Proceedings of Machine Learning Research vol 99:1–59, 2019 32nd Annual Conference on Learning Theory

Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions

Adrien Taylor ADRIEN.TAYLOR@INRIA.FR
INRIA, Département d'informatique de l'ENS, Ecole normale supérieure, CNRS, PSL Research University, Paris, France

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA, Département d'informatique de l'ENS, Ecole normale supérieure, CNRS, PSL Research University, Paris, France

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We provide a novel computer-assisted technique for systematically analyzing first-order methods for optimization. In contrast with previous works, the approach is particularly suited for handling sublinear convergence rates and stochastic oracles. The technique relies on semidefinite programming and potential functions. It allows simultaneously obtaining worst-case guarantees on the behavior of those algorithms, and assisting in choosing appropriate parameters for tuning their worst-case performances. The technique also benefits from comfortable *tightness guarantees*, meaning that unsatisfactory results can be improved only by changing the setting. We use the approach for analyzing deterministic and stochastic first-order methods under different assumptions on the nature of the stochastic noise. Among others, we treat unstructured noise with bounded variance, different noise models arising in over-parametrized expectation minimization problems, and randomized block-coordinate descent schemes.

1. Introduction

In this work, we study methods for solving convex (stochastic) minimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x), \tag{Opt}$$

with $f \in \mathcal{F}$ some class of convex, proper and closed functions. To perform the minimization, we are given access to an approximate first-order oracle $G(x;i) \approx f'(x)$, where i is some random variable uniformly sampled in an index set I. This includes unbiased stochastic oracles satisfying $\mathbb{E}_i G(x;i) = f'(x)$, but also biased oracles used in block-coordinate methods $G(x;i) = \nabla_i f(x)$ (directional derivative along the i^{th} block of coordinates).

We present a generic approach, based on potential functions, for analyzing and designing first-order methods in the case where I is a finite set—two such problems are the *empirical risk minimization* setting where $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ and $G(x;i) = f_i'(x)$, and the block-coordinate setting. Even though most proofs presented in the sequel turned out not to depend on the cardinality of I, and are therefore valid for expectation minimization problems $f(x) = \mathbb{E}_i f(x;i)$, cardinality can play a major role in specific settings (e.g., finite sums or coordinate descent). Therefore, we do not explicitly look for results that are independent of it, but rather note that it naturally does not intervene in, e.g., analyses of stochastic gradient-based methods that we propose in this paper.

1.1. Preliminaries

A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called L-smooth if its gradient satisfies a Lipschitz condition with parameter $L > 0$:

$$\|f'(x) - f'(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d,$$

where $\langle .; . \rangle$ denotes the Euclidean inner product and $\|.\|$ is the induced norm. We denote by $\mathcal{F}_L(\mathbb{R}^d)$ the class of L-smooth convex functions over $\mathbb{R}^d$ and by $\mathcal{F}_L$ the class of smooth convex functions where d is left unspecified. In addition, we denote by $x_\star$ some optimal solution to (Opt), and by $f_\star := f(x_\star)$ the optimal value. Those assumptions are standard in the optimization literature (Polyak, 1987; Nesterov, 2004).

1.2. Contributions

The main contribution of this work is to propose a framework for constructing potentials for first-order stochastic algorithms; in contrast with previous related works on the topic, the technique is specialized for establishing sublinear convergence rates. The methodology benefits from an advantageous tightness property, meaning that it fails only when it is impossible to prove the desired result using potential functions with the chosen structure. The framework allows dealing with, among others, all stochastic settings presented in Table 1. We use the methodology for designing novel analyses of SGD and averaging schemes in different stochastic optimization settings. Based on the methodology for constructing potentials, we propose a complementary automatic parameter selection technique in App. C, whose main idea is roughly to optimize the algorithmic parameters while designing the potentials.

$\mathcal{F}$	Noise model	$\mathbb{E}_i\|G(x;i)\|^2 \leq$	Note	Sections
$f \in \mathcal{F}_L$	$G(x;i) = f'(x)$	—	No noise	3.2.1, D
$f \in \mathcal{F}_L$	$G(x;i)$	$\sigma_\star^2 + 2\rho_1 L(f(x) - f_\star) + \rho_2\|f'(x)\|^2$ for all $x \in \mathbb{R}^d$	Unified variance model	3.2.2, E, G
$f_i \in \mathcal{F}_L$	$G(x;i) = f_i'(x)$	$\sigma_\star^2$ for some $x = x_\star$	Variance at $x_\star$	4, F, H
$f_i \in \mathcal{F}_L$	$G(x;i) = f_i'(x)$	—	Finite sums	(not presented)
$f_i \in \mathcal{F}_L$	$G(x;i) = \nabla_i f(x)$	—	Block-coordinate	I

Table 1: Stochastic settings summary: non-exhaustive list of assumptions on the classes of functions $\mathcal{F}$ and the nature of the noisy oracle $G(x;i)$ that can be directly embedded within the framework. Further examples are discussed in Sec. 5.

1.3. Prior works

In what follows, we take a *worst-case* point of view, that is standard in the optimization and machine learning communities, as in the original works (Nemirovski and Yudin, 1983; Polyak, 1987; Nesterov, 2004).

Stochastic first-order methods. Stochastic approximation algorithms date back to the works of Robbins and Monro (1951), and numerous analyses and improvements can be found in the literature (see e.g., Bottou et al. (2018) and the references therein). Among others, averaging plays a crucial role for improving their convergence guarantees (Ruppert, 1988; Polyak and Juditsky, 1992).

Stochastic methods are usually analyzed using a uniformly bounded variance assumption (i.e., $\mathbb{E}_i\|G(x;i) - f'(x)\|^2 \leq \sigma^2$ for all $x \in \mathbb{R}^d$), or bounded gradients. This intrinsically has limited applicability ranges (e.g., it does not even hold for quadratic minimization), although theoretical guarantees for stochastic methods involving momentum were out of reach so far without such assumptions (Hu et al., 2009; Xiao, 2010; Devolder, 2011; Lan, 2012; Cohen et al., 2018). A better understanding of stochastic gradient methods (in particular, those involving momentum) can therefore only be achieved by studying alternatives to standard assumptions. In particular, let us mention the non-asymptotic analyses of Bach and Moulines (2011) (not relying on the uniformly bounded assumption), and Schmidt and Le Roux (2013); Ma et al. (2018); Vaswani et al. (2019) (strong and weak growth conditions), and numerous works on stochastic methods for quadratic minimization, see e.g., Bach and Moulines (2013); Dieuleveut et al. (2017); Jain et al. (2018a,b).

Potential functions for first-order methods. Potential functions have been used a lot for studying convergence properties of first-order methods. This kind of analyses is typically natural for obtaining linear convergence results—potentials are then often being referred to as Lyapunov functions (Lyapunov and Fuller, 1992), as in the analyses of dynamical systems—, but is typically also used for certifying sublinear convergence rates, as discussed in Sec. 2. As being nicely reviewed by Bansal and Gupta (2017), the use of potential functions is not new in the optimization literature, and is closely related to the machinery of *estimate sequences* (Nesterov, 2004, 2018; Wilson et al., 2016). Successful uses of such techniques include the original developments underlying accelerated gradient (Nesterov, 1983, Theorem 1) and FISTA (Beck and Teboulle, 2009, Lemma 4.1).

Computer-assisted analyses of first-order methods. Recently, linear matrix inequalities (LMI) and semidefinite programming (SDP) (see e.g., Vandenberghe and Boyd (1996)) techniques were used for automatically generating worst-case guarantees for first-order methods. This trend started with *performance estimation* as initiated by Drori and Teboulle (2014) and was taken further in different directions: for designing optimal methods (Drori and Teboulle, 2014, 2016; Kim and Fessler, 2016, 2018c), lower bounds (Drori, 2017), or to be featured with automated tightness guarantees and broader range of applications (Taylor et al., 2017a,c). A competing strategy, inspired on the one hand by performance estimation and on the other one by control theory, was developed by Lessard et al. (2016). This technique is based on *integral quadratic constraints* (IQC) and was initially specialized for obtaining linear convergence rates. This work adapts the performance estimation approach for using potential functions, easing the development of proofs in settings involving sublinear convergence rates. Detailed relations to works in this research stream are provided in App. B.

1.4. Organization

The flow of this work is as follows. First, we summarize the overall methodology and recall the general principle behind potential-based proofs in Sec. 2. After that, the procedure for designing potential functions is presented in Sec. 3; in particular, the methodology is illustrated on gradient descent and a few stochastic variants in the same section. Then, we present simple results obtained with the analysis technique in different stochastic optimization settings (a few samples from Table 1) in Sec. 4 and in the appendix. Most technical tools are presented in appendix, including an automatic parameter selection technique, and the application to accelerated first-order methods, to stochastic optimization under weak growth conditions, and to coordinate descent. The organization and content of the appendix is summarized in App. A.

2. Potential functions for a restricted class of first-order methods

Let us restrict ourselves to a specific class of methods encapsulating SGD, along with possibly averaging and momentum. This restriction is made for readability purposes only. We consider the following class of (stochastic) first-order methods:

$$\begin{aligned} y_{k+1} &= y_k + \alpha_k\,(x_k - y_k) + \alpha'_k\,(z_k - y_k), \\ x_{k+1}^{(i_k)} &= y_{k+1} + \beta_k\,(x_k - y_{k+1}) + \beta'_k\,(z_k - y_{k+1}) - \delta_k G(y_{k+1}; i_k), \\ z_{k+1}^{(i_k)} &= y_{k+1} + \gamma_k\,(x_k - y_{k+1}) + \gamma'_k\,(z_k - y_{k+1}) - \epsilon_k G(y_{k+1}; i_k), \end{aligned} \tag{SFO}$$

where the superscript (i_k) corresponds to the sampled random variable that was used for performing iteration k. In what follows, we provide a generic approach to study its worst-case properties. For now, let us ask the question *how can we prove such an algorithm work?* A possible methodology for showing convergence consists in exhibiting a potential function, also often referred to as a *Lyapunov function*. For example, when $x_{k+1} = x_k - \frac{1}{L} f'(x_k)$ (i.e., gradient descent), it is possible to show (see, e.g., Bansal and Gupta (2017)) that for all $f \in \mathcal{F}_L$ and $k \geq 0$ the inequality $\phi_{k+1}^f(x_{k+1}) \leq \phi_k^f(x_k)$ holds with

$$\phi_k^f(x_k) = k(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2,$$

leading to $N(f(x_N) - f_\star) \leq \phi_N^f \leq \phi_{N-1}^f \leq \ldots \leq \phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2$. Therefore, $f(x_N) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{2N}$. This proof relies on two key ideas: (i) forget how x_k was generated and study only one iteration at a time, and (ii) choose an appropriate sequence of potentials. Such proofs are philosophically simple, but it is generally unclear how to chose such potentials. Choosing an appropriate sequence $\{\phi_k^f\}$ usually requires a lot of intuitions and potentially tedious investigations. Such proofs may therefore be seen as reserved to experts, and the purpose of this work is to alleviate as much as possible this burden, by proposing a systematic way of designing and verifying potentials. All proofs developed hereafter follow the same principles, and reduce to proving inequalities of type

$$\mathbb{E}_{i_k} \phi_{k+1}^f(y_{k+1}, x_{k+1}^{(i_k)}, z_{k+1}^{(i_k)}) \leq \phi_k^f(y_k, x_k, z_k) + e_k, \tag{Pot}$$

for all $f \in \mathcal{F}$ and all x_k, y_k, z_k used to generate $y_{k+1}, x_{k+1}^{(i_k)}, z_{k+1}^{(i_k)}$ with the method of interest. The term e_k is typically used for encapsulating the *variance* of stochastic algorithms. As before, a recursive use of this inequality allows obtaining

$$\mathbb{E}\phi_N^f(y_N, x_N, z_N) \leq \phi_0^f(y_0, x_0, z_0) + \sum_{k=0}^{N-1} e_k,$$

and the game consists in choosing appropriate sequences for ϕ_k^f and e_k. The expectation $\mathbb{E}$ is taken over all sequences of indices $(i_1, i_2, \ldots, i_N)$ with $i_k \in I$. Such convergence results in expectations can then typically be converted to almost sure convergence using, e.g., Robbins-Siegmund supermartingale theorem (Robbins and Siegmund, 1971; Duflo, 1997).

3. Design methodology for potential functions

We propose a systematic way to verify that a given tuple $(\phi^f_{k+1}, \phi^f_k, e_k)$ satisfies inequality (Pot). First of all, it is clear that the set of such acceptable tuples is convex. Even more, when ϕ^f_k, ϕ^f_{k+1} are both *quadratic* functions of the first-order information $G(.;i)$ and the coordinates x and *linear* functions of the function values $f(.)$, then verifying that the tuple satisfies (Pot) can equivalently be formulated as a linear matrix inequality (LMI). This section aims at providing strategies for finding sequences of potentials $\{\phi^f_k\}_k$ based on a few examples and on Proposition 1 that follows.

3.1. Verifying a potential

The main tool we use for designing potentials is summarized through the following proposition. Even though its proof may appear as straightforward (we do not provide it), the main component in our strategies is the possibility of efficiently formulating (1) (for verifying a potential) using LMIs. In the sequel, we present the methodology in a high-level form and delay all LMI formulations to appendix for readability purposes.

Proposition 1 *Let $\mathcal{F}$ be a class of functions, I be an index set, $G(x;i)$ (with $i \in I$) be satisfying one of the noise model of Table 1, (SFO) be the class of algorithms under consideration, and a given tuple $(\phi^f_{k+1}, \phi^f_k, e_k)$. We have*

$$\mathbb{E}_{i_k}\phi^f_{k+1}(y_{k+1}, x^{(i_k)}_{k+1}, z^{(i_k)}_{k+1}) \leq \phi^f_k(y_k, x_k, z_k) + e_k \quad (\mathbb{E}_{i_k} \text{ denotes the expectation over } i_k \in I)$$

for all $d \in \mathbb{N}$, $f \in \mathcal{F}(\mathbb{R}^d)$ and all $(y_k, x_k, z_k) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ if and only if

$$\begin{aligned} 0 \geq \sup_{\substack{d, f, y_k, x_k, z_k \\ \{G(x;i)\}_{i\in I}}} \ & \mathbb{E}_{i_k}\phi^f_{k+1}(y_{k+1}, x^{(i_k)}_{k+1}, z^{(i_k)}_{k+1}) - \phi^f_k(y_k, x_k, z_k) - e_k \\ \text{s.t. } & (y_{k+1}, x^{(i_k)}_{k+1}, z^{(i_k)}_{k+1}) \text{ generated by (SFO) from } (y_k, x_k, z_k) \\ & \{G(x;i)\}_{i\in I} \text{ compatible with } f \text{ and the noise model for all } x \in \operatorname{dom} f \\ & f \in \mathcal{F}(\mathbb{R}^d) \text{ and } f'(x_\star) = 0. \end{aligned} \tag{1}$$

Remark 2 *In many standard settings (including noise models presented in Table 1 and the use of quadratic potentials—see examples below), the decision problem (1) can be reformulated as a LMI. This lossless reformation into a suitable LMI directly follows from the derivations presented by Taylor et al. (2017a, Section 2) for the deterministic setting. Those reformulations can be extended in a straightforward manner to all settings presented in Table 1, so we only present them in appendix for the different examples treated hereafter. As an introductory example, one can find the 3×3 LMI reformulation for gradient descent in App. D.4; the other examples are summarized in App. A.*

When the class of functions $\mathcal{F}$, the method and the noise model are clear from the context, we abusively denote the set of tuples $(\phi^f_{k+1}, \phi^f_k, e_k)$ that satisfies (1) by $\mathcal{V}_k$.

3.2. Strategies for designing sequences of potentials

Taking advantage of Proposition 1, we propose a few strategies for choosing sequences of potential functions $\{\phi^f_k\}_k$ based on two examples. For simplicity, we start in a deterministic setting. The codes used for generating the numerics below are provided in Sec. 5.

3.2.1. EXAMPLE I: GRADIENT DESCENT

Say we want to bound $\|f'(x_N)\|^2$, where x_N is the iterate obtained after performing N iterations of gradient descent $x_{k+1} = x_k - \frac{1}{L} f'(x_k)$ where $f \in \mathcal{F}_L$. Let us choose the family of potentials:

$$\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k \left(f(x_k) - f_\star \right), \tag{2}$$

parametrized by $\{(a_k, b_k, c_k, d_k)\}_k$. The motivation for such a shape is simply to allow all the information available at x_k to be used, and the Kronecker product with the identity "$\cdot \otimes I_d$" corresponds to requiring the potential function to be *isotropic* in the spaces of coordinates and gradients; in other words, the potentials can be written

$$\phi_k^f = a_k \|x_k - x_\star\|^2 + b_k \|f'(x_k)\|^2 + 2c_k \langle f'(x_k); x_k - x_\star \rangle + d_k \left(f(x_k) - f_\star \right).$$

Now let us arbitrarily choose $\phi_0^f = L^2 \|x_0 - x_\star\|^2$ and $\phi_N^f = b_N \|f'(x_N)\|^2$. The main motivation underlying this choice is that this structure may result in $\|f'(x_N)\|^2 \leq \frac{L^2 \|x_0 - x_\star\|^2}{b_N}$ using similar arguments as in Sec. 2. For choosing an appropriate sequence $\{(a_k, b_k, c_k, d_k)\}_k$, we propose to solve the following problem:

$$b_N^{(\text{opt})} = \max_{\phi_1^f, \ldots, \phi_{N-1}^f, b_N} b_N \text{ subject to } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}. \tag{3}$$

This problem can be formulated via semidefinite programming (SDP) with N LMIs of size 3×3 (see App. D), and allows recovering the largest possible valid b_N given structure (2), ϕ_0^f and ϕ_N^f. Based on numerical inspection (details hereafter), one can find the following valid sequence of potentials:

$$\phi_k^f(x_k) = (2k+1)L(f(x_k) - f_\star) + k(k+2)\|f'(x_k)\|^2 + L^2 \|x_k - x_\star\|^2,$$

which directly allows to prove both $f(x_k) - f_\star = O(k^{-1})$ and $\|f'(x_k)\|^2 = O(k^{-2})$ simultaneously. Although simple, this is apparently the first time that convergence in gradient norm is proved using standard techniques that are usually used only for function values: the typical technique for obtaining this rate of convergence for gradient norm was presented by Nesterov (2012b). Let us briefly describe how such a potential can be obtained (the steps can be followed on Figure 1).

1. We started by numerically solving (3) for a few values of N using standard packages (Löfberg, 2004; Mosek, 2010). Approximate numerical results are provided below:

$N =$	1	2	3	4	5	...	100
$b_N^{(\text{opt})} =$	4	9	16	25	36	...	10201

   The solution to (3) numerically appeared to match $b_N = (N+1)^2$. For completeness, note that the relative inaccuracy $b_N^{(\text{opt})}/(N+1)^2 - 1$ observed on the numerical solutions appeared to be an increasing function of N ranging from 10^{-8} for small values of N to 10^{-5} for $N = 100$. Readers interested in transforming those numerics into formal proofs may proceed with steps 2–4.

2. Observe schedules of $\{(a_k, b_k, c_k, d_k)\}_k$ that are numerically obtained by solving (3). The result for $N = 100$ is given in Figure 1 (plain brown).

3. Try to simplify $\{\phi_k^f\}_k$ without loosing too much on $b_N^{(\text{opt})}$ (i.e., keep $b_N^{(\text{opt})}$ large enough). As examples, the first numerical schedules motivated trying to solve (3) under the additional constraint $d_k = (2k+1)L$ (Figure 1, dashed red), then we additionally tried $c_k = 0$ and $a_k = L^2$ (Figure 1, dashed blue). Those two simplifications turned out to be successful; an example of unsuccessful one is obtained by constraining $d_k = 0$ in (3) (Figure 1, dashed purple), which does not allow achieving a large enough value of b_N for proving $O(k^{-2})$ convergence in $\|f'(x_k)\|^2$.

4. Using the numerical inspirations, study one step of the method, i.e., find a feasible point to (1).

Theorem 3 *Let $x_k \in \mathbb{R}^d$, $f \in \mathcal{F}_L$, and $x_{k+1} = x_k - \frac{1}{L}f'(x_k)$. The inequality $\phi_{k+1}^f(x_{k+1}) \leq \phi_k^f(x_k)$ holds with*

$$\phi_k^f = d_k(f(x_k) - f_\star) + b_k\|f'(x_k)\|^2 + L^2\|x_k - x_\star\|^2,$$

and all values $b_k, d_k \geq 0$ such that $b_{k+1} = 2 + \frac{d_k}{L} + b_k$ and $d_{k+1} = 2L + d_k$.

The proof is relatively simple relies on finding a sequence of feasible points to (1). The proof is delayed to App. D; in particular, the choice $d_k = (2k+1)L$ and $b_k = k(k+2)$ is valid for using Theorem 3, and allows recovering the previous claim. A corresponding result for the proximal gradient method is presented in App. D.2, and similar results for accelerated variants are given in Sec. D.3.

Remark 4 *Instead of using $\{(a_k, b_k, c_k, d_k)\}_k$ as variables, one could explicitly require each of those to be a polynomial in k, and use the coefficients of those polynomials as variables in (3). Our choice of using $\{(a_k, b_k, c_k, d_k)\}_k$ allows knowing in advance that the optimal b_N value obtained by solving (3) is the best value of b_N that can be certified given ϕ_0^f and the structure of ϕ_k^f. It could nevertheless be more practical to use the polynomials' coefficients as variables in certain situations.*

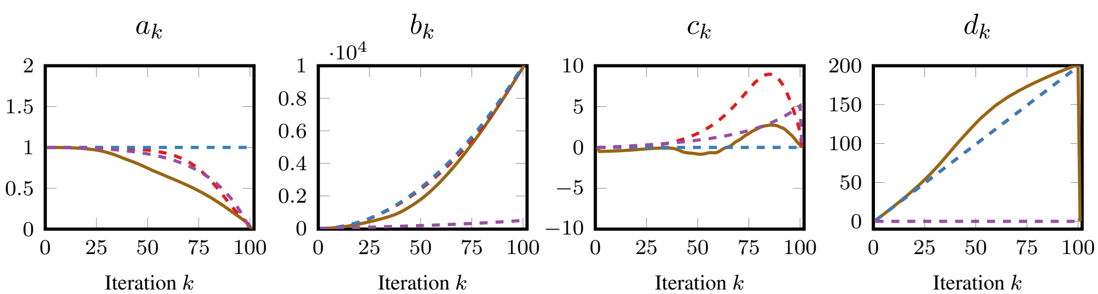

Figure 1: Numerical solution to (3) for $N = 100$ and $L = 1$ (plain brown), forced $d_k = (2k+1)L$ (dashed red), forced $d_k = (2k+1)L$, $c_k = 0$ and $a_k = L^2$ (dashed blue) and forced $d_k = 0$ (dashed purple). Total time: ~ 35 sec. on single core of Intel Core i7 1.8GHz CPU.

3.2.2. EXAMPLE II: STOCHASTIC SMOOTH CONVEX MINIMIZATION, BOUNDED VARIANCE

For studying stochastic methods, we heavily rely on the finite support assumption of the random variable $i_k \in I$. This is crucial, as our main tool is a reformulation of (1) into a LMI where we perform an averaging over the $n = |I|$ possible scenarios.

Stochastic methods are commonly studied using a uniformly bounded variance assumption $\mathbb{E}_i\|G(x;i) - f'(x)\|^2$ over x (Hu et al., 2009; Xiao, 2010; Lan, 2012; Devolder, 2011). This assumption is quite restrictive, but analyses not relying on it often appear to be much more challenging and sometimes even out of reach so far. Nevertheless, this restrictive setting is used for the examples of this section. Other setups are explored in Sec. 4 and in appendix.

In the previous section, the use of Proposition 1 was exemplified for designing a potential function for vanilla gradient method. In the following lines, we provide two alternate ways of choosing sequences of potentials that can be used for stochastic first-order methods, the main additional difficulty being the appearance of a *variance* term e_k in the inequality (Pot). Final consequences of the results of this section are depicted in Table 2 for when (decreasing) step-sizes of the form $\delta_k = (L(1+k)^\alpha)^{-1}$ are used in the stochastic algorithms. For SGD with and without averaging we discuss the differences with Bach and Moulines (2011) below.

Stochastic gradient. Let us apply the methodology to the SGD iteration

$$x_{k+1}^{(i_k)} = x_k - \delta_k G(x_k; i_k),$$

where we choose the following family of potentials:

$$\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top \left[\begin{pmatrix} a_k & c_k \\ c_k & b_k \end{pmatrix} \otimes I_d \right] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k \left(f(x_k) - f_\star \right).$$

For choosing the sequence $\{(a_k, b_k, c_k, d_k, e_k)\}_k$, we arbitrarily start with $\phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2$ and $\phi_N^f = d_N(f(x_N) - f_\star)$ as this may result in a guarantee of the form $\mathbb{E}[f(x_N) - f_\star] \le \frac{L\,\|x_0 - x_\star\|^2}{2d_N} + \sigma^2 \frac{\sum_{k=0}^{N-1} e_k}{d_N}$. We proceed with a two-stage strategy:

$$d_N^{(\text{opt})} = \max_{\substack{\phi_1^f,\ldots,\phi_{N-1}^f, d_N \\ e_0,\ldots,e_{N-1}}} d_N \text{ s.t. } (\phi_0^f, \phi_1^f, e_0) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f, e_{N-1}) \in \mathcal{V}_{N-1},$$

$$\min_{\substack{\phi_1^f,\ldots,\phi_{N-1}^f, \\ d_N, e_0,\ldots,e_{N-1}}} \sum_{k=1}^N e_k \text{ s.t. } d_N = d_N^{(\text{opt})},\ (\phi_0^f, \phi_1^f, e_0) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f, e_{N-1}) \in \mathcal{V}_{N-1}, \quad (4)$$

where the sequence is chosen as the optimal solution to (4), which is formulated using N LMIs of sizes $(2n+1)\times(2n+1)$. Alternatively, one can choose two weights R^2 and σ^2 and solve

$$\min_{V_1,\ldots,V_{N-1}, d_N} \frac{R^2}{d_N} + \frac{\sigma^2}{d_N} \sum_{k=1}^N e_k \text{ subject to } (\phi_0^f, \phi_1^f, e_0) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f, e_{N-1}) \in \mathcal{V}_{N-1},$$

which is also convex. As in the case of the gradient method, let us briefly describe the steps.

