q z - \boldsymbol{e}_{m+i} x_{m+i} + b_i.$$

As such, we define b_i as

$$b_i = -2 \sum_{j=1, j\neq i}^{m-1} \boldsymbol{e}_j x_j + \boldsymbol{e}_{m+i} x_{m+i} - \boldsymbol{e}_m,$$

and we impose that $x > 0$ and that $Qz = [1, 0, \ldots, 0] \in \mathbb{R}^{m+1}$ (may be shown possible by a simple induction argument). This thus gives us that

$$g_t^{(i)} = 5\boldsymbol{e}_i x_i.$$

Moreover, as $x_* = 0$, we have that $g_*^{(i)} = b_i$. Note that $g_t^{(i)}$ has a positive entry in its ith component and 0s everywhere else. Moreover, $g_*^{(i)}$ has a positive component in its $m+i$th component and 0 in all subsequent components. The vector $x_t - x_*$ also has a positive entry in component $2m$. In order to show that P_t is upper triangular with positive diagonal entries, we thus only need to construct $g_t^{(m)}$.

By construction we require $\nabla f(x_*) = 0$, namely that

$$b_m = -\sum_{i=1}^{m-1} b_i = \sum_{i=1}^{m-1} (2m-1)\boldsymbol{e}_i x_i - \sum_{i=1}^{m-1} \boldsymbol{e}_{m+i} x_{m+i} + (m-1)\boldsymbol{e}_m.$$

As such, we may define Q_m to be the identity matrix (which satisfies $\sigma(Q_m) = \{1\}$), such that

$$\nabla f_m(x_t) = g_t^{(m)} = \sum_{i=1}^{m-1} 2m\boldsymbol{e}_i x_i + m\boldsymbol{e}_m.$$

As such, $g_t^{(m)}$ has a positive entry in its mth component and zero entries in all subsequent components. As such, P_t is upper triangular with positive diagonal entries, as wanted.

As such, if $(\mu, L) = (1, 6)$, the result holds. For arbitrary (μ, L), we consider the following transformed functions

$$\tilde{f}_i(x) = \frac{L-\mu}{5}\left(f_i(x) - \frac{1}{2}\|x\|^2\right) + \frac{\mu}{2}\|x\|^2.vg$$

As $f_i \in \mathcal{F}_{1,6}$, we have $\tilde{f}_i \in \mathcal{F}_{\mu,L}$. Moreover, this transformation preserves the property on P_t, hence constructing a Slater point for all (μ, L). □

References

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2022). The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions. *Optimization Letters*, 16(6):1649–1661.

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2023). Conditions for linear convergence of the gradient method for non-convex optimization. *Optimization Letters*, 17(5):1105–1125.

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2024). On the Rate of Convergence of the Difference-of-Convex Algorithm (DCA). *Journal of Optimization Theory and Applications*, 202(1):475–496.

Alacaoglu, A., Malitsky, Y., and Wright, S. J. (2025). Towards Weaker Variance Assumptions for Stochastic Optimization. arXiv preprint arXiv:2504.09951.

Altschuler, J. J. M. . (2018). *Greed, hedging, and acceleration in convex optimization*. PhD thesis, Massachusetts Institute of Technology.

Asi, H. and Duchi, J. C. (2019). Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290.

Bach, F. and Moulines, E. (2011). Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Bauschke, H. H. and Combettes, P. L. (2017). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer International Publishing, Cham.

Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195.

Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient Convergence in Gradient methods with Errors. *SIAM Journal on Optimization*, 10(3):627–642.

Blum, J. R. (1954). Approximation Methods which Converge with Probability one. *The Annals of Mathematical Statistics*, 25(2):382–386.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311.

Bousselmi, N., Hendrickx, J. M., and Glineur, F. (2024a). Interpolation Conditions for Linear Operators and Applications to Performance Estimation Problems. *SIAM Journal on Optimization*, 34(3):3033–3063.

Bousselmi, N., Pustelnik, N., Hendrickx, J. M., and Glineur, F. (2024b). Comparison of Proximal First-Order Primal and Primal-Dual algorithms via Performance Estimation. arXiv preprint arXiv:2403.10209.

Boyd, S. P. and Vandenberghe, L. (2023). *Convex optimization*. Cambridge University Press.

Bubeck, S. (2015). Convex Optimization: Algorithms and Complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357.

Combettes, P. L. and Pesquet, J.-C. (2015). Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248.

Daccache, A. (2019). *Performance estimation of the gradient method with fixed arbitrary step sizes*. PhD thesis, Ecole polytechnique de Louvain.

Das Gupta, S., Freund, R. M., Sun, X. A., and Taylor, A. (2024a). Nonlinear conjugate gradient methods: worst-case convergence rates via computer-assisted analyses. *Mathematical Programming*.