1. Solve (4) for the fixed step-size policy $\delta_k = \frac{1}{L}$ and for a few values of N (number of iterations) and n (cardinality of $\{G(x;1), \ldots, G(x;n)\}$). Approximate numerical results are as follow:

$n =$	2							10						
$N =$	1	2	3	4	5	$\ldots$	100	1	2	3	4	5	$\ldots$	100
$d_N^{(\text{opt})} =$	3	5	7	9	11	$\ldots$	201	3	5	7	9	11	$\ldots$	201
$e_N^{(\text{opt})} =$	1.5	2.5	3.5	4.5	5.5	$\ldots$	100.5	1.5	2.5	3.5	4.5	5.5	$\ldots$	100.5

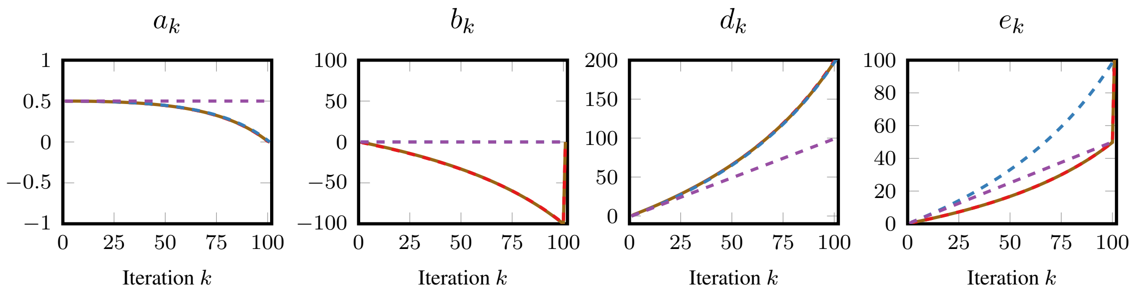

Figure 2: Numerical solution to (4) for $n = 2$, $N = 100$ and $L = 1$ (plain brown), forced $c_k = 0$ (dashed red), forced $c_k = 0$, $b_k = 0$ (dashed blue) and forced $c_k = 0$, $b_k = 0$ and $a_k = \frac{L}{2}$ (dashed purple). Total time: ~ 60 sec. on single core of Intel Core i7 1.8GHz CPU.

2. Observe schedules $\{(a_k, b_k, c_k, d_k, e_k)\}_k$ that are numerically obtained by solving (4). The result for $N = 100$ and $n = 2$ is given in Figure 2 (plain brown).
3. Simplify $\{\phi_k^f\}_k$ as much as possible without loosing too much (i.e., keep d_N large and e_N small). For example, enforcing $c_k = 0$ in (4) (Figure 2, dashed red), enforcing $b_k = c_k = 0$ in (4) (Figure 2, dashed blue) and finally enforcing $b_k = c_k = 0$ and $a_k = \frac{L}{2}$ in (4) (Figure 2, dashed purple).
4. Using numerical inspiration, study one step of the method, i.e., find a feasible point to (1).

Theorem 5 *Let $x_k \in \mathbb{R}^d$, $f \in \mathcal{F}_L$ and $x_{k+1}^{(i_k)} = x_k - \delta_k G(x_k; i_k)$ with $\mathbb{E}_{i_k} \|G(x_k; i_k)\|^2 \le \sigma^2 + \|f'(x_k)\|^2$. The inequality $\mathbb{E}_{i_k}[\phi_{k+1}^f(x_{k+1}^{(i_k)})] \le \phi_k^f(x_k) + e_k\sigma^2$ holds with*

$$\phi_k^f(x_k) = d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2,$$

for all values $d_k, \delta_k \ge 0$ such that $d_{k+1} = d_k + \delta_k L$, $e_k = \frac{\delta_k^2 L}{2}(1 + d_{k+1})$, and $\delta_k d_{k+1} \ge e_k$ (i.e., when step-size δ_k is small enough; in particular, the choice $0 \le \delta_k \le \frac{1}{L}$ is valid).

The proof is provided in App. E. The choice $\delta_k = \frac{1}{L}$, $d_k = k$ and $e_k = \frac{1}{L}\left(\frac{k}{2} + 1\right)$ leads to

$$(k+1)\mathbb{E}_{i_k}[f(x_{k+1}^{(i_k)}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}[\|x_{k+1}^{(i_k)} - x_\star\|^2] \le k\,[f(x_k) - f_\star] + \tfrac{L}{2}[\|x_k - x_\star\|^2] + (\tfrac{k}{2} + 1)\tfrac{\sigma^2}{L},$$

the results are shown on Figure 2, whose dashed purple curves correspond to

$$N\mathbb{E}[f(x_N) - f_\star] + \tfrac{L}{2}\mathbb{E}[\|x_N - x_\star\|^2] \le \tfrac{L}{2}\|x_0 - x_\star\|^2 + \tfrac{N}{4}(N+3)\tfrac{\sigma^2}{L}.$$

The choice $\delta_k = (L(1+k)^\alpha)^{-1}$ leads to the results provided in Table 2 (details in App. E.5). Compared to Bach and Moulines (2011) —which deals with the slightly different setting $f_i \in \mathcal{F}_L$ and bounded variance at $x_\star$: $\mathbb{E}_i\|f_i'(x_\star)\|^2 \le \sigma_\star^2$— we obtained bounds that are valid for all values of α, compared to $\alpha \in (1/2, 1)$, and the same optimal value $\alpha = 2/3$. Although the scope of this new result is more limited through the assumptions, the proof is considerably simpler.

Stochastic gradient with averaging. A standard way to improve theoretical guarantees of stochastic gradient methods is to embed them within an averaging process, as in Ruppert (1988); Polyak and Juditsky (1992); Bach and Moulines (2011), leading to the following iterative process:

$$\begin{aligned} x_{k+1}^{(i_k)} &= x_k - \delta_k G(x_k; i_k), \\ z_{k+1}^{(i_k)} &= \tfrac{1}{k+1}\, x_{k+1}^{(i_k)} + \tfrac{k}{k+1}\, z_k. \end{aligned}$$

The following potential was found using a procedure similar to the ones presented for gradient and stochastic gradient methods. In contrast with previous examples, this bound can only be propagated for decreasing step-sizes (i.e., $\delta_{k+1} \leq \delta_k$). In particular, the choice $\delta_k = (L(1+k)^\alpha)^{-1}$ leads to the results provided in Table 2 (details in App. E.5). Compared to Bach and Moulines (2011), as for SGD, we obtain the same optimal value $\alpha = 1/2$ for averaging, and bounds that are valid for all values of α (compared to only $\alpha \in (1/2, 1)$ before). The proof is delayed to App. E.2, and relied on using N LMIs of sizes $(3+2n) \times (3+2n)$.

Theorem 6 *Consider the following iterative scheme*

$$\begin{aligned} x_{k+1}^{(i_k)} &= x_k - \delta_k G(x_k; i_k), \\ z_{k+1}^{(i_k)} &= \tfrac{1}{d_k+1}\, x_{k+1}^{(i_k)} + \tfrac{d_k}{d_k+1}\, z_k, \end{aligned}$$

for some $d_k, \delta_k \geq 0$. *Assuming* $f \in \mathcal{F}_L$ *and* $\mathbb{E}_{i_k}\|G(x_k; i_k)\|^2 \leq \sigma^2 + \|f'(x_k)\|^2$, *the following inequality holds*

$$\delta_k d_{k+1} L \mathbb{E}_{i_k}[f(z_{k+1}^{(i_k)}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq \delta_k d_k L(f(z_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k \sigma^2,$$

with $d_{k+1} = d_k + 1$, $e_k = \frac{\delta_k^2}{2}\frac{L(1+d_k+L\delta_k)}{1+d_k}$ *and* $\delta_k \leq \frac{1+\sqrt{5}}{2L}$.

Stochastic gradient with primal averaging. Inspired by the numerical step-size selection tool provided in App. C, we propose an alternative to averaging—sometimes referred to as *primal averaging* (Tao et al., 2018)—corresponding to evaluating the stochastic gradient at the averaged iterate, in the particular case of a fixed-step policy $\delta_k = \frac{1}{L}$:

$$\begin{aligned} y_{k+1} &= \tfrac{k}{k+1}\, y_k + \tfrac{1}{k+1}\, x_k, \\ x_{k+1}^{(i_k)} &= x_k - \tfrac{1}{L} G(y_{k+1}; i_k). \end{aligned}$$

The following theorem was obtained through the use of N LMIs of sizes $(3+n) \times (3+n)$.

Theorem 7 *Consider the following iterative scheme*

$$\begin{aligned} y_{k+1} &= \tfrac{d_k}{d_k+\delta_k L} y_k + \tfrac{\delta_k L}{d_k+\delta_k L} x_k \\ x_{k+1}^{(i_k)} &= x_k - \delta_k G(y_{k+1}; i_k) \end{aligned}$$

for some $d_k, \delta_k \geq 0$. *Assuming* $f \in \mathcal{F}_L$ *and* $\mathbb{E}_{i_k}\|G(y_{k+1}; i_k)\|^2 \leq \sigma^2 + \|f'(y_{k+1})\|^2$, *the following inequality holds*

$$d_{k+1}(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k\sigma^2,$$

with $d_{k+1} = d_k + \delta_k L$ *and* $e_k = \frac{L\delta_k^2}{2}$ *when* $\delta_k \leq \frac{1}{L}$.

The proof is provided in App. E.3. In particular, the choice $\delta_k = (L(1+k)^\alpha)^{-1}$ leads to the results provided in Table 2 (details in App. E.5), and an alternate version where we always evaluate the stochastic gradient at the averaged iterate for any step-size policy δ_k is provided in App. E.4.

		$\|x_0 - x_\star\|^2$	σ^2	Optimal α
Vanilla SGD	$\mathbb{E}f(x_k) - f_\star \leq$	$O(k^{\alpha-1})$	$O(k^{1-2\alpha})$	$2/3$
Polyak-Ruppert averaging	$\mathbb{E}f(z_k) - f_\star \leq$	$O(k^{\alpha-1})$	$O(k^{-\alpha})$	$1/2$
Primal averaging	$\mathbb{E}f(y_k) - f_\star \leq$	$O(k^{\alpha-1})$	$O(k^{-\alpha})$	$1/2$

Table 2: Asymptotic rates for SGD, Polyak-Ruppert averaging, and primal averaging under uniformly bounded variance $\mathbb{E}_i\|G(x;i)\|^2 \leq \sigma^2 + \|f'(x)\|^2$ and step-sizes $\delta_k \sim k^{-\alpha}$. A factor $\log k$ was neglected for optimal α's. Details in App. E.5, and momentum in App. E.6.

4. Application to stochastic convex minimization for over-parameterized models

In many modern machine learning settings, models are over-parametrized and allow interpolating the data. This is discussed by Schmidt and Le Roux (2013); Ma et al. (2018); Vaswani et al. (2019) and sometimes analyzed through the use of *growth conditions* (which we discuss in App. G). Alternatively, we model this scenario through the setup

$$\min_{x\in\mathbb{R}^d}\{f(x) \equiv \mathbb{E}_i f_i(x)\},$$

where we assume $f_i \in \mathcal{F}_L$ and that there exists an optimal point $x_\star$ such that $f_i'(x_\star) = 0$ for all $i \in I$. Using the previous methodology, the best worst-case guarantees we could reach for vanilla SGD (without averaging) was achieved by using a decreasing step-size policy, resulting only in a disappointing $O(k^{-1/2})$ guarantee. On the other hand, the following method (inspired by our step-size selection tool in appendix) turned out to be considerably simpler to analyze, while enjoying better worst-case guarantees. As in the previous section, the main idea is to evaluate the stochastic gradient at the averaged iterate instead of the last one (primal averaging). The proof is delayed to App. F, and relied on using our step-size selection technique. The computational cost (of designing N iterations of this method) was that of solving N LMIs of sizes $(3+3n+n^2)\times(3+3n+n^2)$.

Theorem 8 *Let $x_k \in \mathbb{R}^d$, $f_i \in \mathcal{F}_L$ and an optimal point $x_\star$ such that $f_i'(x_\star) = 0$ for all $i \in I$. Then the iterative scheme*

$$\begin{aligned} y_{k+1} &= \tfrac{d_k}{d_k+\delta_k L} y_k + \tfrac{\delta_k L}{d_k+\delta_k L} x_k,\\ x_{k+1}^{(i_k)} &= x_k - \delta_k f_{i_k}'(y_{k+1}), \end{aligned}$$

satisfies

$$d_{k+1}(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2,$$

for all values of $d_k, \delta_k \geq 0$ and

$$d_{k+1} = \begin{cases} d_k + \delta_k L & \text{if } \delta_k \leq \frac{1}{L},\\ d_k + 2\delta_k L - \delta_k^2 L^2 & \text{otherwise.}\end{cases}$$

Using $\delta_k = \frac{1}{L}$ (choice that maximizes d_{k+1}) and $d_0 = 0$ leads to $d_k = k$ and to the algorithm

$$\begin{aligned} y_{k+1} &= \tfrac{k}{k+1} y_k + \tfrac{1}{k+1} x_k,\\ x_{k+1}^{(i_k)} &= x_k - \tfrac{1}{L} f_{i_k}'(y_{k+1}), \end{aligned}$$

for which the bound $\mathbb{E}f(y_k) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{2k}$ holds for all $k \geq 0$.

5. Conclusion

In this work, we showed how to adapt the performance estimation approach to obtain potential-based proofs. Given a first-order methods and a class of (quadratic) potential functions and predefined numbers of iterations, the methodology allows obtaining the *best* worst-case guarantees that can be obtained by a potential-based approach with a given structure—choosing an appropriate structure is therefore the critical point. Hence, if the methodology fails to provide the user a satisfactory worst-case bound, the only possible alternatives for improving the results are to either (i) enrich the class of potential functions, or (ii) add assumptions, or change the problem class. This methodology has the advantage of quickly allowing to assess feasibility of new ideas and to develop simple algorithms for new settings.

Although provided only for unconstrained minimization, the methodology allows dealing with many other settings such as projection, linear optimization operators (a.k.a., Frank-Wolfe or conditional gradient oracles), proximal terms, deterministic noise (bias) and so on. For using the framework, the only requirement is the ability to formulate the verifications of potential inequalities of type "$\mathbb{E}\phi_{k+1}^f \leq \phi_k^f + e_k$" (or sufficient conditions for satisfying it) in a tractable way—and this can be done for many optimization settings (Taylor et al., 2017a) and standard operator classes (Ryu et al., 2018) (e.g., for studying fixed-point iterations for monotone inclusion problems). The current work focuses on smooth problems without strong convexity, but the same tools can be used when strong convexity (or related notions for obtaining linear convergence results, see e.g., Necoara et al. (2018); Karimi et al. (2016)) is involved, as in e.g., Bach and Moulines (2011); Nguyen et al. (2018), where the norms of the stochastic gradients are not assumed to be uniformly bounded. Finally, minor adaptations allow studying algorithms specifically designed for finite sums problems (see e.g., Le Roux et al. (2012); Johnson and Zhang (2013); Defazio et al. (2014); Schmidt et al. (2017); Allen-Zhu (2017); Zhou et al. (2019)).

Acceleration and algorithmic design. In App. C, App. D.3, App. E.6, App. F.2, we discuss techniques for automatic step-size selection in different settings. This is done by adapting the constructive approach to efficient first-order methods by Drori and Taylor (2018) to deal with potential functions. In a few words, the idea is to study methods with *unrealistic line-search procedures* and to deduce, from the analysis, step-size policies for methods of type (SFO) that enjoys the same worst-case guarantees. The technique is also inspired by historical developments related to accelerated first-order methods (Nemirovski, 1982; Nesterov, 1983).

Application to proximal/projected methods. The methodology extends to projected and proximal settings, as previously used in the performance estimation literature (Drori, 2014; Taylor et al., 2017a, 2018b). As an example, we provide a corresponding potential function for the proximal gradient method in App. D.2.

Application to coordinate descent. We illustrate the application of technique to coordinate descent-type schemes in App. I. The assumptions used here differ from standard ones (Nesterov, 2012a), but allows a unified treatment of this kind of methods. We also use this example for illustrating the incorporation of strong convexity within the framework.

Codes

The codes used to generate and validate the results are available at github.com/AdrienTaylor/Potential-functions-for-first-order-methods.

Acknowledgments

The authors would like to thank Nicolas Flammarion, François Glineur and Yurii Nesterov for insightful discussions related on the one hand to stochastic gradient methods and on the other one to potential functions and estimate sequences. The authors also thank the three anonymous referees for their constructive remarks which helped improving the quality of this paper. The authors acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205, 2017.

Daniel Azagra and Carlos Mudarra. An extension theorem for convex functions of class $\mathrm{C}^{1,1}$ on Hilbert spaces. *Journal of Mathematical Analysis and Applications*, 446(2):1167–1182, 2017.

Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 451–459, 2011.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathrm{O}(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, pages 773–781, 2013.

Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods. *preprint arXiv:1712.04581*, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

Michael B. Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1018–1027, 2018.

Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pages 1376–1381, 2018.

Aris Daniilidis, Mounir Haddou, Erwan Le Gruyer, and Olivier Ley. Explicit formulas for $\mathrm{C}^{\{1,1\}}$ Glaeser-Whitney extensions of 1-Taylor fields in Hilbert spaces. *Proceedings of the American Mathematical Society*, 146(10):4487–4495, 2018.

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

Etienne de Klerk, Francois Glineur, and Adrien B. Taylor. Worst-case convergence analysis of gradient and newton methods through semidefinite programming performance estimation. *preprint arXiv:1709.05191*, 2017b.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1646–1654, 2014.

Olivier Devolder. Stochastic first order methods in smooth convex optimization. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics, 2011.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

Yoel Drori. *Contributions to the Complexity Analysis of Optimization Algorithms*. PhD thesis, Tel-Aviv University, 2014.

Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

Yoel Drori. On the properties of convex functions over open sets. *preprint arXiv:1812.02419*, 2018.

Yoel Drori and Adrien B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *preprint arXiv:1803.05676*, 2018.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

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

Marie Duflo. *Random Iterative Models*. Springer-Verlag, Berlin, Heidelberg, 1997.

Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Bin Hu and Laurent Lessard. Dissipativity theory for nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1549–1557, 2017.

Bin Hu, Peter Seiler, and Laurent Lessard. Analysis of approximate stochastic gradient using quadratic constraints and sequential semidefinite programs. *preprint arXiv:1711.00987*, 2017a.

Bin Hu, Peter Seiler, and Anders Rantzer. A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. In *Conference on Learning Theory (COLT)*, pages 1157–1189, 2017b.

Chonghai Hu, Weike Pan, and James T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 781–789, 2009.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory (COLT)*, pages 545–604, 2018a.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018b.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1-2):81–107, 2016.

Donghwan Kim and Jeffrey A Fessler. Another look at the fast iterative shrinkage/thresholding algorithm (fista). *SIAM Journal on Optimization*, 28(1):223–250, 2018a.

Donghwan Kim and Jeffrey A Fessler. Generalizing the optimized gradient method for smooth convex minimization. *SIAM Journal on Optimization*, 28(2):1920–1950, 2018b.

Donghwan Kim and Jeffrey A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *preprint arXiv:1803.06600*, 2018c.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2663–2671, 2012.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Johan Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, 2004.

Aleksandr M. Lyapunov and Anthony T. Fuller. *General Problem of the Stability Of Motion*. Control Theory and Applications Series. Taylor & Francis, 1992. Original text in Russian, 1892.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3325–3334, 2018.

APS Mosek. The MOSEK optimization software. *Online at http://www.mosek.com*, 54, 2010.

Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

Arkadi S. Nemirovski. Orth-method for smooth convex optimization. *Izvestia AN SSSR, Tekhnicheskaya Kibernetika*, 2:937–947, 1982.

Arkadi S. Nemirovski and David B. Yudin. Problem complexity and method efficiency in optimization. *Willey-Interscience, New York*, 1983.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate O($1/k^2$). *Soviet Mathematics Doklady*, 27:372–376, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization : a Basic Course*. Applied optimization. Kluwer Academic Publishing, 2004.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012a.

Yurii Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012b.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer International Publishing, 2018.

Lam M. Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takáč, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *preprint arXiv:1811.12403*, 2018.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software New York, 1987.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Ernest K. Ryu, Adrien B. Taylor, Carolina Bergeling, and Pontus Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *preprint arXiv:1812.00146*, 2018.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *preprint arXiv:1308.6370*, 2013.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Ziqiang Shi and Rujie Liu. Better worst-case complexity analysis of the block coordinate descent method for large scale machine learning. In *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 889–892, 2017.

Jos F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999.

Wei Tao, Zhisong Pan, Gaowei Wu, and Qing Tao. Primal averaging: A new gradient evaluation step to attain the optimal individual convergence. *IEEE Transactions on Cybernetics*, 2018.

Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4897–4906, 2018a.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3): 1283–1313, 2017a.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Performance Estimation Toolbox (PESTO): automated worst-case analysis of first-order optimization methods. In *IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283, 2017b.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2): 307–345, 2017c.

Adrien B. Taylor, Julien M Hendrickx, and François Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018b.

Onur Toker and Hitay Ozbay. On the np-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback. In *1995 Annual American Control Conference (ACC)*, volume 4, pages 2525–2526, 1995.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

Bryan Van Scoy, Randy A. Freeman, and Kevin M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2 (1):49–54, 2018.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996.

Lieven Vandenberghe, V. Ragu Balakrishnan, Ragnar Wallin, Anders Hansson, and Tae Roh. Interior-point algorithms for semidefinite programming problems derived from the kyp lemma. *Positive Polynomials in Control*, pages 579–579, 2005.

Sharan Vaswani, Francis Bach, and Mark Schmidt. 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 (AISTATS)*, pages 1195–1204, 2019.

Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A Lyapunov analysis of momentum methods in optimization. *preprint arXiv:1611.02635*, 2016.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Kaiwen Zhou, Qinghua Ding, Fanhua Shang, James Cheng, Danli Li, and Zhi-Quan Luo. Direct acceleration of saga using sampled negative momentum. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1602–1610, 2019.

Appendix A. How to read the appendix

In this section, we provide a few keys for going through the appendix.

How to read the appendix? Those additional sections provide proofs that were not presented in the core of the paper, and complementary examples of applications. The full content of the appendix is listed in Table 3.

The appendix is divided in a few sections: each of them focuses on a single optimization setup. For example, App. E focuses on stochastic methods under a bounded variance assumption $\mathbb{E}_i\|G(x;i)-f'(x)\|^2\leq\sigma^2$. In each section, we start by presenting the proofs that were not done in the core part of the text (see next paragraphs for discussions on how those proofs were found). Then, for the first few settings, we provide the derivations of the corresponding linear matrix inequalities and the parameter selection technique.

Going through the proofs. The proofs presented in the sequel where *computer-generated*, by numerically solving (1). They all consists in the exact same ideas: reformulating weighted sums of inequalities. In order to generate the proofs, we mostly used specific inequalities; the so-called *interpolation inequalities* (Taylor et al., 2017a,c); for any L-smooth μ-strongly convex function f (notation $f\in\mathcal{F}_{\mu,L}$), those inequalities can be written as

$$\begin{aligned} f(x)\geq f(y)&+\langle f'(y);x-y\rangle\\ &+\tfrac{1}{2\left(1-\frac{\mu}{L}\right)}\left(\tfrac{1}{L}\|f'(x)-f'(y)\|^2+\mu\|x-y\|^2-2\tfrac{\mu}{L}\langle f'(x)-f'(y);x-y\rangle\right),\end{aligned}$$

for all $x,y\in\mathbb{R}^d$; whereas in the L-smooth convex case they simplify to

$$f(x)\geq f(y)+\langle f'(y);x-y\rangle+\tfrac{1}{2L}\|f'(x)-f'(y)\|^2.$$

This choice is essentially motivated by the fact those inequalities are key for reformulating (1) in a tractable way. This is explained in e.g., App. D.4 were we used them for formulating the linear matrix inequalities for the gradient method.

In order to simplify most proofs, we could often directly replace some of those *interpolation inequalities* encoding smoothness and convexity by appropriate uses of either simple convexity inequalities, or with the descent lemma (which are both *weaker* than interpolation conditions):

$$\begin{aligned} f(x)&\geq f(y)+\langle f'(y);x-y\rangle,\\ f(x)&\geq f(y)-\langle f'(x);y-x\rangle-\tfrac{L}{2}\|y-x\|^2.\end{aligned}$$

Obtaining and verifying the proofs. The proofs that were *computer-aided* may seem quite mysterious. However, they can be verified in a systematic manner (essentially verifying that the claimed result can be rewritten as the given weighted sum of inequalities). The weights used in those weighted sums essentially correspond to dual variables used in our reformulation of the problem (1), and can be either guessed based on numerical solutions (see e.g., App. D.4 for the gradient method; for all interpolation inequalities, the weights in the weighted sum is equal to a feasible choice of the corresponding dual variable $\lambda_{i,j}$), or obtained through symbolic computations. Finally, it is possible to *validate* them numerically e.g., by formulating (1) via the performance estimation toolbox (Taylor et al., 2017b).

Computational cost of the approach The computational complexity of the approach can be deduced from that of semidefinite programming, see e.g., discussions in (Vandenberghe et al., 2005). The resulting complexities depend on (i) the structure of the potentials, (ii) the stochastic setting, (iii) the specific method under consideration, and (iv) whether we use the specific structure of the SDP for solving it. Therefore, we do not provide those complexities in the discussions, and rather give the computational time required to execute the different examples using one of the current state of the art solver (Mosek, 2010) on a laptop, along with the sizes of the LMIs at hand.

Section	Content
App. B	Existing methodologies for computer-assisted worst-case analyses.
App. C	High-level explanation of our proposed parameter selection technique.
App. D (no noise)	Proof for the potential for gradient descent (App. D.1). Potential for the proximal gradient method (App. D.2). Automated design of accelerated methods, parameter selection (App. D.3). Linear matrix inequalities for gradient method (App. D.4). Linear matrix inequalities for parameter selection (App. D.5).
App. E (bounded variance)	Potential for stochastic gradient descent (App. E.1). Potential for stochastic gradient descent with averaging (App. E.2). Potential for stochastic gradient descent with primal averaging (App. E.3). Stochastic gradient evaluated at averaged iterate (App. E.4). Convergence rates (App. E.5). Momentum, dual averaging and line-searches (App. E.6). Linear matrix inequalities (App. E.7).
App. F (over-parametrization)	Potential for primal averaging (App. F.1). Parameter selection (App. F.2). Linear matrix inequalities (App. F.3).
App. G (weak growth)	Primal averaging under weak growth conditions (App. G.1).
App. H (variance at $x_\star$)	Primal averaging under bounded variance at $x_\star$ (App. H).
App. I (block-coordinate)	Potential for randomized block-coordinate descent (App. I.1). Potential with strong convexity (App. I.2).

Table 3: Organization of the appendix.

Appendix B. Methodologies for computer-assisted worst-case analyses

As introduced in Sec. 1.3, two competing strategies rely on using semidefinite programming for studying worst-case performances of first-order methods.

- ◇ First, performance estimation problems (PEPs) were introduced by Drori and Teboulle (2014). This methodology relies on formulating the worst-case performance of N iterations of a given first-order method as the solution to an optimization problem. One of the key advantage of this methodology is that it is guaranteed to provide *non-improvable* (or tight) results, due to lossless semidefinite reformulations (Taylor et al., 2017c), while being applicable to a wide range of settings (Drori, 2014; Taylor et al., 2017a). This methodology was initially tailored for studying methods with sublinear convergence rates, but linear rates can be obtained as well, through smaller SDPs (Taylor et al., 2018b; Ryu et al., 2018). The methodology is available through the performance estimation toolbox (Taylor et al., 2017b), and was used to develop optimized methods (Drori and Teboulle, 2014, 2016; Kim and Fessler, 2016, 2018a,b,c; Drori and Taylor, 2018) and lower bounds (Drori, 2017).

  The main attractive features of this framework are that (i) feasible points to primal PEPs correspond to lower bounds (i.e., functions) on which the given algorithms behave badly, whereas (ii) feasible points to the dual PEPs correspond to upper bounds on the worst-case performance of the given methods. The main inherent difficulty is to convert numerics into analytical proofs. This is mostly due to the fact all iterations are treated at once, which typically implies playing with large semidefinite matrices. Those matrices may scale particularly badly in complicated optimization settings, such as in stochastic setups.

- ◇ The second approach is based on *integral quadratic constraints* (IQCs). Their uses for studying optimization methods is due to Lessard et al. (2016). This framework was initially tailored for studying settings with linear convergence rates, through the use of smaller SDPs. Recent works formally linked IQCs with performance estimation (Taylor et al., 2018a), which can be seen as feasible points to PEPs specifically designed to look for Lyapunov functions with the smallest possible linear convergence rate.

  Recent works established that IQCs could also be used for sublinear rates (Hu and Lessard, 2017; Fazlyab et al., 2018), to study stochastic methods (Hu et al., 2017a,b) and to design new first-order algorithms (Van Scoy et al., 2018; Cyrus et al., 2018).

In this work, we use the exact same technique as in PEPs for performing the lossless verification of the "potential inequality" (Pot). The resulting proofs are much simpler, while keeping some *a priori* guarantees, as we know *a priori* that there is no way of ending up with better worst-case guarantees with the same potential-based proof structure. This allows using the methodology with e.g., randomness and stochasticity.

Compared to IQCs, the approach taken here allows (i) studying sublinear rates with more general types of potential functions, beyond standard guarantees on $f(x_N) - f_\star$ and $\|x_0 - x_\star\|$, while only having to specify the content of the potential, (ii) obtaining theoretical guarantees of ending up with the best possible worst-case bounds (for the given class of potentials) for the chosen number of iterations, and (iii) allows dealing with many models involving randomness and stochasticity, even without strong convexity. The lossless reformulation is a key feature of our approach, as it allows guaranteeing that the methodology cannot fail verifying a potential function that is true. Hence, failure is still informative in understanding the behavior of the algorithm at hand.

Appendix C. A parameter selection technique

In this section, we provide a high-level overview of a technique we used for performing automatic step-size selection, whereas the details are delayed to the following sections where the technique is used. The main idea is to try to optimize algorithmic parameters for improving worst-case performance guarantees. However, from the derivations of the linear matrix inequalities of next sections, finding at the same time a valid sequence of potentials and an optimized sequence of steps actually requires solving a set of bilinear matrix inequalities (BMIs), which are intractable in general (Toker and Ozbay, 1995). One way to work around this difficulty is to study a variant of algorithm (SFO) where the parameters $\{(\alpha_k, \alpha'_k, \beta_k, \beta'_k, \delta_k, \gamma_k, \gamma'_k, \epsilon_k)\}_k$ are chosen by appropriate span-search procedures. One can then make use of the technique developed by Drori and Taylor (2018) for formulating (1) (most of the time *relaxed* versions of it) into a LMI (feasibility problem), with the particularity that for any feasible point to this LMI, one can reconstruct an algorithm of the form (SFO) (without span-searches) that achieves the same performances.

The technique relies on two elements:

- ◇ choice of an *idealized* algorithm, typically using (possibly unrealistic) line-searches,
- ◇ choice of a family of potentials that easily allows optimizing the algorithmic parameters while looking for a sequence of valid potentials. The parameters that can not be optimized are replaced by (possibly unrealistic, see below) line-search procedures.

The technique is inspired by Drori and Taylor (2018) and original developments related to accelerated methods Nemirovski (1982); Nesterov (1983). The main thing to keep in mind is that we would ideally want to optimize the algorithmic parameters for improving its worst-case guarantees. For explaining the strategy, let us consider the following example—which we carry out in Sec. D.3—: consider the first-order method given by

$$\begin{aligned} y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\ x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}), \\ z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}), \end{aligned}$$

for which we wish to optimize parameters $\{(\tau_k, \alpha_k, \delta_k, \gamma_k)\}_k$. We also consider a specific family of potentials (discussed hereafter):

$$\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k\left(f(x_k) - f_\star\right) + a_k \|z_k - x_\star\|^2, \tag{5}$$

with $Q_k \in \mathbb{S}^2$ (space of 2×2 symmetric matrices), and picking $\phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2$ and $\phi_N^f = d_N\left(f(x_N) - f_\star\right)$. As our goal is to optimize the parameter schedule $\{(\tau_k, \alpha_k, \delta_k, \gamma_k)\}_k$, a natural thing to try is to solve

$$\max_{\{(\tau_k,\alpha_k,\delta_k,\gamma_k)\}_k} \quad \max_{\phi_1^f,\ldots,\phi_{N-1}^f, d_N} d_N \text{ subject to } (\phi_0^f, \phi_1^f) \in \mathcal{V}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \mathcal{V}_{N-1}. \tag{6}$$

However, although there might be other workarounds, this problem turns out to have N BMIs. Instead of solving this problem, the workaround we propose is to study the algorithm

$$\begin{aligned} y_{k+1} &= \operatorname{argmin}_x \left\{f(x) \text{ subject to } x \in x_k + \operatorname{span}\{z_k - x_k\}\right\}, \\ x_{k+1} &= \operatorname{argmin}_x \left\{f(x) \text{ subject to } x \in y_{k+1} + \operatorname{span}\{f'(y_{k+1})\}\right\}, \\ z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}), \end{aligned} \tag{7}$$

for which one can formulate relaxed versions of (1) (i.e., sufficient conditions for verifying a potential) using ideas developed below. By denoting $\tilde{\mathcal{V}}_k$ the set of pairs (ϕ_k^f, ϕ_{k+1}^f) of potentials that can be verified for (7) with our sufficient conditions (see below), we propose to solve the following alternative to (6):

$$d_N^{(\text{LSearch})} = \max_{\{(\delta_k,\gamma_k)\}_k} \max_{\phi_1^f,\ldots,\phi_{N-1}^f,d_N} d_N \text{ subject to } (\phi_0^f, \phi_1^f) \in \tilde{\mathcal{V}}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \tilde{\mathcal{V}}_{N-1}, \quad (8)$$

from which one can recover a policy $\{(\tau_k, \alpha_k, \delta_k, \gamma_k)\}_k$ with the same $d_N^{(\text{LSearch})}$ is attained, as illustrated below. All steps involved in the analysis of (7)—similar in spirit with those presented for vanilla gradient in Sec. 3.2.1— are presented in Sec. D.3. Before going to the next section, note that it is straightforward to verify $\tilde{\mathcal{V}}_k \subseteq \mathcal{V}_k$, as for any pair (ϕ_k^f, ϕ_{k+1}^f) we have

$$(\phi_k^f, \phi_{k+1}^f) \in \tilde{\mathcal{V}}_k \Rightarrow (\phi_k^f, \phi_{k+1}^f) \in \mathcal{V}_k$$

(in other words, all potentials that can be verified are potentials). However, in general we do not have $\tilde{\mathcal{V}}_k = \mathcal{V}_k$, meaning that the analysis can fail even though a good sequence of potentials with the desired structure might exist—given a chosen structure of potentials, this problem is not present in the analysis framework presented in the core of the paper. Another reason why it might fall is when the method with line-search does not have nice worst-case guarantees.

Transforming line-search procedures to fixed-step policies. Let us provide an example of the use of a method for designing a gradient method with optimal step-size for smooth strongly convex minimization, as provided by de Klerk et al. (2017a). That is, consider the problem of minimizing a smooth strongly convex function $f \in \mathcal{F}_{\mu,L}$. We show how to let the computer choose an appropriate step-size δ_k in a gradient descent scheme, by studying the line-search variant

$$x_{k+1} = \operatorname{argmin}_x \left\{ f(x) \text{ subject to } x \in x_k + \operatorname{span}\{f'(x_k)\} \right\}.$$

For keeping things as simple as possible, let us proceed with the potential: $\phi_k^f = d_k(f(x_k) - f_\star)$; in the following lines, we illustrate the technique for choosing the step-size δ_k achieving $d_{k+1} = \rho^{-1} d_k$ (with $0 < \rho < 1$ being the convergence rate) with the smallest possible ρ. Let us first note that the rate of convergence ρ of the line-search procedure for the family of potentials that we chose satisfies by definition

$$\rho \overset{(\text{def})}{=} \max_{\substack{x_k, x_{k+1}, \\ f \in \mathcal{F}_{\mu,L}}} \frac{f(x_{k+1}) - f_\star}{f(x_k) - f_\star}$$
$$\text{subject to } x_{k+1} = x_k - \delta_k f'(x_k),\ \delta_k = \operatorname{argmin}_{\delta_k \in \mathbb{R}} \left\{ f(x_k - \delta_k f'(x_k)) \right\},$$

which can be upper-bounded by (using optimality conditions of the line-search procedure)

$$\rho \le \max_{\substack{x_k, x_{k+1}, \\ f \in \mathcal{F}_{\mu,L}}} \frac{f(x_{k+1}) - f_\star}{f(x_k) - f_\star} \text{ subject to } \langle f'(x_{k+1}); f'(x_k) \rangle = 0,\ \langle f'(x_{k+1}); x_{k+1} - x_k \rangle = 0. \quad (9)$$

It turns out that (9) often holds with equality, motivating the following developments. As a next step, one can then get an upper bound from the use of a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \le \bar{\rho}(\lambda_1, \lambda_2) \overset{(\text{def})}{=} \max_{\substack{x_k, x_{k+1}, \\ f \in \mathcal{F}_{\mu,L}}} \left\{ \frac{f(x_{k+1}) - f_\star}{f(x_k) - f_\star} + \lambda_1 \langle f'(x_{k+1}); f'(x_k) \rangle + \lambda_2 \langle f'(x_{k+1}); x_{k+1} - x_k \rangle \right\},$$

and from the same pair (λ_1, λ_2), one can create an intermediary problem

$$\rho \leq \max_{\substack{x_k, x_{k+1}, \\ f \in \mathcal{F}_{\mu,L}}} \left\{ \frac{f(x_{k+1}) - f_\star}{f(x_k) - f_\star} \text{ subject to } \lambda_1 \langle f'(x_{k+1}); f'(x_k) \rangle + \lambda_2 \langle f'(x_{k+1}); x_{k+1} - x_k \rangle = 0 \right\}$$
$$\leq \bar{\rho}(\lambda_1, \lambda_2).$$

So, for any pair $\lambda_1, \lambda_2 \in \mathbb{R}$ we get:

- an upper bound $\bar{\rho}(\lambda_1, \lambda_2)$ (possibly $+\infty$ if the choice for λ_1, λ_2 was not appropriate) on ρ,
- as a consequence, all methods satisfying $\langle f'(x_{k+1}); \lambda_1 f'(x_k) + \lambda_2 (x_{k+1} - x_k) \rangle = 0$ benefits from convergence rate at most $\bar{\rho}(\lambda_1, \lambda_2)$. In particular, the method $x_{k+1} = x_k - \frac{\lambda_2}{\lambda_1} f'(x_k)$ satisfies the previous equality.
- Finally, if there exists a choice $\lambda_1^\star, \lambda_2^\star$ such that

$$\rho = \bar{\rho}(\lambda_1^\star, \lambda_2^\star),$$

  then, assuming $\lambda_2^\star \neq 0$ (this is reasonable as otherwise $\langle f'(x_{k+1}); x_{k+1} - x_k \rangle = 0$ would not be used in the analysis), the method $x_{k+1} = x_k - \frac{\lambda_1^\star}{\lambda_2^\star} f'(x_k)$ benefits from the same worst-case guarantee as the line-search procedure. This phenomenon actually (maybe surprisingly) occurs at least in standard smooth and non-smooth convex optimization settings (Drori and Taylor, 2018). In the following sections, we use the same strategy in slightly more complicated settings—where we do not check whether formulations corresponding to (9) should hold with equality or not.

Choice of the potential family. The choice of the family of potentials into consideration plays a crucial role in the parameter selection process. For example, choice (5) for method (7) allows to easily optimize the coefficients δ_k and γ_k. The reason is technical (everything lies in the LMI formulation, which we delay to later sections), but the consequence is relatively simple: the fact $z_k - x_\star$ appears only in a norm allows writing the LMI reformulation of the potential inequality "$\phi_{k+1}^f \leq \phi_k^f$" in a way that easily permits to optimize both δ_k and γ_k (see App. D.5).

Details and extensions. We provide detailed developments relying on this technique in later sections. In its simplest form, the technique is illustrated in the deterministic smooth convex minimization setting in Sec. D.3, among others on algorithm (6).

The technique can also be adapted to other settings, such as stochastic methods, coordinate descent, or finite sums. Examples that are not provided here include the application to noise models satisfying strong growth conditions; in the latter case, the parameter selection technique allows obtaining accelerated methods similar to those of Vaswani et al. (2019).

Appendix D. Gradient method

In this section, we present the proof of Theorem 3 for the gradient method. For simplicity, we present the result for vanilla gradient method in the core of the paper; the proximal case is presented in App. D.2 and the semidefinite reformulation of (1) in App. D.4. Finally, we present the parameter selection technique to devise variants of accelerated gradient methods in App. D.3. The corresponding LMIs are presented in Sec. D.5.

The codes implementing the LMI formulations and numerics presented hereafter are provided in Sec. 5.

D.1. Proof of Theorem 3

The proof follows the same lines as previous works on performance estimation problems (see e.g., de Klerk et al. (2017a, Section 4)), and only consists in reformulating a linear combination of inequalities.

Proof Combine the following inequalities with their corresponding weights.

- ⋄ Convexity and smoothness between x_k and $x_\star$ with weight $\lambda_1 = 2L$

$$f_\star \geq f(x_k) + \langle f'(x_k); x_\star - x_k\rangle + \tfrac{1}{2L}\|f'(x_k)\|^2,$$

- ⋄ convexity and smoothness between x_{k+1} and x_k with weight $\lambda_2 = 2d_k + 2L(2 + b_k)$

$$f(x_k) \geq f(x_{k+1}) + \langle f'(x_{k+1}); x_k - x_{k+1}\rangle + \tfrac{1}{2L}\|f'(x_{k+1}) - f'(x_k)\|^2,$$

- ⋄ convexity between x_k and x_{k+1} with weight $\lambda_3 = 2L(1 + b_k) + d_k$

$$f(x_{k+1}) \geq f(x_k) + \langle f'(x_k); x_{k+1} - x_k\rangle.$$

By substituting $x_{k+1} = x_k - \frac{1}{L}f'(x_k)$, one can easily verify that the corresponding weighted sum can be reformulated exactly as the desired result:

$$\begin{aligned}
0 \geq &\lambda_1 \left[f(x_k) - f_\star + \langle f'(x_k); x_\star - x_k\rangle + \tfrac{1}{2L}\|f'(x_k)\|^2\right] \\
&+ \lambda_2 \left[f(x_{k+1}) - f(x_k) + \langle f'(x_{k+1}); x_k - x_{k+1}\rangle + \tfrac{1}{2L}\|f'(x_{k+1}) - f'(x_k)\|^2\right] \\
&+ \lambda_3 \left[f(x_k) - f(x_{k+1}) + \langle f'(x_k); x_{k+1} - x_k\rangle\right] \\
= &(d_k + 2L)(f(x_{k+1}) - f_\star) + L^2\|x_{k+1} - x_\star\|^2 + (2 + \tfrac{d_k}{L} + b_k)\|f'(x_{k+1})\|^2 \\
&- d_k(f(x_k) - f_\star) - L^2\|x_k - x_\star\|^2 - b_k\|f'(x_k)\|^2,
\end{aligned}$$

leading to

$$\begin{aligned}
(d_k + 2L)(f(x_{k+1}) - f_\star) + L^2\|x_{k+1} - x_\star\|^2 + (2 + \tfrac{d_k}{L} + b_k)\|f'(x_{k+1})\|^2 \\
\leq d_k(f(x_k) - f_\star) + L^2\|x_k - x_\star\|^2 + b_k\|f'(x_k)\|^2.
\end{aligned}$$

■

D.2. Proximal gradient method

In this section, we consider the problem

$$\min_{x\in\mathbb{R}^d} \{F(x) \equiv f(x) + h(x)\},$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and smooth ($f \in \mathcal{F}_L$) and $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed, proper and convex (notation $h \in \mathcal{F}_{0,\infty}$), where the proximal operator of h is readily available:

$$\mathrm{prox}_{\gamma h}(y) = \mathrm{argmin}_{x\in\mathbb{R}^d} \left\{\gamma h(x) + \tfrac{1}{2}\|y-x\|^2\right\}.$$

Nearly the same potential as that for the gradient method holds for the proximal gradient method (note that we have assumed $\mathrm{dom}\, f = \mathbb{R}^d$ for simplicity; other inequalities have to be used if it is not the case (de Klerk et al., 2017b; Drori, 2018)).

Theorem 9 *Let $f \in \mathcal{F}_L$, $h \in \mathcal{F}_{0,\infty}$ (class of closed, proper and convex functions), let $x_k \in \mathbb{R}^d$ be satisfying $\partial h(x_k) \neq \emptyset$, and $x_{k+1} = \mathrm{prox}_{h/L}\left(x_k - \frac{1}{L}f'(x_k)\right)$. There exists $F'(x_k) \in \partial F(x_k)$ such that inequality $\phi^F_{k+1}(x_{k+1}) \leq \phi^F_k(x_k)$ holds with*

$$\phi^F_k = d_k(F(x_k) - F_\star) + b_k\|F'(x_k)\|^2 + L^2\|x_k - x_\star\|^2,$$

for all values $b_k, d_k \geq 0$, $b_{k+1} = 1 + \frac{d_k}{L} + b_k$ and $d_{k+1} = 2L + d_k$. In particular, the above inequality holds when choosing $F'(x_{k+1}) = L(x_k - x_{k+1}) - f'(x_k) + f'(x_{k+1})$ (this choice is natural as it corresponds to using the particular subgradient of h that was used in the proximal operation).

In particular, the previous theorem establishes that

$$\phi^F_k = (2k+1)(F(x_k) - F(x_\star)) + k(k+1)\|F'(x_k)\|^2 + L^2\|x_k - x_\star\|^2,$$

is a potential for the proximal gradient method with step-size $1/L$.

Proof Combine the following inequalities with their corresponding weights. We denote by $s_{k+1} \in \partial h(x_{k+1})$ the specific subgradient used in the proximal operation, i.e., such that $x_{k+1} = x_k - \frac{1}{L}(f'(x_k) + s_{k+1})$, and $s_k \in \partial h(x_k)$ any subgradient of h at x_k. We also specifically choose $F'(x_{k+1}) = f'(x_{k+1}) + s_{k+1} \in \partial F(x_{k+1})$ and $F'(x_k) = f'(x_k) + s_k \in \partial F(x_k)$.

- ◇ Convexity of f between x_k and $x_\star$ with weight $\lambda_1 = 2L$

$$f_\star \geq f(x_k) + \langle f'(x_k); x_\star - x_k\rangle,$$

- ◇ convexity and smoothness of f between x_{k+1} and x_k with weight $\lambda_2 = 2d_k + 2L(1 + b_k)$

$$f(x_k) \geq f(x_{k+1}) + \langle f'(x_{k+1}); x_k - x_{k+1}\rangle + \tfrac{1}{2L}\|f'(x_{k+1}) - f'(x_k)\|^2,$$

- ◇ convexity of f between x_k and x_{k+1} with weight $\lambda_3 = 2Lb_k + d_k$

$$f(x_{k+1}) \geq f(x_k) + \langle f'(x_k); x_{k+1} - x_k\rangle,$$

- ⋄ convexity of h between x_{k+1} and $x_\star$ with weight $\lambda_4 = 2L$

$$h(x_\star) \geq h(x_{k+1}) + \langle s_{k+1}; x_\star - x_{k+1}\rangle,$$

- ⋄ convexity of h between x_k and x_{k+1} with weight $\lambda_5 = 2Lb_k + dk$

$$h(x_k) \geq h(x_{k+1}) + \langle s_{k+1}; x_k - x_{k+1}\rangle,$$

- ⋄ convexity of h between x_{k+1} and x_k with weight $\lambda_6 = 2Lb_k$

$$h(x_{k+1}) \geq h(x_k) + \langle s_k; x_{k+1} - x_k\rangle.$$

By substituting $x_{k+1} = x_k - \frac{1}{L}(f'(x_k) + s_{k+1})$, one can easily verify that the corresponding weighted sum can be reformulated exactly as the desired result plus a positive term:

$$\begin{aligned}
0 \geq &\lambda_1 \left[f(x_k) - f_\star + \langle f'(x_k); x_\star - x_k\rangle\right] \\
&+ \lambda_2 \left[f(x_{k+1}) - f(x_k) + \langle f'(x_{k+1}); x_k - x_{k+1}\rangle + \tfrac{1}{2L}\|f'(x_{k+1}) - f'(x_k)\|^2\right] \\
&+ \lambda_3 \left[f(x_k) - f(x_{k+1}) + \langle f'(x_k); x_{k+1} - x_k\rangle\right] \\
&+ \lambda_4 \left[h(x_{k+1}) - h(x_\star) + \langle s_{k+1}; x_\star - x_{k+1}\rangle\right] \\
&+ \lambda_5 \left[h(x_{k+1}) - h(x_k) + \langle s_{k+1}; x_k - x_{k+1}\rangle\right] \\
&+ \lambda_6 \left[h(x_k) - h(x_{k+1}) + \langle s_k; x_{k+1} - x_k\rangle\right] \\
=&(d_k + 2L)(F(x_{k+1}) - F_\star) + L^2\|x_{k+1} - x_\star\|^2 + (1 + \tfrac{d_k}{L} + b_k)\|F'(x_{k+1})\|^2 \\
&- d_k(F(x_k) - F_\star) - L^2\|x_k - x_\star\|^2 - b_k\|F'(x_k)\|^2 + b_k\|s_{k+1} - s_k\|^2,
\end{aligned}$$

leading to the desired result

$$\begin{aligned}
(d_k + 2L)(F(x_{k+1}) - F_\star) &+ L^2\|x_{k+1} - x_\star\|^2 + (1 + \tfrac{d_k}{L} + b_k)\|F'(x_{k+1})\|^2 \\
&\leq d_k(F(x_k) - F_\star) + L^2\|x_k - x_\star\|^2 + b_k\|F'(x_k)\|^2 - b_k\|s_{k+1} - s_k\|^2, \\
&\leq d_k(F(x_k) - F_\star) + L^2\|x_k - x_\star\|^2 + b_k\|F'(x_k)\|^2.
\end{aligned}$$

■

D.3. Design of accelerated methods

There are many different variants of accelerated gradient methods (see e.g., Tseng (2008)), particularly for smooth unconstrained optimization in the Euclidean setting. Here are two examples that can be obtained through the line-search strategy presented in Sec. C.

The codes implementing the LMI formulations and numerics below are provided in Sec. 5.

Design of a first accelerated method. Let us first apply our step-size selection technique to the following first-order method involving three sequences:

$$\begin{aligned}
y_{k+1} &= (1 - \tau_k)x_k + \tau_k z_k, \\
x_{k+1} &= y_{k+1} - \alpha_k f'(y_{k+1}), \\
z_{k+1} &= (1 - \delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}),
\end{aligned} \tag{10}$$

relying on the alternate version involving line-searches and optimization of the last step:

$$\begin{aligned} y_{k+1} &= \operatorname{argmin}_x \{f(x) \text{ subject to } x \in x_k + \operatorname{span}\{z_k - x_k\}\}, \\ x_{k+1} &= \operatorname{argmin}_x \left\{f(x) \text{ subject to } x \in y_{k+1} + \operatorname{span}\{f'(y_{k+1})\}\right\}, \\ z_{k+1} &= (1-\delta_k) y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}). \end{aligned} \tag{11}$$

We provide our LMI encoding sufficient conditions for verifying "$\phi_{k+1}^f \leq \phi_k^f$" in the next section (Sec. D.5) for the potential

$$\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + a_k \|z_k - x_\star\|^2 + d_k\,(f(x_k) - f_\star), \tag{12}$$

with $Q_k \in \mathbb{S}^2$ and the choice $\phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2$, $\phi_N^f = d_N\,(f(x_N) - f_\star)$. Note that one could add arbitrary sequences to (10), and other states to ϕ_k^f and still use the same tricks (see below for another example).

Let us denote $\tilde{\mathcal{V}}_k$ the set of pairs (ϕ_{k+1}^f, ϕ_k^f) for which we can verify the inequality $\phi_{k+1}^f \leq \phi_k^f$ holds for algorithm (11) (see Sec. D.5). Similar in spirit with the approach of Sec. 3.2.1 for designing a potential for vanilla gradient descent, we can now simulteneously design potential and a sequence of parameters for (10), for example by solving

$$\max_{\{(\delta_k,\gamma_k)\}_k} \max_{\phi_1^f,\ldots,\phi_{N-1}^f, d_N} d_N \text{ subject to } (\phi_0^f, \phi_1^f) \in \tilde{\mathcal{V}}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \tilde{\mathcal{V}}_{N-1}. \tag{13}$$

As an example, we provide the results obtained by solving (13) for $N = 100$ on Figure 3. Carrying a few simplifications in a similar manner to those from Sec. 3.2.1, one can arrive easily arrive to Theorem 10 (simpler expressions below).

Theorem 10 *Let $f \in \mathcal{F}_L$. For all values of $b_k, d_k \geq 0$, the iterates of algorithm (11) with $\delta_k = 1$ and $\gamma_k = \frac{d_{k+1} - d_k}{L}$ satisfy*

$$\begin{aligned} d_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{b_{k+1}}{2L}\|f'(x_{k+1})\|^2 + \tfrac{L}{2}\|z_{k+1} - x_\star\|^2 \\ \leq d_k(f(x_k) - f_\star) + \tfrac{b_k}{2L}\|f'(x_k)\|^2 + \tfrac{L}{2}\|z_k - x_\star\|^2, \end{aligned}$$

for all $d_{k+1}, b_{k+1} \in \mathbb{R}$ satisfying $d_{k+1}^2 - 2(d_k+1)d_{k+1} + \frac{d_k^2}{b_k + d_k} + d_k^2 \leq 0$ (reducing to $d_{k+1} \in [1 + d_k - \sqrt{1 + d_k},\, 1 + d_k + \sqrt{1 + d_k}]$ when $b_k = 0$) and $b_{k+1} \leq d_{k+1}$. In addition, the iterates produced by algorithm (10) with $\alpha_k = \frac{1}{L}$, and $\tau_k = \frac{d_{k+1} - d_k}{d_{k+1}}$ satisfy the same inequality.