Das Gupta, S., Van Parys, B. P. G., and Ryu, E. K. (2024b). Branch-and-bound performance estimation programming: a unified methodology for constructing optimal optimization methods. *Mathematical Programming*, 204(1):567–639.

Davis, D. and Drusvyatskiy, D. (2019). Stochastic Model-Based Minimization of Weakly Convex Functions. *SIAM Journal on Optimization*, 29(1):207–239.

de Klerk, E., Glineur, F., and Taylor, A. B. (2017). On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199.

De Klerk, E., Glineur, F., and Taylor, A. B. (2020). Worst-Case Convergence Analysis of Inexact Gradient and Newton Methods Through Semidefinite Programming Performance Estimation. *SIAM Journal on Optimization*, 30(3):2053–2082.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Dragomir, R.-A., Taylor, A. B., d'Aspremont, A., and Bolte, J. (2022). Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194(1):41–83.

Drori, Y. and Teboulle, M. (2014). Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482.

Drori, Y. and Teboulle, M. (2016). An optimal variant of Kelley's cutting-plane method. *Mathematical Programming*, 160(1-2):321–351.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Eloi, D. (2022). *Worst-case functions for the gradient method with fixed variable step sizes*. PhD thesis, Ecole polytechnique de Louvain.

Fercoq, O. (2024). Defining Lyapunov functions as the solution of a performance estimation saddle point problem. arXiv preprint arXiv:2411.12317.

Garrigos, G. and Gower, R. M. (2024). Handbook of Convergence Theorems for (Stochastic) Gradient Methods. arXiv preprint arXiv:2301.11235.

Garrigos, G., Rosasco, L., and Villa, S. (2023). Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. *Mathematical Programming*, 198(1):937–996.

Gladyshev, E. G. (1965). On Stochastic Approximation. *Theory of Probability & Its Applications*, 10(2):275–278.

Goujaud, B., Taylor, A., and Dieuleveut, A. (2022). Optimal first-order methods for convex functions with a quadratic upper bound. arXiv preprint arXiv:2205.15033.

Goulart, P. J. and Chen, Y. (2024). Clarabel: An interior-point solver for conic programs with quadratic objectives. arXiv preprint arXiv:2405.12762.

Gower, R., Sebbouh, O., and Loizou, N. (2021a). SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation. In *Proceedings of the 24th 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 *Proceedings of the 36th International Conference on Machine Learning*, pages 5200–5209. PMLR.

Gower, R. M., Richtárik, P., and Bach, F. (2021b). Stochastic quasi-gradient methods: variance reduction via Jacobian sketching. *Mathematical Programming*, 188(1):135–192.

Grimmer, B. (2024). Provably Faster Gradient Descent via Long Steps. *SIAM Journal on Optimization*, 34(3):2588–2608.

Guille-Escuret, C., Ibrahim, A., Goujaud, B., and Mitliagkas, I. (2022). Gradient Descent Is Optimal Under Lower Restricted Secant Inequality And Upper Error Bound. In *Advances in Neural Information Processing Systems*. Curran Associates Inc.

Hazan, E. (2016). Introduction to Online Convex Optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325.

Hu, B., Seiler, P., and Lessard, L. (2021). Analysis of biased stochastic gradient descent using sequential semidefinite programs. *Mathematical Programming*, 187(1):383–408.

Khaled, A. and Richtárik, P. (2023). Better Theory for SGD in the Nonconvex World. *Transactions on Machine Learning Research*.

Kim, D. and Fessler, J. A. (2021). Optimizing the Efficiency of First-Order Methods for Decreasing the Gradient of Smooth Convex Functions. *Journal of Optimization Theory and Applications*, 188(1):192–219.

Kim, J. (2025). A Proof of the Exact Convergence Rate of Gradient Descent. arXiv preprint arXiv:2412.04427.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Lessard, L., Recht, B., and Packard, A. (2016). Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. *SIAM Journal on Optimization*, 26(1):57–95.

Lewis, A. S., Luke, D. R., and Malick, J. (2009). Local Linear Convergence for Alternating and Averaged Nonconvex Projections. *Foundations of Computational Mathematics*, 9(4):485–513.

Liu, Z. and Zhou, Z. (2023). Revisiting the Last-Iterate Convergence of Stochastic Gradient Methods. In *Proceedings of The Twelfth International Conference on Learning Representations*.

Luner, A. and Grimmer, B. (2024). Performance Estimation for Smooth and Strongly Convex Sets. arXiv preprint arXiv:2410.14811.

MOSEK, A. (2025). MOSEK Optimizer API for Python. Release 11.0.20.

Moucer, C., Taylor, A., and Bach, F. (2023). A Systematic Approach to Lyapunov Analyses of Continuous-Time Models in Convex Optimization. *SIAM Journal on Optimization*, 33(3):1558–1586.