Before proceeding with the proof, let us note two simple scenarios that are valid for Theorem 10:

- the choice $d_0 = 0$ along with $b_k = 0$ and $d_{k+1} = 1 + d_k + \sqrt{1 + d_k} = O(k^2)$, reaching acceleration (analytical version numerically matching the red curves on Figure 3),
- the choice $b_0 = d_0 = 0$ along with $b_{k+1} = d_{k+1} = 1 + d_k + \sqrt{1 + \frac{3}{2}d_k} = O(k^2)$, reaching acceleration (analytical version numerically matching the blue curves on Figure 3).

Proof Combine the following inequalities with corresponding weights:

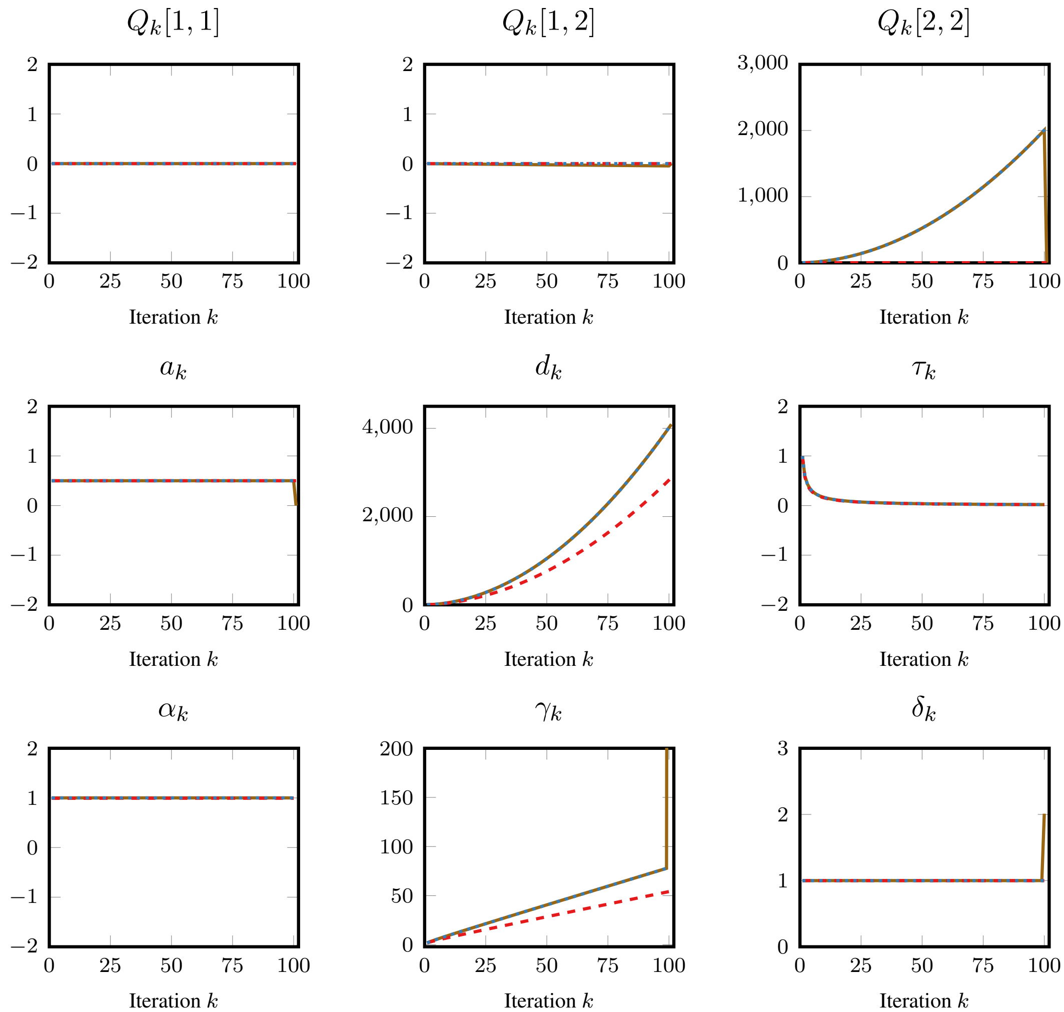

Figure 3: Numerical solution to (13) for $N = 100$ and $L = 1$ (plain brown, large values for δ_{100} and γ_{100} were capped for readability purposes; they are due to the fact we impose no control on z_N with our initial choice ϕ_N^f), forced $a_k = \frac{L}{2}$ and $Q_k = 0$ (dashed red), forced $a_k = \frac{L}{2}$, $Q_k[1,1] = Q_k[1,2] = 0$ and $Q_k[2,2] = \frac{d_k}{2L}$ (dotted blue). Total time: ~ 35 sec. on single core of Intel Core i7 1.8GHz CPU.

◇ smoothness and convexity between $x_\star$ and y_{k+1} with weight $\lambda_1 = d_{k+1} - d_k$

$$f_\star \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1})\|^2,$$

◇ smoothness and convexity between x_k and y_{k+1} with weight $\lambda_2 = d_k$

$$f(x_k) \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(x_k) - f'(y_{k+1})\|^2,$$

◇ smoothness and convexity between y_{k+1} and x_{k+1} with weight $\lambda_3 = d_{k+1}$

$$f(y_{k+1}) \geq f(x_{k+1}) + \langle f'(x_{k+1}); y_{k+1} - x_{k+1}\rangle + \tfrac{1}{2L}\|f'(x_{k+1}) - f'(y_{k+1})\|^2,$$

- $\diamond$ first line-search optimality condition for y_{k+1} with weight $\lambda_4 = d_{k+1}$

$$\langle f'(y_{k+1}); y_{k+1} - x_k \rangle \leq 0,$$

- $\diamond$ second line-search optimality condition for y_{k+1} with weight $\lambda_5 = d_{k+1} - d_k$

$$\langle f'(y_{k+1}); x_k - z_k \rangle \leq 0,$$

- $\diamond$ first line-search optimality condition for x_{k+1} with weight $\lambda_6 = d_{k+1}$

$$\langle f'(x_{k+1}); x_{k+1} - y_{k+1} \rangle \leq 0,$$

- $\diamond$ second line-search optimality condition for x_{k+1} with weight $\lambda_7 = \frac{d_{k+1}}{L}$

$$\langle f'(x_{k+1}); f'(y_{k+1}) \rangle \leq 0.$$

In the case $b_k + d_k > 0$, the weighted sum gives:

$$\begin{aligned}
0 \geq &\lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1} \rangle + \tfrac{1}{2L} \|f'(y_{k+1})\|^2 \right] \\
&+ \lambda_2 \left[f(y_{k+1}) - f(x_k) + \langle f'(y_{k+1}); x_k - y_{k+1} \rangle + \tfrac{1}{2L} \|f'(x_k) - f'(y_{k+1})\|^2 \right] \\
&+ \lambda_3 \left[f(x_{k+1}) - f(y_{k+1}) + \langle f'(x_{k+1}); y_{k+1} - x_{k+1} \rangle + \tfrac{1}{2L} \|f'(x_{k+1}) - f'(y_{k+1})\|^2 \right] \\
&+ \lambda_4 \left[\langle f'(y_{k+1}); y_{k+1} - x_k \rangle \right] \\
&+ \lambda_5 \left[\langle f'(y_{k+1}); x_k - z_k \rangle \right] \\
&+ \lambda_6 \left[\langle f'(x_{k+1}); x_{k+1} - y_{k+1} \rangle \right] \\
&+ \lambda_7 \left[\langle f'(x_{k+1}); f'(y_{k+1}) \rangle \right] \\
= &d_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{b_{k+1}}{2L} \|f'(x_{k+1})\|^2 + \tfrac{L}{2} \|z_{k+1} - x_\star\|^2 \\
&- d_k(f(x_k) - f_\star) - \tfrac{b_k}{2L} \|f'(x_k)\|^2 - \tfrac{L}{2} \|z_k - x_\star\|^2 \\
&+ \tfrac{d_{k+1} - b_{k+1}}{2L} \|f'(x_{k+1})\|^2 + \tfrac{b_k + d_k}{2L} \|f'(x_k) - \tfrac{d_k}{b_k + d_k} f'(y_{k+1})\|^2 \\
&+ \frac{-d_{k+1}^2 + 2(d_k + 1)d_{k+1} - \frac{d_k^2}{b_k + d_k} - d_k^2}{2L} \|f'(y_{k+1})\|^2,
\end{aligned}$$

which can be reformulated as

$$\begin{aligned}
d_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{b_{k+1}}{2L} \|f'(x_{k+1})\|^2 + \tfrac{L}{2} \|z_{k+1} - x_\star\|^2& \\
\leq d_k(f(x_k) - f_\star) + \tfrac{b_k}{2L} \|f'(x_k)\|^2 + \tfrac{L}{2} \|z_k - x_\star\|^2& \\
- \tfrac{d_{k+1} - b_{k+1}}{2L} \|f'(x_{k+1})\|^2 - \tfrac{b_k + d_k}{2L} \|f'(x_k) - \tfrac{d_k}{b_k + d_k} f'(y_{k+1})\|^2& \\
- \frac{-d_{k+1}^2 + 2(d_k + 1)d_{k+1} - \frac{d_k^2}{b_k + d_k} - d_k^2}{2L} \|f'(y_{k+1})\|^2& \\
\leq d_k(f(x_k) - f_\star) + \tfrac{b_k}{2L} \|f'(x_k)\|^2 + \tfrac{L}{2} \|z_k - x_\star\|^2&,
\end{aligned}$$

where the last inequality is valid as soon as $b_k + d_k > 0$ (i.e., at least $b_k > 0$ or $d_k > 0$ hold), $d_{k+1} \geq b_{k+1}$ and $-d_{k+1}^2 + 2(d_k + 1)d_{k+1} - \frac{d_k^2}{b_k + d_k} - d_k^2 \geq 0$, that is, when d_{k+1} lies in the interval

$$\left[\tfrac{b_k d_k + b_k + d_k^2 + d_k - \sqrt{2b_k^2 d_k + b_k^2 + 3b_k d_k^2 + 2b_k d_k + d_k^3 + d_k^2}}{b_k + d_k}, \tfrac{b_k d_k + b_k + d_k^2 + d_k + \sqrt{2b_k^2 d_k + b_k^2 + 3b_k d_k^2 + 2b_k d_k + d_k^3 + d_k^2}}{b_k + d_k} \right].$$

For the case $b_k = 0$ and $d_k \geq 0$ the weighted sum can be written as

$$\begin{aligned} 0 \geq & d_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{b_{k+1}}{2L}\|f'(x_{k+1})\|^2 + \tfrac{L}{2}\|z_{k+1} - x_\star\|^2 - d_k(f(x_k) - f_\star) - \tfrac{L}{2}\|z_k - x_\star\|^2 \\ & + \tfrac{d_{k+1}-b_{k+1}}{2L}\|f'(x_{k+1})\|^2 + \tfrac{d_k}{2L}\|f'(x_k) - f'(y_{k+1})\|^2 \\ & + \tfrac{-d_k^2+2d_kd_{k+1}-d_k-(d_{k+1}-2)d_{k+1}}{2L}\|f'(y_{k+1})\|^2, \end{aligned}$$

and the same simplifications can be done again, as soon as $d_{k+1} \in [1 + d_k - \sqrt{1+d_k},\ 1 + d_k + \sqrt{1+d_k}]$:

$$d_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{b_{k+1}}{2L}\|f'(x_{k+1})\|^2 + \tfrac{L}{2}\|z_{k+1} - x_\star\|^2 \leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2.$$

Finally, note that the same proofs are valid for all methods of the form (10) satisfying

$$\begin{aligned} &\langle f'(y_{k+1}); y_{k+1} - x_k + \tfrac{d_{k+1}-d_k}{d_{k+1}}(x_k - z_k)\rangle \leq 0, \\ &\langle f'(x_{k+1}); x_{k+1} - y_{k+1} + \tfrac{1}{L}f'(y_{k+1})\rangle \leq 0, \end{aligned}$$

which is true in particular when $\alpha_k = \frac{1}{L}$, and $\tau_k = \frac{d_{k+1}-d_k}{d_{k+1}}$. ■

Yet another accelerated method. For the sake of illustration, we want to point out again that some of the choices we made for the design of the previous method were quite arbitrary, and that it was actually not the only way of ending up with an accelerated first-order method. For example, one could start from a method involving only two sequences

$$\begin{aligned} y_{k+1} &= (1 - \tau_k)y_k + \tau_k z_k - \alpha_k f'(y_k), \\ z_{k+1} &= (1 - \delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_k) - \gamma_k' f'(y_{k+1}), \end{aligned} \tag{14}$$

along with the corresponding version involving a span-search

$$\begin{aligned} y_{k+1} &= \operatorname{argmin}_x \left\{ f(x) \text{ subject to } x \in y_k + \operatorname{span}\{(z_k - y_k),\ f'(y_k)\}\right\}, \\ z_{k+1} &= (1 - \delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'(y_k) - \gamma_k' f'(y_{k+1}). \end{aligned} \tag{15}$$

For the potential we choose

$$\phi_k^f = \begin{pmatrix} y_k - x_\star \\ f'(y_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} y_k - x_\star \\ f'(y_k) \end{pmatrix} + a_k\|z_k - x_\star\|^2 + d_k\,(f(y_k) - f_\star),$$

with $Q_k \in \mathbb{S}^2$ and start with the choice $\phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2$, $\phi_N^f = d_N\,(f(x_N) - f_\star)$. Let us denote $\tilde{\mathcal{V}}_k$ the set of pairs (ϕ_{k+1}^f, ϕ_k^f) for which we can verify the inequality $\phi_{k+1}^f \leq \phi_k^f$ holds for algorithm (15). As before, we can now design at the same time a potential and a sequence of parameters for (14), for example by solving

$$\max_{\{(\delta_k,\gamma_k,\gamma_k')\}_k}\ \max_{\phi_1^f,\ldots,\phi_{N-1}^f, d_N} d_N \text{ subject to } (\phi_0^f, \phi_1^f) \in \tilde{\mathcal{V}}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \tilde{\mathcal{V}}_{N-1}. \tag{16}$$

As an example, we provide the results obtained by solving (16) for $N = 100$ on Figure 4. As for the previous cases, it presents a few possible simplifications, leading to the following theorem.

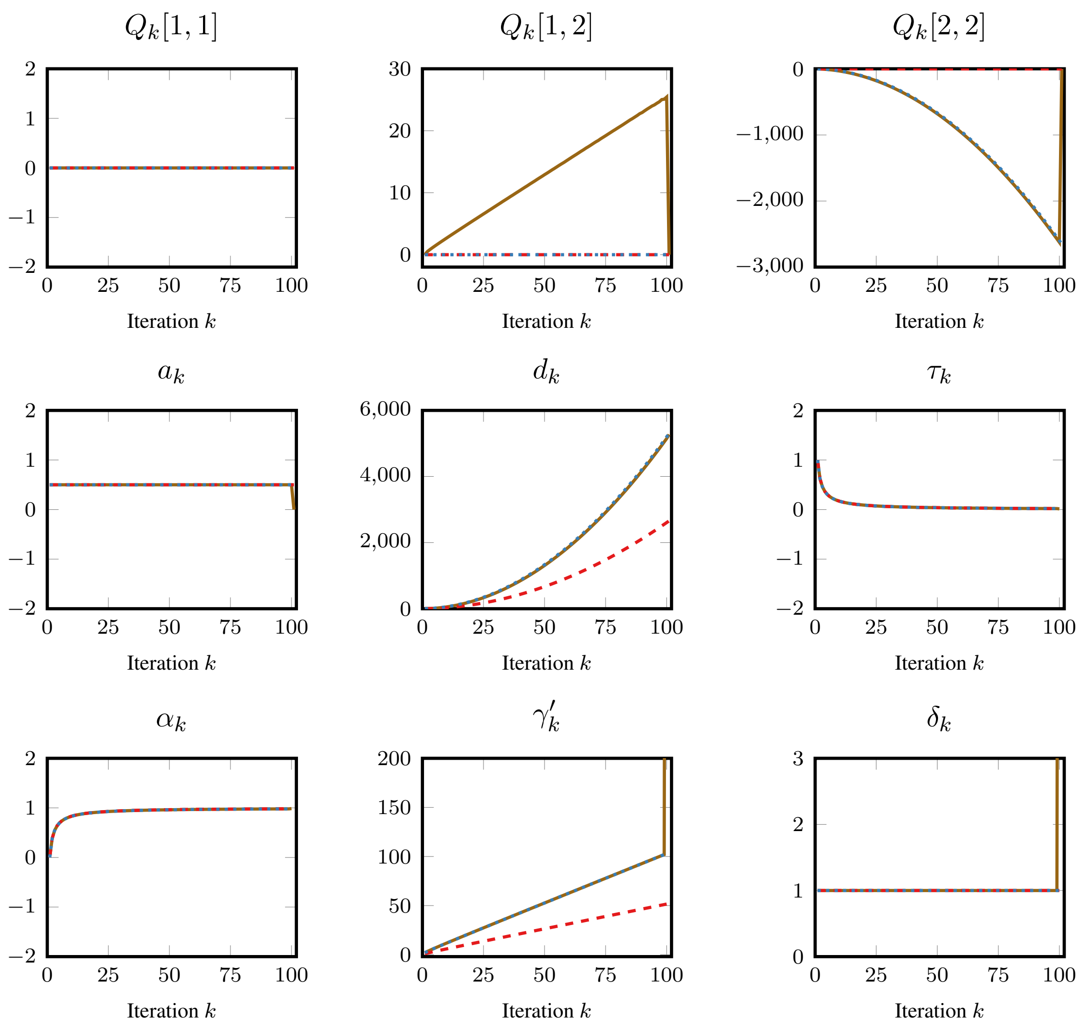

Figure 4: Numerical solution to (16) for $N = 100$ and $L = 1$ (plain brown, large values for δ_{100} and γ_{100} were capped for readability purposes; they are due to the fact we impose no control on z_N with our initial choice ϕ_N^f), forced $a_k = \frac{L}{2}$ and $Q_k = 0$ (dashed red), and forced $a_k = \frac{L}{2}$, $Q_k[1,1] = Q_k[1,2] = 0$ and $Q_k[2,2] = -\frac{d_k}{2L}$ (dotted blue). For convenience, we did not plot γ_k which numerically appeared to be negligible compared to other variables (about 10^{-7}). Total time: $\sim$ 30 sec. on single core of Intel Core i7 1.8GHz CPU.

The following theorem presents one of the possible outcome of the approach (simple choices are presented just after the theorem).

Theorem 11 *Let $f \in \mathcal{F}_L$. For all values of $d_k \geq b_k \geq 0$, the iterates of algorithm (14) with $\delta_k = 1$, $\gamma_k = 0$, and $\gamma'_k = \frac{d_{k+1}-d_k}{L}$ satisfy*

$$d_{k+1}(f(y_{k+1}) - f_\star) - \tfrac{b_{k+1}}{2L}\|f'(y_{k+1})\|^2 + \tfrac{L}{2}\|z_{k+1} - x_\star\|^2 \\ \leq d_k(f(y_k) - f_\star) - \tfrac{b_k}{2L}\|f'(y_k)\|^2 + \tfrac{L}{2}\|z_k - x_\star\|^2$$

for all $d_{k+1}, b_{k+1} \geq 0$ *satisfying* $b_{k+1} + d_{k+1} - (d_k - d_{k+1})^2 \geq 0$. *In addition, the iterates of algorithm* (15) *with* $\alpha_k = \frac{d_k}{d_{k+1}L}$ *and* $\tau_k = \frac{d_{k+1}-d_k}{d_{k+1}}$ *satisfy the same inequality.*

Before proceeding with the proof, let us consider those two simpler cases:

- the choice $d_0 = 0$ along with $b_k = 0$ and $d_{k+1} = \frac{1}{2} + d_k + \sqrt{\frac{1}{4} + d_k} = O(k^2)$, reaching acceleration (analytical version numerically matching the red curves on Figure 3).
- The choice $b_0 = d_0 = 0$ along with $b_{k+1} = d_{k+1} = 1+d_k+\sqrt{1+2d_k} = O(k^2)$, reaching acceleration (analytical version numerically matching the blue curves on Figure 3), for example for the point $x_k = y_k - \frac{1}{L}f'(y_k)$ for which we get $f(x_k) - f_\star \leq f(y_k) - f_\star - \frac{1}{2L}\|f'(y_k)\|^2 \leq \frac{L}{2d_k}\|z_0 - x_\star\|^2$.

Proof Combine the following inequalities with corresponding weights:

- smoothness and convexity between $x_\star$ and y_{k+1} with weight $\lambda_1 = d_{k+1} - d_k$:
$$f_\star \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1})\|^2,$$
- smoothness and convexity between y_k and y_{k+1} with weight $\lambda_2 = d_k$:
$$f(y_k) \geq f(y_{k+1}) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_k) - f'(y_{k+1})\|^2,$$
- first line-search optimality condition for y_{k+1} with weight $\lambda_3 = d_{k+1}$
$$\langle f'(y_{k+1}); y_{k+1} - y_k\rangle \leq 0,$$
- second line-search optimality condition for y_{k+1} with weight $\lambda_4 = d_{k+1} - d_k$
$$\langle f'(y_{k+1}); y_k - z_k\rangle \leq 0,$$
- third line-search optimality condition for y_{k+1} with weight $\lambda_5 = \frac{d_k}{L}$
$$\langle f'(y_{k+1}); f(y_k)\rangle \leq 0.$$

The weighted sum can be rewritten as

$$\begin{aligned}
0 \geq &\lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1})\|^2\right] \\
&+ \lambda_2 \left[f(y_{k+1}) - f(y_k) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_k) - f'(y_{k+1})\|^2\right] \\
&+ \lambda_3 \left[\langle f'(y_{k+1}); y_{k+1} - y_k\rangle\right] \\
&+ \lambda_4 \left[\langle f'(y_{k+1}); y_k - z_k\rangle\right] \\
&+ \lambda_5 \left[\langle f'(y_{k+1}); f(y_k)\rangle\right] \\
=&d_{k+1}(f(y_{k+1}) - f_\star) - \tfrac{b_{k+1}}{2L}\|f'(y_{k+1})\|^2 + \tfrac{L}{2}\|z_{k+1} - x_\star\|^2 \\
&- d_k(f(y_k) - f_\star) + \tfrac{b_k}{2L}\|f'(y_k)\|^2 - \tfrac{L}{2}\|z_k - x_\star\|^2 \\
&+ \tfrac{(d_k - b_k)}{2L}\|f'(y_k)\|^2 + \tfrac{\left(b_{k+1} - (d_k - d_{k+1})^2 + d_{k+1}\right)}{2L}\|f'(y_{k+1})\|^2,
\end{aligned}$$

which, in turn, gives:

$$\begin{aligned}
&d_{k+1}(f(y_{k+1})-f_\star)-\tfrac{b_{k+1}}{2L}\|f'(y_{k+1})\|^2+\tfrac{L}{2}\|z_{k+1}-x_\star\|^2\\
&\quad\leq d_k(f(y_k)-f_\star)-\tfrac{b_k}{2L}\|f'(y_k)\|^2+\tfrac{L}{2}\|z_k-x_\star\|^2\\
&\qquad-\tfrac{(d_k-b_k)}{2L}\|f'(y_k)\|^2-\tfrac{b_{k+1}-(d_k-d_{k+1})^2+d_{k+1}}{2L}\|f'(y_{k+1})\|^2\\
&\quad\leq d_k(f(y_k)-f_\star)-\tfrac{b_k}{2L}\|f'(y_k)\|^2+\tfrac{L}{2}\|z_k-x_\star\|^2,
\end{aligned}$$

where the last inequality is valid as soon as $d_k \geq b_k$, and $b_{k+1}-(d_k-d_{k+1})^2+d_{k+1}\geq 0$. The same bound is valid for all methods satisfying

$$\langle f'(y_{k+1}); d_{k+1}(y_{k+1}-y_k)+(d_{k+1}-d_k)(y_k-z_k)+\tfrac{d_k}{L}f'(y_k)\rangle \leq 0,$$

that include algorithm (14) when $\alpha_k=\frac{d_k}{d_{k+1}L}$ and $\tau_k=\frac{d_{k+1}-d_k}{d_{k+1}}$. ■

D.4. Linear matrix inequalities for the gradient method

In this section, we provide the LMI that was used in Sec. 3.2.1, i.e., for vanilla gradient descent. One can extend this LMI to the projected/proximal case using the tools by Taylor et al. (2017a), resulting in proofs such as in previous section. The code implementing the LMI formulation below is provided in Sec. 5.

Let us recall that the target is to reformulate (1). In this case, it corresponds to

$$\begin{aligned}
0\geq \max_{f,x_k,x_{k+1},x_\star}\ & \phi^f_{k+1}(x_{k+1})-\phi^f_k(x_k)\\
\text{s.t. } & x_{k+1}=x_k-\tfrac{1}{L}f'(x_k),\\
& f\in\mathcal{F}_L \text{ and } f'(x_\star)=0,
\end{aligned}$$

for when

$$\phi^f_k=\begin{pmatrix}x_k-x_\star\\ f'(x_k)\end{pmatrix}^\top\left[\begin{pmatrix}a_k & c_k\\ c_k & b_k\end{pmatrix}\otimes I_d\right]\begin{pmatrix}x_k-x_\star\\ f'(x_k)\end{pmatrix}+d_k\left(f(x_k)-f_\star\right). \tag{17}$$

They first key step in the reformulation is standard from the performance estimation literature and consists in using a discrete version of the variable f, as follows

$$\begin{aligned}
0\geq \max_{\substack{f_k,f_{k+1},f_\star,\\ g_k,g_{k+1},g_\star,\\ x_k,x_{k+1},x_\star}}\ & \phi^f_{k+1}(x_{k+1})-\phi^f_k(x_k),\\
\text{s.t. } & x_{k+1}=x_k-\tfrac{1}{L}g_k,\\
& \exists f\in\mathcal{F}_L \text{ such that } f(x_i)=f_i \text{ and } f'(x_i)=g_i \text{ for all } i\in\{k,k+1,\star\},\\
& g_\star=0.
\end{aligned}\tag{18}$$

The existence constraint is often referred to as an *interpolation constraint* and can be reformulated using appropriate quadratic inequalities (Taylor et al., 2017c, Theorem 4) — this is also sometimes referred to as *extensions* (Azagra and Mudarra, 2017; Daniilidis et al., 2018). For smooth convex functions, this can be reformulated as

$$\exists f \in \mathcal{F}_L \text{ such that } f(x_i) = f_i \text{ and } f'(x_i) = g_i \text{ for all } i \in \{k, k+1, \star\}$$
$$\Leftrightarrow f_i \geq f_j + \langle g_j; x_i - x_j\rangle + \tfrac{1}{2L}\|g_i - g_j\|^2 \text{ for all } i, j \in \{k, k+1, \star\}.$$

This allows reformulating (18) as a linear matrix inequality. For doing that, let us choose without loss of generality $x_\star = g_\star = 0$ and $f_\star = 0$ and introduce two matrices P and F:

$$P = [\, x_k \quad g_k \quad g_{k+1} \,], \quad F = [\, f_k \quad f_{k+1} \,],$$

and the corresponding Gram matrix $G = P^\top P \succeq 0$, that is,

$$G = \begin{pmatrix} \|x_k\|^2 & \langle g_k; x_k\rangle & \langle g_{k+1}; x_k\rangle \\ \langle g_k; x_k\rangle & \|g_k\|^2 & \langle g_k; g_{k+1}\rangle \\ \langle g_{k+1}; x_k\rangle & \langle g_k; g_{k+1}\rangle & \|g_{k+1}\|^2 \end{pmatrix}.$$

Let us introduce the following shorthand notations for picking up elements in the Gram matrix G and in F: we choose $\mathbf{x}_k, \mathbf{x}_{k+1}, \mathbf{x}_\star, \mathbf{g}_k, \mathbf{g}_{k+1}, \mathbf{g}_\star \in \mathbb{R}^3$ and $\mathbf{f}_k, \mathbf{f}_{k+1}, \mathbf{f}_\star \in \mathbb{R}^2$ such that

$$\begin{aligned} x_k &= P\mathbf{x}_k, & x_{k+1} &= P\mathbf{x}_{k+1}, & x_\star &= P\mathbf{x}_\star, \\ g_k &= P\mathbf{g}_k, & g_{k+1} &= P\mathbf{g}_{k+1}, & g_\star &= P\mathbf{g}_\star, \\ f_k &= F\mathbf{f}_k, & f_{k+1} &= F\mathbf{f}_{k+1}, & f_\star &= F\mathbf{f}_\star. \end{aligned}$$

In other words, by letting e_i be the unit vector with 1 as its i^{th} component we have $\mathbf{x}_k := e_1 \in \mathbb{R}^3$, $\mathbf{g}_k := e_2 \in \mathbb{R}^3$, $\mathbf{g}_{k+1} := e_3 \in \mathbb{R}^3$ along with

$$\mathbf{x}_{k+1} := \mathbf{x}_k - \tfrac{1}{L}\mathbf{g}_k \in \mathbb{R}^3, \quad \mathbf{x}_\star = \mathbf{g}_\star := 0 \in \mathbb{R}^3,$$

and $\mathbf{f}_k := e_1 \in \mathbb{R}^2$, $\mathbf{f}_{k+1} := e_2 \in \mathbb{R}^2$ and $\mathbf{f}_\star := 0 \in \mathbb{R}^2$. Those notations allow conveniently writing scalar products by picking up elements in G. For example, $\langle g_{k+1}; x_{k+1} - x_\star\rangle$ can be written as

$$\langle g_{k+1}; x_{k+1} - x_\star\rangle = (P\mathbf{g}_{k+1})^\top P(\mathbf{x}_{k+1} - \mathbf{x}_\star) = \operatorname{Trace}\left(G\mathbf{g}_{k+1}(\mathbf{x}_{k+1} - \mathbf{x}_\star)^\top\right).$$

Also, one can equivalently rewrite the inequality

$$f_i - f_j - \langle g_j; x_i - x_j\rangle - \tfrac{1}{2L}\|g_i - g_j\|^2 \geq 0,$$

as

$$F(\mathbf{f}_i - \mathbf{f}_j) + \operatorname{Trace}\left(G \begin{pmatrix} \mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j \end{pmatrix} M \begin{pmatrix} \mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j \end{pmatrix}^\top\right) \geq 0,$$

where

$$M := \tfrac{1}{2}\begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1/L & 1/L \\ -1 & 1 & 1/L & -1/L \end{pmatrix}. \tag{19}$$

We can also rewrite

$$\phi_k^f(x_k) = d_k F(\mathbf{f}_k - \mathbf{f}_\star) + \mathrm{Trace}\left(G\begin{pmatrix}\mathbf{x}_k - \mathbf{x}_\star & \mathbf{g}_k\end{pmatrix}\begin{pmatrix}a_k & c_k\\ c_k & b_k\end{pmatrix}\begin{pmatrix}\mathbf{x}_k^\top - \mathbf{x}_\star^\top\\ \mathbf{g}_k^\top\end{pmatrix}\right),$$
$$\begin{aligned}\phi_{k+1}^f(x_{k+1}) &= d_{k+1} F(\mathbf{f}_{k+1} - \mathbf{f}_\star)\\ &\quad + \mathrm{Trace}\left(G\begin{pmatrix}\mathbf{x}_{k+1} - \mathbf{x}_\star & \mathbf{g}_{k+1}\end{pmatrix}\begin{pmatrix}a_{k+1} & c_{k+1}\\ c_{k+1} & b_{k+1}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{k+1}^\top - \mathbf{x}_\star^\top\\ \mathbf{g}_{k+1}^\top\end{pmatrix}\right),\end{aligned}$$

leading to the following reformulation of (18) given by

$$\begin{aligned}0 \geq \max_{G,\, F}\ & \phi_{k+1}^f(x_{k+1}) - \phi_k^f(x_k),\\ \text{s.t. } & G \succeq 0,\\ & F(\mathbf{f}_i - \mathbf{f}_j) + \mathrm{Trace}\left(G\begin{pmatrix}\mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix} M \begin{pmatrix}\mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix}^\top\right) \geq 0\\ & \quad \text{for all } i,j \in \{k, k+1, \star\},\end{aligned}$$

whose feasibility (for given parameters $\{(a_k, b_k, c_k, d_k), (a_{k+1}, b_{k+1}, c_{k+1}, d_{k+1})\}$ and L) can be verified with standard semidefinite packages (Sturm, 1999; Mosek, 2010). Let us provide the corresponding linear matrix inequality, which is simply obtained by dualizing the previous formulation. By associating one multiplier for each *interpolation constraint*

$$F(\mathbf{f}_i - \mathbf{f}_j) + \mathrm{Trace}\left(G\begin{pmatrix}\mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix} M \begin{pmatrix}\mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix}^\top\right) \geq 0 \quad : \lambda_{i,j},$$

one can arrive to the final LMI for gradient descent (we use the notation $\mathbb{S}^3$ to denote the set of 3×3 symmetric matrices).

The inequality $\phi_{k+1}^f(x_{k+1}) \leq \phi_k^f(x_k)$ (ϕ_k^f defined in (17)) holds for all $x_k \in \mathbb{R}^d$, all $f \in \mathcal{F}_L(\mathbb{R}^d)$, all $d \in \mathbb{N}$ and $x_{k+1} = x_k - \frac{1}{L} f'(x_k)$ if and only if there exists $\{\lambda_{i,j}\}_{i,j\in I_k}$, with $I_k := \{k, k+1, \star\}$, such that

$$\begin{aligned}&\lambda_{i,j} \geq 0 \text{ for all } i,j \in I_k,\\ &d_{k+1}(\mathbf{f}_{k+1} - \mathbf{f}_\star) - d_k(\mathbf{f}_k - \mathbf{f}_\star) + \sum_{i,j\in I_k} \lambda_{i,j}(\mathbf{f}_i - \mathbf{f}_j) = 0 \text{ (linear constraint in } \mathbb{R}^2\text{)},\\ &V_{k+1} - V_k + \sum_{i,j\in I_k} \lambda_{i,j} M_{i,j} \preceq 0 \text{ (linear matrix inequality in } \mathbb{S}^3\text{)},\end{aligned}$$

with

$$\begin{aligned}V_k &:= \begin{pmatrix}\mathbf{x}_k - \mathbf{x}_\star & \mathbf{g}_k\end{pmatrix}\begin{pmatrix}a_k & c_k\\ c_k & b_k\end{pmatrix}\begin{pmatrix}\mathbf{x}_k^\top - \mathbf{x}_\star^\top\\ \mathbf{g}_k^\top\end{pmatrix} \in \mathbb{S}^3,\\ V_{k+1} &:= \begin{pmatrix}\mathbf{x}_{k+1} - \mathbf{x}_\star & \mathbf{g}_{k+1}\end{pmatrix}\begin{pmatrix}a_{k+1} & c_{k+1}\\ c_{k+1} & b_{k+1}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{k+1}^\top - \mathbf{x}_\star^\top\\ \mathbf{g}_{k+1}^\top\end{pmatrix} \in \mathbb{S}^3,\\ M_{i,j} &:= \begin{pmatrix}\mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix} M \begin{pmatrix}\mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix}^\top \in \mathbb{S}^3.\end{aligned}$$

D.5. Linear matrix inequalities for the design procedure

We provide the LMI formulation of $\tilde{\mathcal{V}}_k$ for algorithm (10). The code implementing the LMI formulation below is provided in Sec. 5.

We proceed mostly like in Sec. D.4 for reformulating (1). That is, we now reformulate

$$\begin{aligned} 0 \geq \max_{\substack{f_{y_k}, f_{x_k}, f_{x_{k+1}}, f_\star, \\ g_{y_k}, g_{x_k}, g_{x_{k+1}}, g_\star, \\ z_k, y_k, x_k, x_{k+1}, x_\star}} \; & \phi^f_{k+1}(x_{k+1}, z_{k+1}) - \phi^f_k(x_k, z_k), \\ \text{s.t. } & \langle g_{y_{k+1}}; y_{k+1} - x_k \rangle = 0, \quad \langle g_{y_{k+1}}; z_k - x_k \rangle = 0, \\ & \langle g_{x_{k+1}}; y_{k+1} - x_{k+1} \rangle = 0, \quad \langle g_{x_{k+1}}; g_{y_{k+1}} \rangle = 0, \\ & z_{k+1} = (1-\delta_k) y_{k+1} + \delta_k z_k - \gamma_k f'(y_{k+1}), \\ & \exists f \in \mathcal{F}_L \text{ such that } f(x_i) = f_i \text{ and } f'(x_i) = g_i \text{ for all } (x_i, g_i, f_i) \in S, \\ & g_\star = 0, \end{aligned} \tag{20}$$

where S is the discrete version of f:

$$S := \{(x_k, g_{x_k}, f_{x_k}), (x_{k+1}, g_{x_{k+1}}, f_{x_{k+1}}), (y_{k+1}, g_{y_{k+1}}, f_{y_{k+1}}), (x_\star, g_\star, f_\star)\}.$$

Note again that (20) is not an exact reformulation of (1) for the method (11): it is only a sufficient condition for (Pot) to be satisfied for the method (11) and all $d \in \mathbb{N}$, $f \in \mathcal{F}_L$ and $(x_k, z_k) \in \mathbb{R}^d \times \mathbb{R}^d$.

SDP reformulation. As in Sec. D.4, we use a vector F and a Gram matrix $G = P^\top P \succeq 0$ for encoding the problem, with P and F defined as

$$P := [\, z_k \quad x_k \quad x_{k+1} \quad y_{k+1} \quad g_{x_k} \quad g_{x_{k+1}} \quad g_{y_{k+1}} \,], \quad F := [\, f_{x_k} \quad f_{x_{k+1}} \quad f_{y_{k+1}} \,].$$

For denoting all points whose gradient and functions values are needed within the formulation, we re-organize them and denote

$$\begin{aligned} (w_\star, g_\star, f_\star) &:= (x_\star, g_\star, f_\star), & (w_1, g_1, f_1) &:= (x_k, g_{x_k}, f_{x_k}), \\ (w_2, g_2, f_2) &:= (x_{k+1}, g_{x_{k+1}}, f_{x_{k+1}}), & (w_3, g_3, f_3) &:= (y_{k+1}, g_{y_{k+1}}, f_{y_{k+1}}). \end{aligned}$$

Let us denote $\mathbf{w}_\star$, $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{g}_\star$, $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, $\mathbf{z}_k \in \mathbb{R}^7$ and $\mathbf{f}_\star, \mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3 \in \mathbb{R}^3$ the vectors such that

$$\begin{aligned} x_k &= P\mathbf{w}_1, & x_{k+1} &= P\mathbf{w}_2, & y_{k+1} &= P\mathbf{w}_3, \\ g_{x_k} &= P\mathbf{g}_1, & g_{x_{k+1}} &= P\mathbf{g}_2, & g_{y_{k+1}} &= P\mathbf{g}_3, \\ f_{x_k} &= P\mathbf{f}_1, & f_{x_{k+1}} &= P\mathbf{f}_2, & f_{y_{k+1}} &= P\mathbf{f}_3. \end{aligned}$$

along with $z_k = P\mathbf{z}_k$, $\mathbf{w}_\star = \mathbf{g}_\star = 0$ and $\mathbf{f}_\star = 0$. More precisely, we define $\mathbf{z}_k := e_1 \in \mathbb{R}^7$

$$\begin{aligned} \mathbf{w}_1 &:= e_2 \in \mathbb{R}^7, & \mathbf{w}_2 &:= e_3 \in \mathbb{R}^7, & \mathbf{w}_3 &:= e_4 \in \mathbb{R}^7, \\ \mathbf{g}_1 &:= e_5 \in \mathbb{R}^7, & \mathbf{g}_2 &:= e_6 \in \mathbb{R}^7, & \mathbf{g}_3 &:= e_7 \in \mathbb{R}^7, \\ \mathbf{f}_1 &:= e_1 \in \mathbb{R}^3, & \mathbf{f}_2 &:= e_2 \in \mathbb{R}^3, & \mathbf{f}_3 &:= e_3 \in \mathbb{R}^3. \end{aligned}$$

We encode *interpolation constraints* and the corresponding *multiplier* $\lambda_{i,j}$ as before, for all points for which we needed to use the gradient and function values, i.e., for $x_\star$, x_k, x_{k+1} and y_{k+1}:

$$F(\mathbf{f}_i - \mathbf{f}_j) + \operatorname{Trace}\left(G \begin{pmatrix} \mathbf{w}_i & \mathbf{w}_j & \mathbf{g}_i & \mathbf{g}_j \end{pmatrix} M \begin{pmatrix} \mathbf{w}_i & \mathbf{w}_j & \mathbf{g}_i & \mathbf{g}_j \end{pmatrix}^\top \right) \geq 0 \quad : \lambda_{i,j},$$

where $i,j \in \{\star,1,2,3\}$. Next, the four line-search conditions can be encoded as $\mathrm{Trace}(G\,A_i)=0$ (for $i=1,\ldots,4$) with

$$\begin{aligned}
A_1 &= \tfrac{1}{2}\left(\mathbf{g}_3(\mathbf{w}_3-\mathbf{w}_1)^\top + (\mathbf{w}_3-\mathbf{w}_1)\mathbf{g}_3^\top\right), \text{for } \langle f'(y_{k+1}); y_{k+1}-x_k\rangle = \mathrm{Trace}(A_1 G)=0,\\
A_2 &= \tfrac{1}{2}\left(\mathbf{g}_3(\mathbf{z}_k-\mathbf{w}_1)^\top + (\mathbf{z}_k-\mathbf{w}_1)\mathbf{g}_3^\top\right), \text{for } \langle f'(y_{k+1}); z_k-x_k\rangle = \mathrm{Trace}(A_2 G)=0,\\
A_3 &= \tfrac{1}{2}\left(\mathbf{g}_2(\mathbf{w}_3-\mathbf{w}_2)^\top + (\mathbf{w}_3-\mathbf{w}_2)\mathbf{g}_2^\top\right), \text{for } \langle f'(x_{k+1}); y_{k+1}-x_{k+1}\rangle = \mathrm{Trace}(A_3 G)=0,\\
A_4 &= \tfrac{1}{2}\left(\mathbf{g}_2\mathbf{g}_3^\top + \mathbf{g}_3\mathbf{g}_2^\top\right), \text{for } \langle f'(x_{k+1}); f'(y_{k+1})\rangle = \mathrm{Trace}(A_4 G)=0,
\end{aligned}$$

and we denote the corresponding multipliers by $\mu_1,\ldots,\mu_4$. Finally, we define

$$\mathbf{z}_{k+1} := (1-\delta_k)\mathbf{w}_3 + \delta_k\mathbf{z}_k - \gamma_k\mathbf{g}_3.$$

The LMI formulation is, again, a dual to the SDP reformulation of (20) (see below):

$$\begin{aligned}
0 \geq \max_{G,\,F}\ & \phi^f_{k+1}(x_{k+1},z_{k+1}) - \phi^f_k(x_k,z_k),\\
\text{s.t. } & \mathrm{Trace}(G\,A_1)=0,\ \mathrm{Trace}(G\,A_2)=0,\ \mathrm{Trace}(G\,A_3)=0,\ \mathrm{Trace}(G\,A_4)=0,\\
& F(\mathbf{f}_i-\mathbf{f}_j) + \mathrm{Trace}\left(G\begin{pmatrix}\mathbf{w}_i & \mathbf{w}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix} M \begin{pmatrix}\mathbf{w}_i & \mathbf{w}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix}^\top\right) \geq 0\\
& \quad \text{for all } i,j\in\{\star,1,2,3\}.
\end{aligned}$$

We use $\mathbb{S}^7$ to denote the set of 7×7 symmetric matrices, and use M as defined in (19).

Given two triplets (a_k,d_k,Q_k), $(a_{k+1},d_{k+1},Q_{k+1})$ and a pair (δ_k,γ_k), if there exists $\{\lambda_{i,j}\}_{i,j\in I_k}$, with $I_k:=\{\star,1,2,3\}$ and $\{\mu_i\}_{i\in\{1,2,3,4\}}$ such that

$$\begin{aligned}
&\lambda_{i,j}\geq 0 \text{ for all } i,j\in I_k,\\
&d_{k+1}(\mathbf{f}_{k+1}-\mathbf{f}_\star) - d_k(\mathbf{f}_k-\mathbf{f}_\star) + \sum_{i,j\in I_k}\lambda_{i,j}(\mathbf{f}_i-\mathbf{f}_j) = 0 \text{ (linear constraint in } \mathbb{R}^3),\\
&V_{k+1}-V_k + \sum_{i,j\in I_k}\lambda_{i,j}M_{i,j} + \sum_{i\in\{1,2,3,4\}}\mu_i\,A_i \preceq 0 \text{ (linear matrix inequality in } \mathbb{S}^7),
\end{aligned}$$

with

$$\begin{aligned}
V_k &:= a_k(\mathbf{z}_k-\mathbf{x}_\star)(\mathbf{z}_k-\mathbf{x}_\star)^\top + \begin{pmatrix}\mathbf{w}_1-\mathbf{w}_\star & \mathbf{g}_1\end{pmatrix} Q_k \begin{pmatrix}\mathbf{w}_1^\top-\mathbf{w}_\star^\top\\ \mathbf{g}_1^\top\end{pmatrix}\in\mathbb{S}^7,\\
V_{k+1} &:= a_{k+1}(\mathbf{z}_{k+1}-\mathbf{x}_\star)(\mathbf{z}_{k+1}-\mathbf{x}_\star)^\top + \begin{pmatrix}\mathbf{w}_2-\mathbf{w}_\star & \mathbf{g}_2\end{pmatrix} Q_{k+1} \begin{pmatrix}\mathbf{w}_2^\top-\mathbf{w}_\star^\top\\ \mathbf{g}_2^\top\end{pmatrix}\in\mathbb{S}^7,\\
M_{i,j} &:= \begin{pmatrix}\mathbf{w}_i & \mathbf{w}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix} M \begin{pmatrix}\mathbf{w}_i & \mathbf{w}_j & \mathbf{g}_i & \mathbf{g}_j\end{pmatrix}^\top\in\mathbb{S}^7,
\end{aligned}$$

then the inequality $\phi^f_{k+1}(x_{k+1})\leq\phi^f_k(x_k)$ (ϕ^f_k defined in (12)) holds for all $f\in\mathcal{F}_L(\mathbb{R}^d)$, all $d\in\mathbb{N}$, all $x_k,z_k\in\mathbb{R}^d$, and x_{k+1},z_{k+1} generated by method (11).

There are several ways of optimizing the parameters (δ_k, γ_k) in the process (an alternative to what we propose below is to use an appropriate Schur complement), for example

$$a_{k+1}(\mathbf{z}_{k+1}-\mathbf{x}_\star)(\mathbf{z}_{k+1}-\mathbf{x}_\star)^\top = a_k \begin{pmatrix} \mathbf{w}_3 \\ \mathbf{z}_k - \mathbf{w}_3 \\ \mathbf{g}_3 \end{pmatrix}^\top \begin{pmatrix} 1 \\ \delta_k \\ -\gamma_k \end{pmatrix} \begin{pmatrix} 1 \\ \delta_k \\ -\gamma_k \end{pmatrix}^\top \begin{pmatrix} \mathbf{w}_3 \\ \mathbf{z}_k - \mathbf{w}_3 \\ \mathbf{g}_3 \end{pmatrix},$$

$$= \begin{pmatrix} \mathbf{w}_3 \\ \mathbf{z}_k - \mathbf{w}_3 \\ \mathbf{g}_3 \end{pmatrix}^\top \underbrace{\begin{pmatrix} a_k & a_k\delta_k & -a_k\gamma_k \\ a_k\delta_k & a_k\delta_k^2 & -a_k\delta_k\gamma_k \\ -a_k\gamma_k & -a_k\delta_k\gamma_k & a_k\gamma_k^2 \end{pmatrix}}_{S_k:=} \begin{pmatrix} \mathbf{w}_3 \\ \mathbf{z}_k - \mathbf{w}_3 \\ \mathbf{g}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{w}_3 \\ \mathbf{z}_k - \mathbf{w}_3 \\ \mathbf{g}_3 \end{pmatrix}^\top S_k \begin{pmatrix} \mathbf{w}_3 \\ \mathbf{z}_k - \mathbf{w}_3 \\ \mathbf{g}_3 \end{pmatrix},$$

where $\operatorname{Rank} S_k = 1$, $S_k \succeq 0$ and $S_k[1,1] = a_k$. Now, one can pick $S_k \succeq 0$ as new variable, drop the rank constraint, and keep constraining $S_k[1,1] = a_k$. There is always a feasible solution to the LMI with $\operatorname{Rank} S_k = 1$ if the LMI is feasible (S_k intervenes only on the side $. \preceq 0$), leading to

$$V_{k+1} = S_k + \begin{pmatrix} \mathbf{w}_2 - \mathbf{w}_\star & \mathbf{g}_2 \end{pmatrix} Q_k \begin{pmatrix} \mathbf{w}_2^\top - \mathbf{w}_\star^\top \\ \mathbf{g}_2^\top \end{pmatrix}$$

in the previous LMI, under the additional constraints $S_k \succeq 0$ and $S_k[1,1] = a_k$. One can then recover $\delta_k = \frac{S_k[1,2]}{S_k[1,1]}$ and $\gamma_k = -\frac{S_k[1,3]}{S_k[1,1]}$, along with $\tau_k = -\frac{\mu_2}{\mu_1}$ and $\alpha_k = \frac{\mu_4}{\mu_3}$.

Appendix E. Stochastic gradients under bounded variance

The proofs of this section follow the same ideas as that of App. D: we only reformulate weighted sums of inequalities.

E.1. Proof of Theorem 5

Proof Combine the following inequalities with their corresponding weights.

- Convexity between $x_\star$ and x_k with weight $\lambda_1 = \delta_k L$

$$f_\star \geq f(x_k) + \langle f'(x_k); x_\star - x_k \rangle,$$

- averaged smoothness between x_k and x_{k+1} with weight $\lambda_2 = d_{k+1} = d_k + \delta_k L$

$$\mathbb{E}_{i_k} f(x_{k+1}^{(i_k)}) \leq f(x_k) + \mathbb{E}_{i_k} \langle f'(x_k); x_{k+1}^{(i_k)} - x_k \rangle + \tfrac{L}{2} \mathbb{E}_{i_k} \|x_{k+1}^{(i_k)} - x_k\|^2,$$

- bounded variance at x_k with weight $\lambda_3 = e_k = \frac{\delta_k^2 L}{2}(1 + d_{k+1})$

$$\mathbb{E}_{i_k} \|G(x_k; i_k) - f'(x_k)\|^2 \leq \sigma^2.$$

The weighted sum can be reformulated as

$$\begin{aligned}
0 \geq & \lambda_1 \left[f(x_k) - f_\star + \langle f'(x_k); x_\star - x_k \rangle \right] \\
& + \lambda_2 \left[f(x_k) - \mathbb{E}_{i_k} f(x_{k+1}^{(i_k)}) + \mathbb{E}_{i_k} \langle f'(x_k); x_{k+1}^{(i_k)} - x_k \rangle + \tfrac{L}{2} \mathbb{E}_{i_k} \|x_{k+1}^{(i_k)} - x_k\|^2 \right] \\
& + \lambda_3 \left[\mathbb{E}_{i_k} \|G(x_k; i_k) - f'(x_k)\|^2 - \sigma^2 \right] \\
= & d_{k+1} \mathbb{E}_{i_k} [f(x_{k+1}^{(i_k)}) - f_\star] + \tfrac{L}{2} \mathbb{E}_{i_k} \|x_{k+1}^{(i_k)} - x_\star\|^2 \\
& - d_k (f(x_k) - f_\star) - \tfrac{L}{2} \|x_k - x_\star\|^2 - e_k \sigma^2 + (\delta_k d_{k+1} - e_k) \|f'(x_k)\|^2,
\end{aligned}$$

where rearrangment of the last inequality leads to

$$\begin{aligned}
& d_{k+1} \mathbb{E}_{i_k} [f(x_{k+1}^{(i_k)}) - f_\star] + \tfrac{L}{2} \mathbb{E}_{i_k} \|x_{k+1}^{(i_k)} - x_\star\|^2 \\
& \quad \leq d_k (f(x_k) - f_\star) + \tfrac{L}{2} \|x_k - x_\star\|^2 + e_k \sigma^2 - (\delta_k d_{k+1} - e_k) \|f'(x_k)\|^2 \\
& \quad \leq d_k (f(x_k) - f_\star) + \tfrac{L}{2} \|x_k - x_\star\|^2 + e_k \sigma^2,
\end{aligned}$$

as soon as $\delta_k d_{k+1} - e_k \geq 0$. ■

E.2. Proof of Theorem 6

Proof Combine the following inequalities with their corresponding weights.