Necoara, I., Richtárik, P., and Patrascu, A. (2019). Randomized Projection Methods for Convex Feasibility: Conditioning and Convergence Rates. *SIAM Journal on Optimization*, 29(4):2814–2852.

Nedić, A. (2010). Random projection algorithms for convex set intersection problems. In *49th IEEE Conference on Decision and Control*, pages 7655–7660.

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

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nguyen, L., Nguyen, P. H., Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. (2018). SGD and Hogwild! Convergence Without the Bounded Gradients Assumption. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3750–3758. PMLR.

Patrascu, A. and Necoara, I. (2018). Nonasymptotic Convergence of Stochastic Proximal Point Methods for Constrained Convex Optimization. *Journal of Machine Learning Research*, 18(198):1–42.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Quarteroni, A., Sacco, R., and Saleri, F. (2007). *Numerical Mathematics*. Number 37 in Texts in Applied Mathematics. Springer, Berlin, second edition.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456. PMLR.

Richtárik, P., Sadiev, A., and Demidovich, Y. (2024). A Unified Theory of Stochastic Proximal Point Methods without Smoothness. arXiv preprint arXiv:2405.15941.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Rotaru, T., Glineur, F., and Patrinos, P. (2022). Tight Convergence Rates of the Gradient Method on Smooth Hypoconvex Functions. arXiv preprint arXiv:2203.00775.

Rotaru, T., Glineur, F., and Patrinos, P. (2024). Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions. arXiv preprint arXiv:2406.17506.

Ryu, E. K., Taylor, A. B., Bergeling, C., and Giselsson, P. (2020). Operator Splitting Performance Estimation: Tight Contraction Factors and Optimal Parameter Selection. *SIAM Journal on Optimization*, 30(3):2251–2271.

Scherer, C. W., Ebenbauer, C., and Holicki, T. (2023). Optimization Algorithm Synthesis Based on Integral Quadratic Constraints: A Tutorial. In *62nd IEEE Conference on Decision and Control*, pages 2995–3002.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112.

Schmidt, M. and Roux, N. L. (2013). Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition. arXiv preprint arXiv:1308.6370.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147. PMLR.

Tan, S. S. Y., Varvitsiotis, A., and Tan, V. Y. F. (2021). Analysis of Optimization Algorithms via Sum-of-Squares. *Journal of Optimization Theory and Applications*, 190(1):56–81.

Taylor, A. and Bach, F. (2019). Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Proceedings of the 32nd Conference on Learning Theory*, pages 2934–2992. PMLR.

Taylor, A., Scoy, B. V., and Lessard, L. (2018a). Lyapunov Functions for First-Order Methods: Tight Automated Convergence Guarantees. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4897–4906. PMLR.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2017a). Exact Worst-Case Performance of First-Order Methods for Composite Convex Optimization. *SIAM Journal on Optimization*, 27(3):1283–1313.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2017b). Smooth Strongly Convex Interpolation and Exact Worst-case Performance of First-order Methods. *Mathematical Programming*, 161(1):307–345.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. (2018b). Exact Worst-Case Convergence Rates of the Proximal Gradient Method for Composite Convex Minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476.

Teboulle, M. and Vaisbourd, Y. (2023). An elementary approach to tight worst case complexity analysis of gradient based methods. *Mathematical Programming*, 201(1):63–96.

Tovmasyan, Z., Malinovsky, G., Condat, L., and Richtárik, P. (2025). Revisiting Stochastic Proximal Point Methods: Generalized Smoothness and Similarity. arXiv preprint arXiv:2502.03401.

Traoré, C., Apidopoulos, V., Salzo, S., and Villa, S. (2024). Variance Reduction Techniques for Stochastic Proximal Point Algorithms. *Journal of Optimization Theory and Applications*, 203(2):1910–1939.

Upadhyaya, M., Banert, S., Taylor, A. B., and Giselsson, P. (2025). Automated tight Lyapunov analysis for first-order methods. *Mathematical Programming*, 209(1):133–170.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR.

Xiao, L. (2009). Dual Averaging Method for Regularized Stochastic Learning and Online Optimization. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Zamani, M., Abbaszadehpeivasti, H., and de Klerk, E. (2024). The exact worst-case convergence rate of the alternating direction method of multipliers. *Mathematical Programming*, 208(1):243–276.

Zamani, M. and Glineur, F. (2023). Exact convergence rate of the last iterate in subgradient methods. arXiv preprint arXiv:2307.11134.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2020). Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *Proceedings of The Eigth International Conference on Learning Representations*.

Zhang, Z., Lee, J. D., Du, S. S., and Chen, Y. (2024). Anytime Acceleration of Gradient Descent. arXiv preprint arXiv:2411.17668.

Zhou, K., Tian, L., So, A. M.-C., and Cheng, J. (2022). Practical Schemes for Finding Near-Stationary Points of Convex Finite-Sums. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 3684–3708. PMLR.