- Smoothness and convexity between x_k and $x_\star$ with weight $\lambda_1 = \delta_k L$

$$f_\star \geq f(x_k) + \langle f'(x_k); x_\star - x_k \rangle + \tfrac{1}{2L} \|f'(x_k)\|^2,$$

- $\diamond$ averaged smoothness and convexity between x_k and $z_{k+1}^{(i_k)}$ with weight $\lambda_2 = \delta_k L$

$$f(x_k) \geq \mathbb{E}_{i_k} f(z_{k+1}^{(i_k)}) + \mathbb{E}_{i_k} \langle f'(z_{k+1}^{(i_k)}); x_k - z_{k+1}^{(i_k)} \rangle + \tfrac{1}{2L} \mathbb{E}_{i_k} \| f'(x_k) - f'(z_{k+1}^{(i_k)}) \|^2,$$

- $\diamond$ averaged smoothness and convexity between z_k and $z_{k+1}^{(i_k)}$ with weight $\lambda_3 = d_k \delta_k L$

$$f(z_k) \geq \mathbb{E}_{i_k} f(z_{k+1}^{(i_k)}) + \mathbb{E}_{i_k} \langle f'(z_{k+1}^{(i_k)}); z_k - z_{k+1}^{(i_k)} \rangle + \tfrac{1}{2L} \mathbb{E}_{i_k} \| f'(z_k) - f'(z_{k+1}^{(i_k)}) \|^2,$$

- $\diamond$ bounded variance at x_k with weight $\lambda_4 = e_k = \frac{\delta_k^2 L}{2} \frac{1+d_k+L\delta_k}{1+d_k}$

$$\mathbb{E}_{i_k} \| G(x_k; i_k) - f'(x_k) \|^2 \leq \sigma^2.$$

The weighted sum can be reformulated as

$$\begin{aligned}
0 \geq & \lambda_1 \left[f(x_k) - f_\star + \langle f'(x_k); x_\star - x_k \rangle + \tfrac{1}{2L} \| f'(x_k) \|^2 \right] \\
& + \lambda_2 \left[\mathbb{E}_{i_k} f(z_{k+1}^{(i_k)}) - f(x_k) + \mathbb{E}_{i_k} \langle f'(z_{k+1}^{(i_k)}); x_k - z_{k+1}^{(i_k)} \rangle + \tfrac{1}{2L} \mathbb{E}_{i_k} \| f'(x_k) - f'(z_{k+1}^{(i_k)}) \|^2 \right] \\
& + \lambda_3 \left[\mathbb{E}_{i_k} f(z_{k+1}^{(i_k)}) - f(z_k) + \mathbb{E}_{i_k} \langle f'(z_{k+1}^{(i_k)}); z_k - z_{k+1}^{(i_k)} \rangle + \tfrac{1}{2L} \mathbb{E}_{i_k} \| f'(z_k) - f'(z_{k+1}^{(i_k)}) \|^2 \right] \\
& + \lambda_4 \left[\mathbb{E}_{i_k} \| G(x_k; i_k) - f'(x_k) \|^2 - \sigma^2 \right] \\
= & \delta_k d_{k+1} L \mathbb{E}_{i_k} [f(z_{k+1}^{(i_k)}) - f_\star] + \tfrac{L}{2} \mathbb{E}_{i_k} \| x_{k+1}^{(i_k)} - x_\star \|^2 - \delta_k d_k L (f(z_k) - f_\star) - \tfrac{L}{2} \| x_k - x_\star \|^2 - e_k \sigma^2 \\
& + \tfrac{(d_k+1)\delta_k}{2} \mathbb{E}_{i_k} \| (\tfrac{1}{d_k+1} - 1) f'(z_k) - \tfrac{1}{d_k+1} f'(x_k) + \tfrac{\delta_k L}{d_k+1} G(x_k; i_k) + f'(z_{k+1}^{(i_k)}) \|^2 \\
& + \tfrac{d_k \delta_k}{2(d_k+1)} \| f'(z_k) + (\delta_k L - 1) f'(x_k) \|^2 \\
& + \tfrac{\delta_k (1 + \delta_k L (1 - \delta_k L))}{2} \| f'(x_k) \|^2.
\end{aligned}$$

Again, rearranging the terms leads to:

$$\begin{aligned}
\delta_k d_{k+1} & L \mathbb{E}_{i_k} [f(z_{k+1}^{(i_k)}) - f_\star] + \tfrac{L}{2} \mathbb{E}_{i_k} \| x_{k+1}^{(i_k)} - x_\star \|^2 \\
& \leq \delta_k d_k L (f(z_k) - f_\star) + \tfrac{L}{2} \| x_k - x_\star \|^2 + e_k \sigma^2 \\
& \quad - \tfrac{(d_k+1)\delta_k}{2} \mathbb{E}_{i_k} \| (\tfrac{1}{d_k+1} - 1) f'(z_k) - \tfrac{1}{d_k+1} f'(x_k) + \tfrac{\delta_k L}{d_k+1} G(x_k; i_k) + f'(z_{k+1}^{(i_k)}) \|^2 \\
& \quad - \tfrac{d_k \delta_k}{2(d_k+1)} \| f'(z_k) + (\delta_k L - 1) f'(x_k) \|^2 - \tfrac{\delta_k (1 + \delta_k L (1 - \delta_k L))}{2} \| f'(x_k) \|^2 \\
& \leq \delta_k d_k L (f(z_k) - f_\star) + \tfrac{L}{2} \| x_k - x_\star \|^2 + e_k \sigma^2,
\end{aligned}$$

where the last inequality follows from $\delta_k L (1 - \delta_k L) + 1 \geq 0$ (i.e., $\delta_k \leq \frac{1+\sqrt{5}}{2L}$). ■

E.3. Proof of Theorem 7

Proof Combine the following inequalities with their corresponding weights.

- $\diamond$ Convexity and smoothness between y_{k+1} and $x_\star$ with weight $\lambda_1 = \delta_k L$

$$f_\star \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1} \rangle + \tfrac{1}{2L} \| f'(y_{k+1}) \|^2,$$

- ⋄ convexity and smoothness between y_{k+1} and y_k with weight $\lambda_2 = d_k$

$$f(y_k) \geq f(y_{k+1}) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1}) - f'(y_k)\|^2,$$

- ⋄ bounded variance at y_{k+1} with weight $\lambda_3 = e_k = \frac{\delta_k^2 L}{2}$

$$\mathbb{E}_{i_k}\|G(y_{k+1}; i_k) - f'(y_{k+1})\|^2 \leq \sigma^2.$$

The weighted sum can be reformulated as

$$\begin{aligned}
0 \geq& \lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1})\|^2\right] \\
&+ \lambda_2 \left[f(y_{k+1}) - f(y_k) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1}) - f'(y_k)\|^2\right] \\
&+ \lambda_3 \left[\mathbb{E}_{i_k}\|G(y_{k+1}; i_k) - f'(y_{k+1})\|^2 - \sigma^2\right] \\
=&(d_k + \delta_k L)[f(y_{k+1}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 - d_k(f(y_k) - f_\star) - \tfrac{L}{2}\|x_k - x_\star\|^2 - e_k\sigma^2 \\
&+ \tfrac{\delta_k}{2}(1 - \delta_k L)\|f'(y_{k+1})\|^2 + \tfrac{d_k}{2L}\|f'(y_{k+1}) - f'(y_k))\|^2.
\end{aligned}$$

Rearranging the terms allows obtaining:

$$\begin{aligned}
(d_k + \delta_k L)[f(y_{k+1}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2& \\
\leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k\sigma^2& \\
\quad - \tfrac{\delta_k}{2}(1 - \delta_k L)\|f'(y_{k+1})\|^2 - \tfrac{d_k}{2L}\|f'(y_{k+1}) - f'(y_k))\|^2&, \\
\leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k\sigma^2&,
\end{aligned}$$

where the last inequality follows from $\delta_k \leq \frac{1}{L}$. ■

E.4. Evaluation of the stochastic gradient at the true averaged iterate

The target algorithm we study here is as follows

$$\begin{aligned}
y_{k+1} &= \tfrac{k}{k+1}y_k + \tfrac{1}{k+1}x_k, \\
x_{k+1}^{(i_k)} &= x_{k+1} - \delta_k G(y_{k+1}; i_k).
\end{aligned}$$

Theorem 12 *Consider the following iterative scheme*

$$\begin{aligned}
y_{k+1} &= \tfrac{d_k}{d_k+1}y_k + \tfrac{1}{d_k+1}x_k, \\
x_{k+1}^{(i_k)} &= x_{k+1} - \delta_k G(y_{k+1}; i_k),
\end{aligned}$$

for some $d_k \geq 0$ *and* $0 \leq \delta_k \leq \frac{1}{L}$. *Assuming* $f \in \mathcal{F}_L$ *the following inequality holds*

$$d_{k+1}\delta_k L[f(y_{k+1}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq d_k\delta_k L(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k\sigma^2,$$

with $d_{k+1} = d_k + 1$ *and* $e_k = \frac{L\delta_k^2}{2}$.

Proof Combine the following inequalities with their corresponding weights.

- $\diamond$ Convexity and smoothness between y_{k+1} and $x_\star$ with weight $\lambda_1 = \delta_k L$

$$f_\star \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1})\|^2,$$

- $\diamond$ convexity and smoothness between y_{k+1} and y_k with weight $\lambda_2 = d_k\delta_k L$

$$f(y_k) \geq f(y_{k+1}) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1}) - f'(y_k)\|^2,$$

- $\diamond$ bounded variance at y_{k+1} with weight $\lambda_3 = e_k = \frac{\delta_k^2 L}{2}$

$$\mathbb{E}_{i_k}\|G(y_{k+1}; i_k) - f'(y_{k+1})\|^2 \leq \sigma^2.$$

The corresponding weighted sum can be reformulated as

$$\begin{aligned}
0 \geq &\lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1})\|^2\right] \\
&+ \lambda_2 \left[f(y_{k+1}) - f(y_k) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle + \tfrac{1}{2L}\|f'(y_{k+1}) - f'(y_k)\|^2\right] \\
&+ \lambda_3 \left[\mathbb{E}_{i_k}\|G(y_{k+1}; i_k) - f'(y_{k+1})\|^2 - \sigma^2\right] \\
=&(d_k+1)\delta_k L(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 - d_k\delta_k L(f(y_k) - f_\star) - \tfrac{L}{2}\|x_k - x_\star\|^2 - e_k\sigma^2 \\
&+ \tfrac{\delta_k}{2}(1 - \delta_k L)\|f'(y_{k+1})\|^2 + \tfrac{\delta_k d_k}{2}\|f'(y_k) - f'(y_{k+1})\|^2.
\end{aligned}$$

After rearranging the terms, we reach:

$$\begin{aligned}
(d_k+1)\delta_k L(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 - d_k\delta_k \\
\leq d_k\delta_k L(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k\sigma^2 \\
- \tfrac{\delta_k}{2}(1 - \delta_k L)\|f'(y_{k+1})\|^2 - \tfrac{\delta_k d_k}{2}\|f'(y_k) - f'(y_{k+1})\|^2 \\
\leq d_k\delta_k L(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + e_k\sigma^2
\end{aligned}$$

where the last inequality follows from $\delta_k \leq \frac{1}{L}$. ■

E.5. Convergence rates

In this section, we use the following two facts (they can easily be recovered by upper and lower bounding the sums with appropriate integrals)

1. $\sum_{t=1}^{N} t^{-\alpha} = O(N^{1-\alpha})$ for $\alpha \neq 1$,
2. $\sum_{t=1}^{N} t^{-1} = O(\log N)$.

for obtaining asymptotic rates for the previous methods when $\delta_k = (L(1+k)^\alpha)^{-1}$.

Stochastic gradient descent. From Theorem 5, we have

$$(d_k+\delta_k L)\mathbb{E}_{i_k}[f(x_{k+1}^{(i_k)})-f_\star]+\tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)}-x_\star\|^2 \\ \leq d_k(f(x_k)-f_\star)+\tfrac{L}{2}\|x_k-x_\star\|^2+\tfrac{\delta_k^2 L}{2}(1+d_k+\delta_k L)\sigma^2.$$

The choice $d_0=0$ leads to

$$\left(\sum_{t=0}^{N-1} L\delta_t\right)\mathbb{E}(f(x_N)-f_\star)\leq \frac{L}{2}\|x_0-x_\star\|^2+\frac{\sigma^2}{2}\sum_{k=0}^{N-1}\left[L^2\delta_k^2\left(1+\sum_{t=0}^{k} L\delta_t\right)\right].$$

For the choice $\delta_k=(L(1+k)^\alpha)^{-1}$, the different terms behave as follows:

- ⋄ $\sum_{k=0}^{N-1} L\delta_k \sim N^{1-\alpha}$ when $\alpha\neq 1$,
- ⋄ $\sum_{k=0}^{N-1} L^2\delta_k^2 \sim N^{1-2\alpha}$ for $\alpha\neq 1/2$,
- ⋄ $\sum_{k=0}^{N-1}\left[L^2\delta_k^2\sum_{t=0}^{k} L\delta_t\right]\sim \sum_{k=0}^{N-1} k^{1-3\alpha}\sim N^{2-3\alpha}$ when $\alpha\neq 1$ and $\alpha\neq 2/3$,

Under the same restrictions on α, we also get:

$$\left(\sum_{k=0}^{N-1} L\delta_k\right)^{-1}\sim N^{\alpha-1},\quad \frac{\sum_{k=0}^{N-1} L^2\delta_k^2}{\sum_{t=0}^{N-1} L\delta_t}\sim N^{-\alpha},\quad \frac{\sum_{k=0}^{N-1}\left[L^2\delta_k^2\sum_{t=0}^{k} L\delta_t\right]}{\sum_{k=0}^{N-1} L\delta_k}\sim N^{1-2\alpha},$$

and hence

$$\mathbb{E}f(x_N)-f_\star\leq \frac{L}{2}\|x_0-x_\star\|^2\,O(N^{\alpha-1})+\frac{\sigma^2}{2}\left[O\left(N^{-\alpha}\right)+O\left(N^{1-2\alpha}\right)\right].$$

Stochastic gradient descent with averaging. From Theorem 6, we have

$$(d_k+1)L\delta_k\mathbb{E}_{i_k}[f(z_{k+1}^{(i_k)})-f_\star]+\tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)}-x_\star\|^2 \\ \leq d_k L\delta_k(f(z_k)-f_\star)+\tfrac{L}{2}\|x_k-x_\star\|^2+\tfrac{\delta_k^2}{2}\tfrac{L(1+d_k+L\delta_k)}{1+d_k}\sigma^2.$$

Let us choose $d_0=0$ and define $d_{k+1}:=\frac{\delta_k}{\delta_{k+1}}(d_k+1)$, $D_k:=L\delta_{k-1}d_{k-1}+L\delta_{k-1}$ (with $D_0:=0$); one can write

$$D_{k+1}\mathbb{E}_{i_k}[f(z_{k+1}^{(i_k)})-f_\star]+\tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)}-x_\star\|^2\leq D_k(f(z_k)-f_\star)+\tfrac{L}{2}\|x_k-x_\star\|^2+\tfrac{\delta_k^2}{2}\tfrac{L(D_{k+1}+L^2\delta_k^2)}{D_{k+1}}\sigma^2.$$

In addition, we get $D_{k+1}=L\delta_k d_k+L\delta_k=L\delta_k\frac{\delta_{k-1}}{\delta_k}(d_{k-1}+1)+L\delta_k=D_k+L\delta_k=\sum_{t=0}^{k} L\delta_t$, and arrive to the final guarantee

$$D_N\mathbb{E}[f(z_N)-f_\star]\leq \tfrac{L}{2}\|x_0-x_\star\|^2+\tfrac{\sigma^2}{2L}\sum_{k=1}^{N}\delta_{k-1}^2 L^2\tfrac{D_k+L^2\delta_{k-1}^2}{D_k},$$

from which one can arrive to the following rates:

$$\mathbb{E}f(z_N)-f_\star\leq \tfrac{L}{2}\|x_0-x_\star\|^2\,O(N^{\alpha-1})+\tfrac{\sigma^2}{2L}[O(N^{-\alpha})+O(N^{-1-2\alpha})],$$

where we used the following estimates of the rates for the different terms:

- ◇ $D_N = \sum_{k=0}^{N-1} L\delta_k \sim N^{1-\alpha}$ when $\alpha \neq 1$,
- ◇ $\sum_{k=1}^{N} L^2\delta_{k-1}^2 \sim N^{1-2\alpha}$ when $\alpha \neq 1/2$,
- ◇ $\sum_{k=1}^{N} \frac{L^4\delta_{k-1}^4}{D_k} \sim \sum_{k=1}^{N} k^{-1-3\alpha} \sim N^{-3\alpha}$ when $\alpha \neq 1$ and $\alpha \neq 0$,

and under the same restrictions on α:

$$D_N^{-1} \sim N^{\alpha-1}, \quad \frac{\sum_{k=1}^{N} L^2\delta_{k-1}^2}{D_N} \sim N^{-\alpha}, \quad \frac{\sum_{k=1}^{N} \frac{L^4\delta_{k-1}^4}{D_k}}{D_N} \sim N^{-1-2\alpha}.$$

Stochastic gradient descent with primal averaging. From Theorem 7, we have

$$(d_k + \delta_k L)[f(y_{k+1}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 + \tfrac{L\delta_k^2}{2}\sigma^2,$$

hence choosing $d_0 = 0$ and using the same estimates as before, we get

$$\mathbb{E}f(y_N) - f_\star \leq \tfrac{L}{2}\|x_0 - x_\star\|^2 O(N^{\alpha-1}) + \tfrac{\sigma^2}{2L} O(N^{-\alpha}).$$

E.6. Better worst-case guarantees and rates

Different techniques can be used for obtaining stochastic methods with improved worst-case bounds. Among others, one can (i) assume the number of iterations to be fixed in advance, (ii) assume the domain to be compact, or (iii) use more past information — essentially through use of a *dual averaging* scheme. Algorithms using (iii) do not directly fit within our framework, which assumes (very) limited memory through (SFO)—we believe it could be adapted, but let it for further investigations. The previous points (i) and (ii) are used by Lan (2012), point (ii) is used e.g., by Hu et al. (2009). Point (iii) is used by Devolder (2011, Section 6) and by Xiao (2010, Section 7.1 and Appendix D).

In this section, we briefly discuss two ways of improving the complexity results using namely techniques (i) and an alternative to (iii) for using more past information by adding a new sequence to (SFO). We start by assuming a known number of iterations N. In this setting, we get the following results.

Fixed number of iterations. Consider a stochastic gradient scheme with primal averaging and a constant step-size $\delta_k = \delta$. Denoting $R = \|x_0 - x_\star\|$ and choosing $\delta = \frac{1}{\sqrt{N}\frac{\sigma}{R} + L}$, we get

$$\mathbb{E}f(y_N) - f_\star \leq \tfrac{R^2}{2\delta N} + \tfrac{\delta\sigma^2}{2} = \tfrac{L^2R^3}{2LRN + 2N^{3/2}\sigma} + \tfrac{R\sigma}{\sqrt{N}} \leq \tfrac{LR^2}{2N} + \tfrac{R\sigma}{\sqrt{N}}$$

(note though that the optimal step-size in that setting is $\delta = \frac{R}{\sqrt{N}\sigma}$ (when $\frac{R}{\sqrt{N}\sigma} \leq \frac{1}{L}$ following the statement of Theorem 7), leading to $\mathbb{E}f(y_N) - f_\star \leq \frac{R\sigma}{\sqrt{N}}$).

Dual averaging. Through the use of the dual averaging technique (iii) Devolder (2011, Section 6) and Xiao (2010, Section 7.1 and Appendix D), one can observe that x_0 plays an important role as its weight grows in the dual averaging process (through coefficients β_i's in both (Devolder, 2011) and (Xiao, 2010)). In a certain sense, one can interpret x_0 as a *magnet* or *anchor* whose goal is to stabilize (and slow down) the iterative process. As such, our framework does not allow treating estimate sequences with the damping with x_0 (recall that the whole point of potential-based proofs is to forget how the current iterate was obtained). Still, let us have a glance at the answers provided by the framework without dual averaging.

Momentum without dual averaging nor damping. Using the line-search design procedure, we obtain the following methods that appears to be symptomatic of all accelerated methods without damping so far: a huge error accumulation. This phenomenon is not surprising nor new. Using the previous parameter selection technique, we arrive to the following theorem, which establishes the (negative and non-surprising) upper bound obtained through our framework, the main message being that the accumulation term $\sum_{k=0}^{N-1} e_k = O(N^3)$ in both cases presented below (whereas $d_N = O(N^2)$) can hardly be avoided while using *pure acceleration* without damping—usually achieved by appropriate dual averaging schemes (Xiao, 2010; Devolder, 2011).

Theorem 13 *Let $k \in \mathbb{N}$, $d_k, d_{k+1} \geq 0$, $f \in \mathcal{F}_L$ and the algorithms*

$$\begin{aligned} y_{k+1} &= \mathrm{argmin}_x \left\{ f(x) \textit{ subject to } x \in x_k + \mathrm{span}\{z_k - x_k\} \right\}, \\ x_{k+1}^{(i_k)} &= \mathrm{argmin}_x \left\{ f(x) \textit{ subject to } x \in y_{k+1} + \mathrm{span}\{G(y_{k+1}; i_k)\} \right\}, \\ z_{k+1}^{(i_k)} &= z_k - \tfrac{d_{k+1}-d_k}{L} G(y_{k+1}; i_k), \end{aligned}$$

and

$$\begin{aligned} y_{k+1} &= x_k + \left(1 - \tfrac{d_k}{d_{k+1}}\right)(z_k - x_k), \\ x_{k+1}^{(i_k)} &= y_{k+1} - \tfrac{\eta_k}{d_{k+1}} G(y_{k+1}; i_k), \\ z_{k+1}^{(i_k)} &= z_k - \tfrac{d_{k+1}-d_k}{L} G(y_{k+1}; i_k). \end{aligned}$$

In both cases, we have

$$d_{k+1} \mathbb{E}_{i_k}(f(x_{k+1}^{(i_k)}) - f_\star) + \tfrac{L}{2} \mathbb{E}_{i_k} \|z_{k+1}^{(i_k)} - x_\star\|^2 \leq d_k (f(x_k) - f_\star) + \tfrac{L}{2} \|z_k - x_\star\|^2 + e_k \sigma^2$$

for all choices of $e_k, \eta_k \geq 0$ satisfying both $\frac{d_{k+1}-d_k}{2L} - e_k + \eta_k \geq 0$ and $2 d_{k+1} e_k L - d_{k+1}(d_{k+1} - d_k)^2 - \eta_k^2 L^2 \geq 0$. In particular, the following three choices are valid:

- *$d_k = \frac{k(k+1)}{4}$ for all $k \geq 0$ together with $\eta_k = \frac{(k+1)(k+2)}{4L}$ and $e_k = \frac{(k+1)(k+3)}{4L}$, or*
- *$d_k = \frac{k(k+1)}{4}$ for all $k \geq 0$ together with the slightly better $\eta_k = \frac{(k+1)\left(k+2-\sqrt{3}\sqrt{k+2}\right)}{4L}$, $e_k = \frac{(k+1)\left(k+3-\sqrt{3}\sqrt{k+2}\right)}{4L}$, or*
- *(primal averaging) $d_{k+1} = d_k + \delta_k L$ together with $\eta_k = 0$, $e_k = \frac{L\delta_k^2}{2}$ when $\delta_k \leq \frac{1}{L}$.*

Proof Perform a weighted sum of the following inequalities:

- smoothness and convexity between $x_\star$ and y_{k+1} with weight $\lambda_1 = d_{k+1} - d_k$

$$f_\star \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1} \rangle + \tfrac{1}{2L} \|f'(y_{k+1})\|^2,$$

- averaged smoothness and convexity between y_{k+1} and $x_{k+1}^{(i_k)}$ with weight $\lambda_2 = d_{k+1}$

$$f(y_{k+1}) \geq \mathbb{E}_{i_k} f(x_{k+1}^{(i_k)}) + \mathbb{E}_{i_k} \langle f'(x_{k+1}^{(i_k)}); y_{k+1} - x_{k+1}^{(i_k)} \rangle + \tfrac{1}{2L} \mathbb{E}_{i_k} \|f'(y_{k+1}) - f'(x_{k+1}^{(i_k)})\|^2,$$

- convexity between x_k and y_{k+1} with weight $\lambda_3 = d_k$

$$f(x_k) \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_k - y_{k+1} \rangle,$$

◇ first line-search optimality condition of y_{k+1} with weight $\lambda_4 = d_{k+1}$

$$\langle f'(y_{k+1}); x_k - y_{k+1}\rangle \geq 0,$$

◇ second line-search optimality condition of y_{k+1} with weight $\lambda_5 = d_{k+1} - d_k$

$$\langle f'(y_{k+1}); z_k - x_k\rangle \geq 0,$$

◇ first (averaged) line-search optimality condition for $x_{k+1}^{(i_k)}$ with weight $\lambda_6 = d_{k+1}$

$$\mathbb{E}_{i_k}\langle f'(x_{k+1}^{(i_k)}); y_{k+1} - x_{k+1}^{(i_k)}\rangle \geq 0,$$

◇ second (averaged) line-search optimality condition for $x_{k+1}^{(i_k)}$ with weight $\lambda_7 = \eta_k$:

$$\mathbb{E}_{i_k}\langle f'(x_{k+1}^{(i_k)}); G(y_{k+1}; i_k)\rangle \leq 0,$$

◇ bounded variance at y_{k+1} with weight $\lambda_8 = e_k$

$$\mathbb{E}_{i_k}\|G(y_{k+1}; i_k) - f'(y_{k+1})\|^2 \leq \sigma^2.$$

Now, the weighted sum corresponds to

$$\begin{aligned}
&d_{k+1}\mathbb{E}_{i_k}(f(x_{k+1}^{(i_k)}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|z_{k+1}^{(i_k)} - x_\star\|^2 \\
&\quad\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2 + e_k\sigma^2 \\
&\qquad - \tfrac{d_{k+1}}{2L}\mathbb{E}_{i_k}\|\tfrac{\eta_k L}{d_{k+1}}G(y_{k+1}; i_k) + f'(x_{k+1}^{(i_k)}) - f'(y_{k+1})\|^2 \\
&\qquad - \left(\tfrac{d_{k+1}-d_k}{2L} - e_k + \eta_k\right)\|f'(y_{k+1})\|^2 \\
&\qquad - \tfrac{2d_{k+1}e_kL - d_{k+1}(d_{k+1}-d_k)^2 - \eta_k^2L^2}{2d_{k+1}L}\mathbb{E}_{i_k}\|G(y_{k+1}; i_k)\|^2,
\end{aligned}$$

therefore, this weighted sum corresponds to the statement of the theorem when the three conditions are met:

$$\begin{aligned}
\tfrac{d_{k+1}}{2L} &\geq 0, \\
\left(\tfrac{d_{k+1}-d_k}{2L} - e_k + \eta_k\right) &\geq 0, \\
\tfrac{2d_{k+1}e_kL - d_{k+1}(d_{k+1}-d_k)^2 - \eta_k^2L^2}{2d_{k+1}L} &\geq 0,
\end{aligned}$$

which are easy to check in the case of primal averaging. Another possible choice is to set $e_k = \frac{d_{k+1}-d_k}{2L} + \eta_k$ and η_k such that $\frac{2d_{k+1}e_kL - d_{k+1}(d_{k+1}-d_k)^2 - \eta_k^2L^2}{2d_{k+1}L} \geq 0$; in particular, when $d_k = \frac{k(k+1)}{4}$ for all $k \geq 0$ we get

$$\eta_k \in \left[\tfrac{(k+1)\left(k+2-\sqrt{3}\sqrt{k+2}\right)}{4L}, \tfrac{(k+1)\left(k+2+\sqrt{3}\sqrt{k+2}\right)}{4L}\right],$$

a possible choice is $\eta_k = \frac{(k+1)(k+2)}{4L}$, resulting in a critical accumulation in the variance term

$$\begin{aligned}
&d_{k+1}\mathbb{E}_{i_k}(f(x_{k+1}^{(i_k)}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|z_{k+1}^{(i_k)} - x_\star\|^2\\
&\quad\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2 + \tfrac{(k+1)(k+3)}{4L}\sigma^2\\
&\qquad - \tfrac{(k+1)(k+2)}{8L}\mathbb{E}_{i_k}\|f'(x_{k+1}^{(i_k)}) - f'(y_{k+1}) + G(y_{k+1}; i_k)\|^2\\
&\qquad - \tfrac{3(k+1)}{8L}\mathbb{E}_{i_k}\|G(y_{k+1}; i_k)\|^2\\
&\quad\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2 + \tfrac{(k+1)(k+3)}{4L}\sigma^2,
\end{aligned}$$

where the accumulation arrives to $\frac{\sigma^2}{4L}\sum_{k=0}^{N-1}(k+1)(k+3) = O(N^3)$. It is also possible to choose the better (smaller) $\eta_k = \frac{(k+1)\left(k+2-\sqrt{3}\sqrt{k+2}\right)}{4L}$ leading to

$$\begin{aligned}
&d_{k+1}\mathbb{E}_{i_k}(f(x_{k+1}^{(i_k)}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|z_{k+1}^{(i_k)} - x_\star\|^2\\
&\quad\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2 + \tfrac{(k+1)\left(k+3-\sqrt{3}\sqrt{k+2}\right)}{4L}\sigma^2\\
&\qquad - \tfrac{(k+1)(k+2)}{8L}\mathbb{E}_{i_k}\|f'(x_{k+1}^{(i_k)}) - f'(y_{k+1}) + \left(1 - \tfrac{\sqrt{3}}{\sqrt{k+2}}\right)G(y_{k+1}; i_k)\|^2\\
&\quad\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|^2 + \tfrac{(k+1)\left(k+3-\sqrt{3}\sqrt{k+2}\right)}{4L}\sigma^2,
\end{aligned}$$

but the result remains philosophically the same. The proof is valid for all methods proposed above as they all satisfy both conditions

$$\begin{aligned}
&\langle f'(y_{k+1}); d_{k+1}(x_k - y_k) + (d_{k+1} - d_k)(z_k - x_k)\rangle \geq 0,\\
&\mathbb{E}_{i_k}\langle f'(x_{k+1}^{(i_k)}); d_{k+1}(y_{k+1} - x_{k+1}^{(i_k)}) - \eta_k G(y_{k+1}; i_k)\rangle \geq 0.
\end{aligned}$$

■

E.7. Linear matrix inequalities

In this section, we provide the linear matrix inequalities for the basic stochastic gradient method $x_{k+1}^{(i_k)} = x_k - \delta_k G(x_k; i_k)$ with the following family of potentials

$$\phi_k^f = \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ f'(x_k) \end{pmatrix} + d_k\,(f(x_k) - f_\star). \tag{21}$$

with $Q_k \in \mathbb{S}^2$. The following lines can serve for reproducing Figure 1; the code implementing the following lines is provided in Sec. 5.

We use the following matrices P and F for encoding (stochastic) gradients and function values:

$$\begin{aligned}
P &= [\ x_k \mid G(x_k; 1)\ \dots\ G(x_k; n) \mid f'(x_{k+1}^{(1)})\ \dots\ f'(x_{k+1}^{(n)})\ ]\\
F &= [\ f(x_k) \mid f(x_{k+1}^{(1)})\ \dots\ f(x_{k+1}^{(n)})\ ].
\end{aligned}$$

Define the following vectors for selecting entries in P and F: $\mathbf{x}_\star, \mathbf{x}_0,\ \mathbf{x}_1,\ \dots,\ \mathbf{x}_n, \mathbf{G}_1,\ \dots,\ \mathbf{G}_n$, $\mathbf{g}_\star,\ \mathbf{g}_0,\ \mathbf{g}_1,\ \dots,\ \mathbf{g}_n \in \mathbb{R}^{1+2n}$ and $\mathbf{f}_\star,\ \mathbf{f}_0, \mathbf{f}_1,\ \dots,\ \mathbf{f}_n \in \mathbb{R}^{1+n}$ such that

$$\begin{aligned}
x_k = P\mathbf{x}_0,\ f'(x_k) = P\mathbf{g}_0,\ f(x_k) = F\mathbf{f}_0,\ G(x_k; i) = P\mathbf{G}_i,\\
x_{k+1}^{(i)} = P\mathbf{x}_i,\ f'(x_{k+1}^{(i)}) = P\mathbf{g}_i,\ f(x_{k+1}^{(i)}) = P\mathbf{f}_i,
\end{aligned}$$

for $i \in \{1, \ldots, n\}$ and $x_\star = P\mathbf{x}_\star,\ g_\star = P\mathbf{g}_\star,\ f_\star = F\mathbf{f}_\star$. More precisely we choose the following vectors in $\mathbb{R}^{1+2n}$

$$\begin{aligned}
&\mathbf{x}_0 := e_1, \quad \mathbf{G}_i := e_{1+i} \quad (\text{for } i \in \{1, \ldots, n\}), \quad \mathbf{g}_0 := \tfrac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i,\\
&\mathbf{x}_i := \mathbf{x}_0 - \delta_k \mathbf{G}_i \quad (\text{for } i \in \{1, \ldots, n\}), \quad \mathbf{g}_i := e_{1+n+i} \quad (\text{for } i \in \{1, \ldots, n\}),\\
&\mathbf{x}_\star := 0, \quad \mathbf{g}_\star := 0,
\end{aligned}$$

and the following vectors in $\mathbb{R}^{n+1}$: $\mathbf{f}_0 := e_1$, $\mathbf{f}_i := e_{1+i}$ (for $i \in \{1, \ldots, n\}$), $\mathbf{f}_\star = 0$. We also encode variance using a matrix A_{var} such that

$$\operatorname{Trace}(A_{\text{var}} P^\top P) = \mathbb{E}_i \|G(x_k; i) - f'(x_k)\|^2,$$

in others words $A_{\text{var}} := \frac{1}{n}\sum_{i=1}^{n}(\mathbf{G}_i - \mathbf{g}_0)(\mathbf{G}_i - \mathbf{g}_0)^\top$. We use $\mathbb{S}^{1+2n}$ to denote the set of $(1+2n) \times (1+2n)$ symmetric matrices, and use M as defined in (19).

Given two doublets (d_k, Q_k), (d_{k+1}, Q_{k+1}) and a step-size δ_k, the inequality $\mathbb{E}_{i_k} \phi_{k+1}^f(x_{k+1}^{(i_k)}) \leq \phi_k^f(x_k) + e_k \sigma^2$ (ϕ_k^f defined in (21)) holds for all $f \in \mathcal{F}_L(\mathbb{R}^d)$, all $d \in \mathbb{N}$, all $x_k \in \mathbb{R}^d$ and all sets $\{G(x_k; i)\}_{i=1,\ldots,n} \subset \mathbb{R}^d$ satisfying $\mathbb{E}_i G(x_k; i) = f'(x_k)$ and $\mathbb{E}_i \|G(x_k; i) - f'(x_k)\|^2 \leq \sigma^2$ used to generate $x_{k+1}^{(i_k)}$ with $x_{k+1}^{(i_k)} = x_k - \delta_k G(x_k; i_k)$ if and only if there exists $\{\lambda_{i,j}\}_{i,j \in I}$, with $I := \{\star, 0, 1, \ldots, n\}$ and $e_k \in \mathbb{R}$ such that

$$\begin{aligned}
&\lambda_{i,j} \geq 0 \text{ for all } i, j \in I \text{ and } e_k \geq 0,\\
&d_{k+1} \tfrac{1}{n} \sum_{i=1}^{n} (\mathbf{f}_i - \mathbf{f}_\star) - d_k(\mathbf{f}_0 - \mathbf{f}_\star) + \sum_{i,j \in I} \lambda_{i,j}(\mathbf{f}_i - \mathbf{f}_j) = 0 \text{ (linear constraint in } \mathbb{R}^{n+1}),\\
&\tfrac{1}{n} \sum_{i=1}^{n} V_{k+1}^{(i)} - V_k - e_k A_{\text{var}} + \sum_{i,j \in I} \lambda_{i,j} M_{i,j} \preceq 0 \text{ (linear matrix inequality in } \mathbb{S}^{1+2n}),
\end{aligned}$$

with

$$\begin{aligned}
V_k &:= \begin{pmatrix} \mathbf{x}_0 - \mathbf{x}_\star & \mathbf{g}_0 \end{pmatrix} Q_k \begin{pmatrix} \mathbf{x}_0^\top - \mathbf{x}_\star^\top \\ \mathbf{g}_0^\top \end{pmatrix} \in \mathbb{S}^{1+2n},\\
V_{k+1}^{(i)} &:= \begin{pmatrix} \mathbf{x}_i - \mathbf{x}_\star & \mathbf{g}_i \end{pmatrix} Q_{k+1} \begin{pmatrix} \mathbf{x}_i^\top - \mathbf{x}_\star^\top \\ \mathbf{g}_i^\top \end{pmatrix} \in \mathbb{S}^{1+2n},\\
M_{i,j} &:= \begin{pmatrix} \mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j \end{pmatrix} M \begin{pmatrix} \mathbf{x}_i & \mathbf{x}_j & \mathbf{g}_i & \mathbf{g}_j \end{pmatrix}^\top \in \mathbb{S}^{1+2n}.
\end{aligned}$$

Parameter selection. For adapting the strategy to our parameter selection technique, we refer to App. D.5, as the exact same tricks apply, though a bit more tedious in the stochastic setting.

Appendix F. Stochastic gradients for over-parameterized models

This section contains the proof of Theorem 8, a discussion on how we ended up with primal averaging using the parameter selection technique, and an LMI formulation for studying primal averaging.

F.1. Proof of Theorem 8

Proof Let us start with the case $\delta_k \leq \frac{1}{L}$ and $d_{k+1} = d_k + \delta_k L$. As before, the proof consists in linearly combining the following inequalities.

- ⋄ Averaged smoothness and convexity between y_{k+1} and $x_\star$ with weight $\lambda_1 = \delta_k L$:
$$f_\star \geq f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2,$$
- ⋄ convexity between x_{k+1} and $x_\star$ with weight $\lambda_2 = d_k$:
$$f(y_k) \geq f(y_{k+1}) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle.$$

The weighted sum of those inequalities can be rewritten as
$$\begin{aligned}0 \geq& \lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2\right]\\ &+ \lambda_2 \left[f(y_{k+1}) - f(y_k) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle\right]\\ =&(d_k + \delta_k L)(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x^{(i_k)}_{k+1} - x_\star\|^2 - d_k(f(y_k) - f_\star) - \tfrac{L}{2}\|x_k - x_\star\|^2\\ &+ \tfrac{\delta_k}{2}\left(1 - \delta_k L\right)\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2,\end{aligned}$$
which can be rearranged to
$$\begin{aligned}(d_k + \delta_k L)(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x^{(i_k)}_{k+1} - x_\star\|^2&\\ \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 - \tfrac{\delta_k}{2}\left(1 - \delta_k L\right)\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2&\\ \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2,&\end{aligned}$$
where the last inequality follows from $\delta_k \leq \frac{1}{L}$. In the case where $\delta_k \geq \frac{1}{L}$, we use instead $d_{k+1} = d_k + 2\delta_k L - \delta_k^2 L^2$ and compensate the nonpositive term in the righthand side by using an additional inequality:

- ⋄ averaged smoothness between y_{k+1} and $x_\star$ with weight $\lambda_3 = \delta_k L(\delta_k L - 1)$ (which is nonnegative as $0 \leq \delta_k \leq \frac{1}{L}$)
$$f(y_{k+1}) \geq f_\star + \tfrac{1}{2L}\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2.$$

Adding this inequality to the previous sum, we get:
$$\begin{aligned}0 \geq& \lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle + \tfrac{1}{2L}\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2\right]\\ &+ \lambda_2 \left[f(y_{k+1}) - f(y_k) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle\right]\\ &+ \lambda_3 \left[f_\star - f(y_{k+1}) + \tfrac{1}{2L}\mathbb{E}_{i_k}\|f'_{i_k}(y_{k+1})\|^2\right]\\ =&(d_k + 2\delta_k L - \delta_k^2 L^2)(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x^{(i_k)}_{k+1} - x_\star\|^2 - d_k(f(y_k) - f_\star) - \tfrac{L}{2}\|x_k - x_\star\|^2,\end{aligned}$$
which can be reformulated as
$$(d_k + 2\delta_k L - \delta_k^2 L^2)(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x^{(i_k)}_{k+1} - x_\star\|^2 \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2.$$

■

F.2. Parameter selection

In this section, we quickly discuss how we ended up with primal averaging for overparametrized models (see Theorem 8). Essentially, we used the first parameter selection technique from App. C, which we adapted to

$$\begin{aligned} y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\ x_{k+1}^{(i_k)} &= y_{k+1} - \alpha_k f'_{i_k}(y_{k+1}), \\ z_{k+1}^{(i_k)} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'_{i_k}(y_{k+1}), \end{aligned}$$

for which we wish to optimize the parameters $\{(\tau_k, \alpha_k, \delta_k, \gamma_k)\}_k$. Instead, we consider the line-search version

$$\begin{aligned} y_{k+1} &= \operatorname{argmin}_x \left\{ f(x) \text{ subject to } x \in x_k + \operatorname{span}\{z_k - x_k\}\right\}, \\ x_{k+1} &= \operatorname{argmin}_x \left\{ f(x) \text{ subject to } x \in y_{k+1} + \operatorname{span}\{f'_{i_k}(y_{k+1})\}\right\}, \\ z_{k+1} &= (1-\delta_k)y_{k+1} + \delta_k z_k - \gamma_k f'_{i_k}(y_{k+1}), \end{aligned}$$

along with the following family of potentials

$$\begin{aligned} \phi_k^f =& q_{1,k}\, \|x_k - x_\star\|^2 + q_{2,k}\, \|f'(x_k)\|^2 + q_{3,k}\, \mathbb{E}_i \|f'_i(x_k)\|^2 + q_{4,k}\, \langle f'(x_k); x_k - x_\star \rangle \\ &+ d_k\, (f(x_k) - f_\star) + a_k \|z_k - x_\star\|^2, \end{aligned}$$

which was chosen based on expected symmetries with respect to i_k. By picking $\phi_0^f = \frac{L}{2}\|x_0 - x_\star\|^2$ and $\phi_N^f = d_N\,(f(x_N) - f_\star)$, we naturally arrived to primal averaging by solving

$$\max_{\{(\delta_k,\gamma_k)\}_k} \quad \max_{\phi_1^f,\ldots,\phi_{N-1}^f, d_N} d_N \text{ subject to } (\phi_0^f, \phi_1^f) \in \tilde{\mathcal{V}}_0, \ldots, (\phi_{N-1}^f, \phi_N^f) \in \tilde{\mathcal{V}}_{N-1}, \tag{22}$$

as done in (8); the numerical results are shown on Figure 5.

F.3. Linear matrix inequalities for over-parametrized models

The derivations of the linear matrix inequalities follow from the exact same line as in the previous sections. As an example, we provide the LMI formulation for analyzing primal averaging:

$$\begin{aligned} y_{k+1} &= \tau_k y_k + (1-\tau_k)x_k, \\ x_{k+1}^{(i_k)} &= x_k - \gamma_k f'_{i_k}(y_{k+1}), \end{aligned} \tag{23}$$

along with the following family of potentials

$$\phi_k^f = \begin{pmatrix} x_k - x_\star \\ y_k - x_\star \\ f'_1(y_k) - f'_1(x_\star) \\ \vdots \\ f'_n(y_k) - f'_n(x_\star) \end{pmatrix}^\top [Q_k \otimes I_d] \begin{pmatrix} x_k - x_\star \\ y_k - x_\star \\ f'_1(y_k) - f'_1(x_\star) \\ \vdots \\ f'_n(y_k) - f'_n(x_\star) \end{pmatrix} + d_k\,(f(y_k) - f_\star), \tag{24}$$

which can be simplified using appropriate symmetry arguments, but we simply formulate the problem without using them for tutorial purposes, and simply choose $Q_k \in \mathbb{S}^{2+n}$. One could additionally

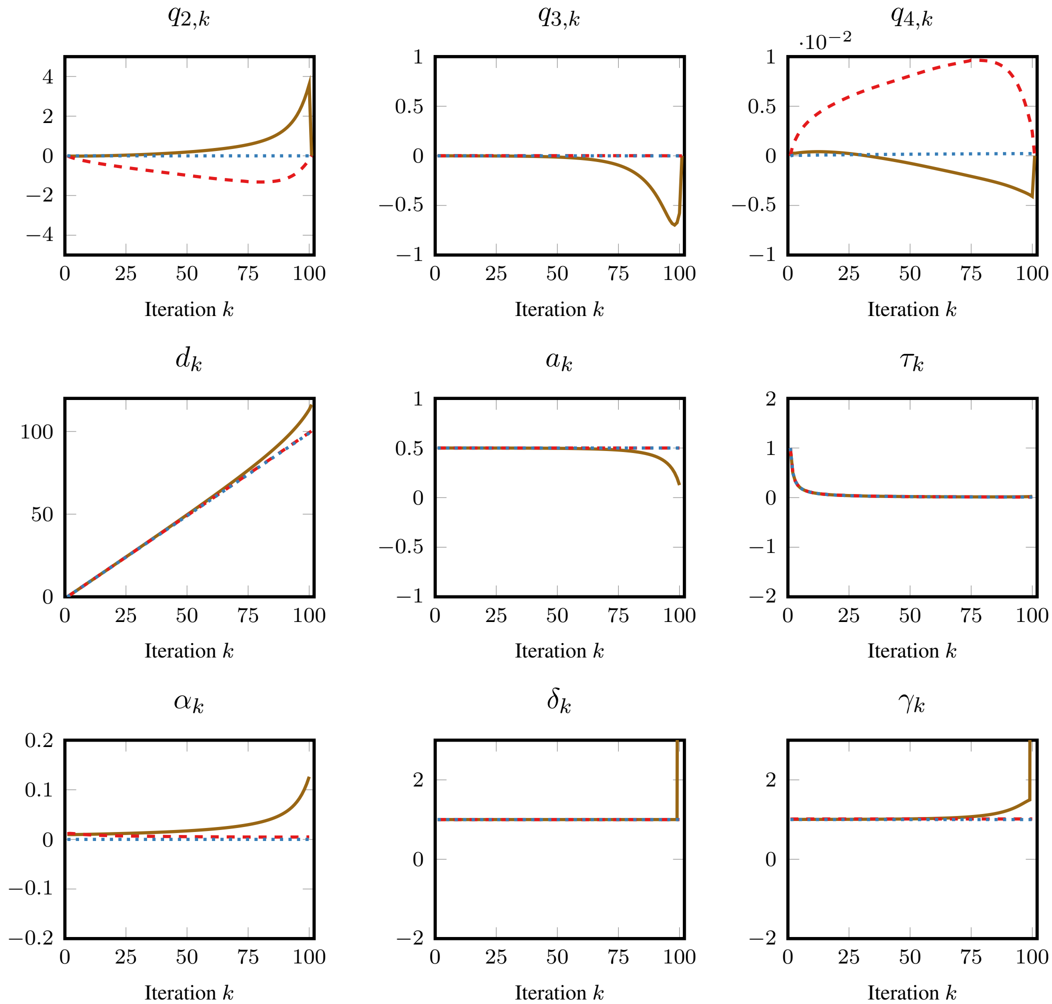

Figure 5: Numerical solution to (22) for $N = 100$, $n = 2$ and $L = 1$ (plain brown, large values for δ_{100} and γ_{100} were capped for readability purposes; they are due to the fact we impose no control on z_N with our initial choice ϕ_N^f), forced $a_k = \frac{L}{2}$ (dashed red), forced $a_k = \frac{L}{2}$ and $\alpha_k = 0$ (dotted blue). For convenience we do not show the values for $q_{1,k}$; they numerically appeared to be negligible for all values of k (about 10^{-7}) in all three scenarios. Total computational time: ~ 90 sec. on single core of Intel Core i7 1.8GHz CPU.

use function values at $x_{k+1}^{(i_k)}$ in the formulation, however, this did not improve the results in our experiments while increasing quite a bit the size of the Gram matrix by adding n^2 rows and columns to it (corresponding to all gradients

$$f_1'(x_{k+1}^{(1)}), \ldots, f_n'(x_{k+1}^{(1)}), \ldots, f_1'(x_{k+1}^{(n)}), \ldots, f_n'(x_{k+1}^{(n)})).$$

In order to reformulate (1) we therefore only need to encode function values F and gradient/coordinate in a matrix P, which we use for formulating the Gram matrix $G = P^\top P \succeq 0$:

$$\begin{aligned} P &= [\, x_k \mid y_k \mid f_1'(y_k) \ \ldots \ f_n'(y_k) \mid f_1'(y_{k+1}) \ \ldots \ f_n'(y_{k+1}) \,], \\ F &= [\, f_1(y_k) \ \ldots \ f_n(y_k) \mid f_1(y_{k+1}) \ \ldots \ f_n(y_{k+1}) \,]. \end{aligned}$$

We also denote by $\mathbf{y}_\star,\ \mathbf{x}_k,\ \mathbf{y}_k,\ \mathbf{g}_{i,\star},\ \mathbf{g}_{i,k},\ \mathbf{g}_{i,k+1}\in\mathbb{R}^{2+2n}$ (for all $i\in\{1,\ldots,n\}$) such that

$$x_\star=P\mathbf{y}_\star,\ x_k=P\mathbf{x}_k,\ y_k=P\mathbf{y}_k,\ f_i'(x_\star)=P\mathbf{g}_{i,\star},\ f_i'(y_k)=P\mathbf{g}_{i,k},\ f_i'(y_{k+1})=P\mathbf{g}_{i,k+1},$$

that is $\mathbf{y}_\star=\mathbf{g}_{i,\star}:=0$, $\mathbf{x}_k:=e_1$, $\mathbf{y}_k:=e_2$, $\mathbf{g}_{i,k}:=e_{2+i}$, $\mathbf{g}_{i,k+1}:=e_{2+n+i}$. For iteration $k+1$:

$$\begin{aligned}\mathbf{y}_{k+1}&:=\tau_k\mathbf{y}_k+(1-\tau_k)\mathbf{x}_k,\\ \mathbf{x}_{k+1}^{(l)}&:=\mathbf{x}_k-\gamma_k\mathbf{g}_{l,k},\end{aligned}$$

in order to have $y_{k+1}=P\mathbf{y}_{k+1}$ and $x_{k+1}^{(l)}=P\mathbf{x}_{k+1}^{(l)}$. Finally, we do the same for function values: $\mathbf{f}_{i,\star},\ \mathbf{f}_{i,k},\ \mathbf{f}_{i,k+1}\in\mathbb{R}^{2n}$ (for all $i\in\{1,\ldots,n\}$) with

$$f_\star=F\mathbf{f}_\star,\ f_i(y_k)=F\mathbf{f}_{i,k},\ f_i(y_{k+1})=F\mathbf{f}_{i,k+1},$$

that is $\mathbf{f}_{i,\star}:=0$, $\mathbf{f}_{i,k}:=e_i$ and $\mathbf{f}_{i,k+1}:=e_{n+i}$. For each function f_l (with $l\in\{1,\ldots,n\}$) we introduce a set of multipliers $\{\lambda_{i,j}^{(l)}\}_{i,j\in I_k}$ with $I_k=\{\star,k,k+1\}$. We use $\mathbb{S}^{2+2n}$ is the space of $(2+2n)\times(2+2n)$ symmetric matrices, and use M as defined in (19).

Given two doublets (d_k,Q_k) and (d_{k+1},Q_{k+1}), a pair (τ_k,γ_k), and some $n\in\mathbb{N}$, the inequality $\mathbb{E}_{i_k}\phi_{k+1}^f(x_{k+1}^{(i_k)},y_{k+1})\leq\phi_k^f(x_k,y_k)$ (ϕ_k^f defined in (24)) holds for all $f(x)=\frac{1}{n}\sum_{i=1}^n f_i(x)$ such that $f_i\in\mathcal{F}_L(\mathbb{R}^d)$, for all $d\in\mathbb{N}$, and all $x_k,y_k\in\mathbb{R}^d$ used to generate $x_{k+1}^{(i_k)},y_{k+1}$ with method (23) if and only if for all $l\in\{1,\ldots,n\}$ there exists $\{\lambda_{i,j}^{(l)}\}_{i,j\in I_k}$ with $I_k=\{\star,k,k+1\}$ such that

$$\lambda_{i,j}^{(l)}\geq 0 \text{ for all } i,j\in I_k \text{ and all } l\in\{1,\ldots,n\},$$

$$\tfrac{d_{k+1}}{n}\sum_{l=1}^n(\mathbf{f}_{l,k+1}-\mathbf{f}_{l,\star})-\tfrac{d_k}{n}\sum_{l=1}^n(\mathbf{f}_{l,k}-\mathbf{f}_{l,\star})+\sum_{\substack{i,j\in I_k\\ l\in\{1,\ldots,n\}}}\lambda_{i,j}^{(l)}(\mathbf{f}_{l,i}-\mathbf{f}_{l,j})=0 \text{ (in } \mathbb{R}^{2n}),$$

$$\tfrac{1}{n}\sum_{l\in\{1,\ldots,n\}}V_{k+1}^{(l)}-V_k+\sum_{\substack{i,j\in I_k\\ l\in\{1,\ldots,n\}}}\lambda_{i,j}^{(l)}M_{i,j}^{(l)}\preceq 0 \text{ (linear matrix inequality in } \mathbb{S}^{2+2n}),$$

with

$$V_k:=\begin{pmatrix}\mathbf{x}_k-\mathbf{y}_\star & \mathbf{y}_k-\mathbf{y}_\star & \mathbf{g}_{1,k} & \ldots & \mathbf{g}_{n,k}\end{pmatrix}Q_k\begin{pmatrix}\mathbf{x}_k^\top-\mathbf{y}_\star^\top\\ \mathbf{y}_k^\top-\mathbf{y}_\star^\top\\ \mathbf{g}_{1,k}^\top\\ \vdots\\ \mathbf{g}_{n,k}^\top\end{pmatrix}\in\mathbb{S}^{2+2n},$$

$$V_{k+1}^{(l)}:=\begin{pmatrix}\mathbf{x}_{k+1}^{(l)}-\mathbf{y}_\star & \mathbf{y}_{k+1}-\mathbf{y}_\star & \mathbf{g}_{1,k+1} & \ldots & \mathbf{g}_{n,k+1}\end{pmatrix}Q_{k+1}\begin{pmatrix}\mathbf{x}_{k+1}^{(l)\top}-\mathbf{y}_\star^\top\\ \mathbf{y}_{k+1}^\top-\mathbf{y}_\star^\top\\ \mathbf{g}_{1,k+1}^\top\\ \vdots\\ \mathbf{g}_{n,k+1}^\top\end{pmatrix}\in\mathbb{S}^{2+2n},$$

$$M_{i,j}^{(l)}:=\begin{pmatrix}\mathbf{y}_i & \mathbf{y}_j & \mathbf{g}_{l,i} & \mathbf{g}_{l,j}\end{pmatrix}M\begin{pmatrix}\mathbf{y}_i & \mathbf{y}_j & \mathbf{g}_{l,i} & \mathbf{g}_{l,j}\end{pmatrix}^\top\in\mathbb{S}^{2+2n}.$$

Appendix G. Stochastic gradients under weak growth conditions

This section deals with weak growth conditions. It is straightforward to adapt the methodology to e.g., strong growth conditions. However, since recent works already covered the topic in a comprehensive way (see e.g., Vaswani et al. (2019) and the references therein), we chose not to cover it.

Under similar motivations as those of the Sec. 3.2.2, consider minimizing $f \in \mathcal{F}_L$

$$\min_{x\in\mathbb{R}^d} f(x),$$

with an unbiased gradient estimate $G(x;i)$ satisfying a weak growth condition for some $\rho \geq 1$:

$$\mathbb{E}_i G(x;i) = f'(x), \quad \mathbb{E}_i \|G(x;i)\|^2 \leq 2\rho L(f(x) - f_\star),$$

for all $x \in \mathbb{R}^d$. In that setting, it was shown by Vaswani et al. (2019) that averaging was actually reaching the $O(k^{-1})$ convergence for $\mathbb{E} f(\bar{x}_k) - f_\star$. The best method we could obtain with the previous methodology achieves a $O(k^{-1})$ convergence for function values at its last iterate.

Theorem 14 *Let $x_k \in \mathbb{R}^d$, $f \in \mathcal{F}_L$ and an unbiased stochastic oracle satisfying a weak growth condition:*

$$\mathbb{E}_i G(x;i) = f'(x), \quad \mathbb{E}_i \|G(x;i)\|^2 \leq 2\rho L(f(x) - f_\star)$$

for all $x \in \mathbb{R}^d$ and $i \in I$. Then, the iterative scheme

$$\begin{aligned} y_{k+1} &= \tfrac{d_k}{d_k+\delta_k L} y_k + \tfrac{\delta_k L}{d_k+\delta_k L} x_k, \\ x_{k+1}^{(i_k)} &= x_k - \delta_k G(y_{k+1}; i_k), \end{aligned}$$

satisfies

$$d_{k+1}[f(y_{k+1}) - f_\star] + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq d_k(f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2,$$

for all values of $d_k, \delta_k \geq 0$ and $d_{k+1} = d_k + \delta_k L - \rho\delta_k^2 L^2$.

Using $\delta_k = \frac{1}{2\rho L}$ (choice that maximizes d_{k+1}) and $d_0 = 0$ leads to $d_k = \frac{k}{4\rho}$ and to the algorithm

$$\begin{aligned} y_{k+1} &= \tfrac{k}{k+2} y_k + \tfrac{2}{k+2} x_k, \\ x_{k+1}^{(i_k)} &= x_k - \tfrac{1}{2\rho L} G(y_{k+1}; i_k), \end{aligned}$$

for which the bound $\mathbb{E} f(y_k) - f_\star \leq \frac{2\rho L\|x_0 - x_\star\|^2}{k}$ holds for $k \geq 0$. As in the previous section (overparametrized models, App. F), the use of primal averaging for weak growth conditions appeared naturally through the use of the parameter selection technique starting from

$$\begin{aligned} y_{k+1} &= (1-\tau_k)x_k + \tau_k z_k, \\ x_{k+1}^{(i_k)} &= y_{k+1} - \alpha_k G(y_{k+1}; i_k), \\ z_{k+1}^{(i_k)} &= (1-\delta_k) y_{k+1} + \delta_k z_k - \gamma_k G(y_{k+1}; i_k), \end{aligned}$$

for which we wish to optimize the policy $\{(\tau_k, \alpha_k, \delta_k, \gamma_k)\}_k$. Using the line-search workaround

$$\begin{aligned} y_{k+1} &= \operatorname{argmin}_x \{f(x) \text{ subject to } x \in x_k + \operatorname{span}\{z_k - x_k\}\}, \\ x_{k+1}^{(i_k)} &= \operatorname{argmin}_x \{f(x) \text{ subject to } x \in y_{k+1} + \operatorname{span}\{G(y_{k+1}; i_k)\}\}, \\ z_{k+1}^{(i_k)} &= (1 - \delta_k) y_{k+1} + \delta_k z_k - \gamma_k G(y_{k+1}; i_k), \end{aligned}$$

the parameters of primal averaging appeared as a feasible point to the parameter selection technique.

Before going into next section, let us briefly mention that this setting can be embedded within similar LMIs as before, by encoding the $G(x;i)$'s for all i's and all x's where a stochastic gradient is used. Then, one can rely on interpolation conditions for constraining gradients and stochastic gradients (stochastic gradients are embedded within interpolation conditions by averaging them $f'(x) = \frac{1}{n}\sum_{i=1}^n G(x;i)$). For example, for analyzing a stochastic gradient scheme $x_{k+1}^{(i_k)} = x_k - \delta_k G(x_k; i_k)$ under a weak growth condition, for a potential $\phi_k^f = d_k\,(f(x_k) - f_\star) + \frac{L}{2}\|x_k - x_\star\|^2$, one can use a Gram matrix $G = P^\top P \succeq 0$ with

$$P = [\; x_k \mid G(x_k; 1) \;\ldots\; G(x_k; n) \mid f'(x_{k+1}^{(1)}) \;\ldots\; f'(x_{k+1}^{(n)}) \;],$$

along with function values $F = [\; f(x_k) \mid f(x_{k+1}^{(1)}) \;\ldots\; f(x_{k+1}^{(n)}) \;]$; the subtlety in that setting is to include A_{var} and a_{var} for encoding the variance of $G(x_k;i)$'s, with

$$A_{\text{var}} = \frac{1}{n}\sum_{i=1}^n e_{1+i} e_{1+i}^\top \in \mathbb{S}^{1+2n}, \quad a_{\text{var}} = 2\rho L\, e_1 \in \mathbb{R}^{n+2},$$

for requiring $\operatorname{Trace}(A_{\text{var}} G) = \mathbb{E}_i \|G(x_k;i)\|^2 \le 2\rho L (f(x_k) - f_\star) = F a_{\text{var}}$.

G.1. Proof of Theorem 14

Proof Combine the following inequalities with their corresponding weights.

- ◇ Convexity between $x_\star$ and y_{k+1} with weight $\lambda_1 = \delta_k L$,

$$f_\star \ge f(y_{k+1}) + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle,$$

- ◇ convexity between y_k and y_{k+1} with weight $\lambda_2 = d_k$,

$$f(y_k) \ge f(y_{k+1}) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle,$$

- ◇ weak growth condition with weight $\lambda_3 = \frac{\delta_k^2 L}{2}$

$$\mathbb{E}_{i_k}\|G(y_{k+1}; i_k)\|^2 \le 2\rho L (f(y_{k+1}) - f_\star).$$

The weighted sum of those inequalities yields the desired result:

$$\begin{aligned} 0 \ge\, &\lambda_1 \left[f(y_{k+1}) - f_\star + \langle f'(y_{k+1}); x_\star - y_{k+1}\rangle\right] + \lambda_2 \left[f(y_{k+1}) - f(y_k) + \langle f'(y_{k+1}); y_k - y_{k+1}\rangle\right] \\ &+ \lambda_3 \left[\mathbb{E}_{i_k}\|G(y_{k+1}; i_k)\|^2 - 2\rho L (f(y_{k+1}) - f_\star)\right] \\ =&(d_k + \delta_k L - \rho \delta_k^2 L^2)(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 - d_k (f(y_k) - f_\star) - \tfrac{L}{2}\|x_k - x_\star\|^2, \end{aligned}$$

which can be rearranged to the desired:

$$(d_k + \delta_k L - \rho \delta_k^2 L^2)(f(y_{k+1}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \le d_k (f(y_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2.$$

■

Appendix H. Bounded variance at optimum

As in the previous sections, it is possible to apply the methodology and the parameter selection technique for designing a method to solve

$$\min_{x\in\mathbb{R}^d}\{f(x)\equiv\ \mathbb{E}_i f_i(x)\},$$

with $\mathbb{E}_i\|f'(x_\star)\|^2\leq\sigma_\star^2$ (over-parametrized models from Sec. 4 arise as particular case with $\sigma_\star=0$). Again, we obtain a primal averaging scheme, but with smaller allowed step-sizes.

Theorem 15 *Let $x_k\in\mathbb{R}^d$, $f_i\in\mathcal{F}_L$ and an optimal point $x_\star$ such that $\mathbb{E}_i\|f_i'(x_\star)\|^2\leq\sigma_\star^2$. Then the iterative scheme*

$$\begin{aligned}y_{k+1}&=\tfrac{d_k}{d_k+\delta_k L}y_k+\tfrac{\delta_k L}{d_k+\delta_k L}x_k,\\ x_{k+1}^{(i_k)}&=x_k-\delta_k f_{i_k}'(y_{k+1}),\end{aligned}$$

satisfies

$$d_{k+1}(f(y_{k+1})-f_\star)+\tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)}-x_\star\|^2\leq d_k(f(y_k)-f_\star)+\tfrac{L}{2}\|x_k-x_\star\|^2+e_k\sigma_\star^2,$$

for all values $d_k,d_{k+1},\delta_k,e_k\geq 0$ satisfying $0\leq\delta_k<\frac{1}{L}$, $d_k\geq 0$, $d_{k+1}\leq d_k+\delta_k L$ and $e_k=\frac{\delta_k^2 L}{2(1-\delta_k L)}$.

Proof Let us reformulate the following weighted sum.

- ◇ Averaged smoothness and convexity between y_{k+1} and $x_\star$ with weight $\lambda_1=\delta_k L$:

$$f_\star\geq f(y_{k+1})+\langle f'(y_{k+1});x_\star-y_{k+1}\rangle+\tfrac{1}{2L}\mathbb{E}_{i_k}\|f_{i_k}'(y_{k+1})-f_{i_k}'(x_\star)\|^2,$$

- ◇ convexity between x_{k+1} and $x_\star$ with weight $\lambda_2=d_k$:

$$f(y_k)\geq f(y_{k+1})+\langle f'(y_{k+1});y_k-y_{k+1}\rangle,$$

- ◇ bounded variance of the gradients at $x_\star$ with weight $\lambda_3=e_k$

$$\mathbb{E}_{i_k}\|f_{i_k}'(x_\star)\|^2\leq\sigma_\star^2.$$

The weighted sum of those inequalities can be rewritten as

$$\begin{aligned}0\geq&\lambda_1\left[f(y_{k+1})-f_\star+\langle f'(y_{k+1});x_\star-y_{k+1}\rangle+\tfrac{1}{2L}\mathbb{E}_{i_k}\|f_{i_k}'(y_{k+1})\|^2\right]\\&+\lambda_2\left[f(y_{k+1})-f(y_k)+\langle f'(y_{k+1});y_k-y_{k+1}\rangle\right]\\=&(d_k+\delta_k L)(f(y_{k+1})-f_\star)+\tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)}-x_\star\|^2-d_k(f(y_k)-f_\star)-\tfrac{L}{2}\|x_k-x_\star\|^2\\&+\tfrac{\delta_k}{2-2\delta_k L}\mathbb{E}_{i_k}\|f_{i_k}'(x_\star)+(\delta_k L-1)f_{i_k}'(y_{k+1})\|^2+\tfrac{2e_k(\delta_k L-1)+\delta_k^2 L}{2\delta_k L-2}\mathbb{E}_{i_k}\|f_{i_k}'(x_\star)\|^2-e_k\sigma_\star^2,\end{aligned}$$

which can be rearranged to (note that e_k is chosen such that $\frac{2e_k(\delta_k L-1)+\delta_k^2 L}{2\delta_k L-2}=0$)

$$\begin{aligned}&(d_k+\delta_k L)(f(y_{k+1})-f_\star)+\tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)}-x_\star\|^2\\&\leq d_k(f(y_k)-f_\star)+\tfrac{L}{2}\|x_k-x_\star\|^2+e_k\sigma_\star^2-\tfrac{\delta_k}{2-2\delta_k L}\mathbb{E}_{i_k}\|f_{i_k}'(x_\star)+(\delta_k L-1)f_{i_k}'(y_{k+1})\|^2\\&\leq d_k(f(y_k)-f_\star)+\tfrac{L}{2}\|x_k-x_\star\|^2+e_k\sigma_\star^2,\end{aligned}$$

where the last inequality follows from $\delta_k<\frac{1}{L}$. ■

Appendix I. Randomized block-coordinate descent

In this section, we illustrate the use of the methodology for block-coordinate type schemes. In contrast with standard references on the topic (see e.g., Nesterov (2012a); Richtárik and Takáč (2014); Fercoq and Richtárik (2015)), we make use of the global Lipschitz constant of the function to be minimized, instead of Lipschitz constants of the individual blocks. This choice is made for convenience and illustrative purpose, and the results presented here can be adapted for using the Lipschitz constants of the blocks (see e.g., Shi and Liu (2017) where it is done for cyclic coordinate descent), or even for dealing with (separable) proximal operators (Richtárik and Takáč, 2014).

The setting is as follows: consider the problem

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f \in \mathcal{F}_{\mu,L}$, and x is partitioned into n blocks: $x = \sum_{i=1}^n \mathbf{U}_i\, x$, with the partition described by $[\mathbf{U}_1\, \mathbf{U}_2\, \ldots\, \mathbf{U}_n] = I_d$. For solving the problem, we use $x_{k+1}^{(i_k)} = x_k - \delta_k \mathbf{U}_{i_k} f'(x_k)$, where i_k is chosen uniformly at random in the set $\{1,\ldots,n\}$, δ_k is a step-size, and $\mathbf{U}_i f'(x_k) = \nabla_i f(x_k)$ is the directional derivate along the i$^{\text{th}}$ block of coordinates. We do not detail the LMI formulations here but rather note that they follow the exact same lines as in the previous sections after adapting the Gram matrix G, formulating it with $G = P^\top P \succeq 0$ and

$$P = [x_0 \,|\, \mathbf{U}_1 f'(x_k)\, \ldots\, \mathbf{U}_n f'(x_k) \,|\, f'(x_{k+1}^{(1)})\, \ldots\, f'(x_{k+1}^{(n)})].$$

I.1. Block-coordinate descent for smooth convex minimization

Theorem 16 *Let $x_k \in \mathbb{R}^d$, $f \in \mathcal{F}_L$, and $x_{k+1}^{(i_k)} = x_k - \delta_k \mathbf{U}_{i_k} f'(x_k)$ with $0 \leq \delta_k \leq \frac{1}{L}$. The inequality $\mathbb{E}_{i_k}\phi_{k+1}^f(x_{k+1}^{(i_k)}) \leq \phi_k^f(x_k)$ holds with*

$$\phi_k^f = d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2,$$

for all values $d_k \geq 1$, and $d_{k+1} = d_k + \frac{\delta_k L}{n}$.

Proof Once more, the proof consists in linear combinations of inequalities:

- ⋄ convexity between $x_\star$ and x_k with weight $\lambda_1 = \frac{\delta_k L}{n}$:

$$f_\star \geq f(x_k) + \langle f'(x_k); x_\star - x_k\rangle,$$

- ⋄ averaged smoothness between $x_{k+1}^{(i)}$ and x_k with weight $\lambda_2 = d_k + \frac{\delta_k L}{n}$:

$$\mathbb{E}_{i_k} f(x_{k+1}^{(i_k)}) \leq f(x_k) + \mathbb{E}_{i_k}\langle f'(x_k); x_{k+1}^{(i_k)} - x_k\rangle + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_k\|^2.$$

Recalling that $f'(x_k) = \sum_{i=1}^n \mathbf{U}_i f'(x_k)$ (and therefore $\sum_{i=1}^n \|\mathbf{U}_i f(x_k)\|^2 = \|f(x_k)\|^2$) and that $x_{k+1}^{(i_k)} = x_k - \delta_k \mathbf{U}_{i_k} f'(x_k)$, the weighted sum can be reformulated as

$$\begin{aligned}
0 \geq &\lambda_1 \left[f(x_k) - f_\star + \langle f'(x_k); x_\star - x_k\rangle\right] \\
&+ \lambda_2 \left[\mathbb{E}_{i_k} f(x_{k+1}^{(i_k)}) - f(x_k) - \mathbb{E}_{i_k}\langle f'(x_k); x_{k+1}^{(i_k)} - x_k\rangle - \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_k\|^2\right] \\
=&(d_k + \tfrac{\delta_k L}{n})\mathbb{E}_{i_k}(f(x_{k+1}^{(i_k)}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \\
&- d_k(f(x_k) - f_\star) - \tfrac{L}{2}\|x_k - x_\star\|^2 + \tfrac{\delta_k}{n}\left((1 - \tfrac{\delta_k L}{2})(d_k + \tfrac{\delta_k L}{n}) - \tfrac{\delta_k L}{2}\right)\|f'(x_k)\|^2,
\end{aligned}$$

which can in turn be reorganized as

$$\begin{aligned}&(d_k + \tfrac{\delta_k L}{n})\mathbb{E}_{i_k}(f(x_{k+1}^{(i_k)}) - f_\star) + \tfrac{L}{2}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2\\ &\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2 - \tfrac{\delta_k}{n}\left((1 - \tfrac{\delta_k L}{2})(d_k + \tfrac{\delta_k L}{n}) - \tfrac{\delta_k L}{2}\right)\|f'(x_k)\|^2,\\ &\leq d_k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|^2,\end{aligned}$$

where the last inequality follows from $\left((1 - \frac{\delta_k L}{2})(d_k + \frac{\delta_k L}{n}) - \frac{\delta_k L}{2}\right) \geq 0$, which can be verified as follows. Define

$$\phi(\delta_k) := \left((1 - \tfrac{\delta_k L}{2})(d_k + \tfrac{\delta_k L}{n}) - \tfrac{\delta_k L}{2}\right),$$

which is a negative definite quadratic in δ_k. In addition, we have $\phi(\frac{1}{L}) = \frac{1}{2}\left(d_k + \frac{1}{n} - 1\right) > 0$ (because $d_k \geq 1$ by assumption) and $\phi(0) = d_k > 0$. Since by assumption we had $0 \leq \delta_k \leq \frac{1}{L}$, this discussion completes the proof. ■

I.2. Block-coordinate descent for smooth strongly convex minimization

The following result was obtained using the same methodology as before: finding a feasible point to (1). We obtain a simple linear convergence expressed in terms of distance to optimality:

$$\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 \leq \rho^2\|x_k - x_\star\|^2,$$

with $\rho^2 := \max\left\{\left(\frac{(\delta_k\mu-1)^2+n-1}{n}\right), \left(\frac{(\delta_k L-1)^2+n-1}{n}\right)\right\}$.

Theorem 17 *Let $x_k \in \mathbb{R}^d$, $f \in \mathcal{F}_{\mu,L}$ (class of L-smooth μ-strongly convex functions), and $x_{k+1}^{(i_k)} = x_k - \delta_k \mathbf{U}_{i_k} f'(x_k)$. The inequality $\mathbb{E}_{i_k}\phi_{k+1}^f(x_{k+1}^{(i_k)}) \leq \phi_k^f(x_k)$ holds with*

$$\phi_k^f = a_k\|x_k - x_\star\|^2,$$

for all values $a_k > 0$, and $a_{k+1} = a_k \max\left\{\left(\frac{(\delta_k\mu-1)^2+n-1}{n}\right), \left(\frac{(\delta_k L-1)^2+n-1}{n}\right)\right\}^{-1}$.

Proof As in the previous sections, the proof consists in a linear combination of inequalities:

- ⋄ smoothness and strong convexity between x_k and $x_\star$ with weight λ

$$\begin{aligned}f_\star \geq &f(x_k) + \langle f'(x_k); x_\star - x_k\rangle\\ &+ \tfrac{1}{2(1-\frac{\mu}{L})}\left(\tfrac{1}{L}\|f'(x_k)\|^2 + \mu\|x_k - x_\star\|^2 - 2\tfrac{\mu}{L}\langle f'(x_k); x_k - x_\star\rangle\right),\end{aligned}$$

- ⋄ smoothness and strong convexity between $x_\star$ and x_k with weight λ

$$f(x_k) \geq f_\star + \tfrac{1}{2(1-\frac{\mu}{L})}\left(\tfrac{1}{L}\|f'(x_k)\|^2 + \mu\|x_k - x_\star\|^2 - 2\tfrac{\mu}{L}\langle f'(x_k); x_k - x_\star\rangle\right).$$

Summing up those two inequalities leads to

$$0 \geq \lambda\left[\langle f'(x_k); x_\star - x_k\rangle + \tfrac{1}{(1-\frac{\mu}{L})}\left(\tfrac{1}{L}\|f'(x_k)\|^2 + \mu\|x_k - x_\star\|^2 - 2\tfrac{\mu}{L}\langle f'(x_k); x_k - x_\star\rangle\right)\right],$$

which we reformulate below. The proof is split in two cases:

◇ $\rho^2 = \frac{(\delta_k\mu-1)^2+n-1}{n}$; in that case set $\lambda = \frac{2\delta_k}{n}(1-\delta_k\mu)$. Recalling that $f'(x_k) = \sum_{i_k=1}^n \mathbf{U}_{i_k} f'(x_k)$ and that $x_{k+1}^{(i_k)} = x_k - \delta_k \mathbf{U}_{i_k} f'(x_k)$, we get

$$\begin{aligned}0 \geq \lambda & \left[\langle f'(x_k); x_\star - x_k\rangle + \tfrac{1}{(1-\frac{\mu}{L})}\left(\tfrac{1}{L}\|f'(x_k)\|^2 + \mu\|x_k - x_\star\|^2 - 2\tfrac{\mu}{L}\langle f'(x_k); x_k - x_\star\rangle\right)\right] \\ &= \tfrac{1}{n}\sum_{i_k=1}^n \|x_k - \delta_k \mathbf{U}_{i_k} f'(x_k) - x_\star\|^2 - \tfrac{(1-\delta_k\mu)^2+n-1}{n}\|x_k - x_\star\|^2 \\ &\quad + \tfrac{\delta_k(2-\delta_k(L+\mu))}{n(L-\mu)}\|\mu(x_k - x_\star) - f'(x_k)\|^2,\end{aligned}$$

which can be reorganized as

$$\begin{aligned}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 &\leq \tfrac{(1-\delta_k\mu)^2+n-1}{n}\|x_k - x_\star\|^2 - \tfrac{\delta_k(2-\delta_k(L+\mu))}{n(L-\mu)}\|\mu(x_k - x_\star) - f'(x_k)\|^2, \\ &\leq \tfrac{(1-\delta_k\mu)^2+n-1}{n}\|x_k - x_\star\|^2,\end{aligned}$$

where the last inequality is valid because

$$\rho^2 = \tfrac{(\delta_k\mu-1)^2+n-1}{n} = \max\left\{\left(\tfrac{(\delta_k\mu-1)^2+n-1}{n}\right), \left(\tfrac{(\delta_k L-1)^2+n-1}{n}\right)\right\}$$

in this setting (see choice of ρ), and therefore $\delta_k \leq \frac{2}{L+\mu}$.

◇ $\rho^2 = \frac{(\delta_k\mu-1)^2+n-1}{n}$; in that case set $\lambda = \frac{2\delta_k}{n}(\delta_k L - 1)$. By recalling again that $f'(x_k) = \sum_{i_k=1}^n \mathbf{U}_{i_k} f'(x_k)$ and that $x_{k+1}^{(i_k)} = x_k - \delta_k \mathbf{U}_{i_k} f'(x_k)$, we get

$$\begin{aligned}0 \geq \lambda & \left[\langle f'(x_k); x_\star - x_k\rangle + \tfrac{1}{(1-\frac{\mu}{L})}\left(\tfrac{1}{L}\|f'(x_k)\|^2 + \mu\|x_k - x_\star\|^2 - 2\tfrac{\mu}{L}\langle f'(x_k); x_k - x_\star\rangle\right)\right] \\ &= \tfrac{1}{n}\sum_{i_k=1}^n \|x_k - \delta_k \mathbf{U}_{i_k} f'(x_k) - x_\star\|^2 - \tfrac{(1-\delta_k L)^2+n-1}{n}\|x_k - x_\star\|^2 \\ &\quad + \tfrac{\delta_k(-2+\delta_k(L+\mu))}{n(L-\mu)}\|L(x_k - x_\star) - f'(x_k)\|^2,\end{aligned}$$

which can be reorganized as

$$\begin{aligned}\mathbb{E}_{i_k}\|x_{k+1}^{(i_k)} - x_\star\|^2 &\leq \tfrac{(1-\delta_k L)^2+n-1}{n}\|x_k - x_\star\|^2 - \tfrac{\delta_k(-2+\delta_k(L+\mu))}{n(L-\mu)}\|L(x_k - x_\star) - f'(x_k)\|^2, \\ &\leq \tfrac{(1-\delta_k L)^2+n-1}{n}\|x_k - x_\star\|^2,\end{aligned}$$

where the last inequality is valid because

$$\rho^2 = \tfrac{(\delta_k L-1)^2+n-1}{n} = \max\left\{\left(\tfrac{(\delta_k\mu-1)^2+n-1}{n}\right), \left(\tfrac{(\delta_k L-1)^2+n-1}{n}\right)\right\},$$

in this setting (see choice of ρ), and therefore $\delta_k \geq \frac{2}{L+\mu}$.

■