HAL
archives-ouvertes.fr

Stochastic Composite Least-Squares Regression with Convergence Rate O(1/n)

Nicolas Flammarion, Francis Bach

▶ **To cite this version:**

Nicolas Flammarion, Francis Bach. Stochastic Composite Least-Squares Regression with Convergence Rate O(1/n). Proceedings of The 30th Conference on Learning Theory, (COLT), 2017, Amsterdam, Netherlands. 2017. <hal-01472867>

HAL Id: hal-01472867
https://hal.inria.fr/hal-01472867

Submitted on 21 Feb 2017

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

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

Stochastic Composite Least-Squares Regression with convergence rate $O(1/n)$

Nicolas Flammarion and Francis Bach
INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France
nicolas.flammarion@ens.fr, francis.bach@ens.fr

February 21, 2017

Abstract

We consider the minimization of composite objective functions composed of the expectation of quadratic functions and an arbitrary convex function. We study the stochastic dual averaging algorithm with a constant step-size, showing that it leads to a convergence rate of $O(1/n)$ without strong convexity assumptions. This thus extends earlier results on least-squares regression with the Euclidean geometry to (a) all convex regularizers and constraints, and (b) all geometries represented by a Bregman divergence. This is achieved by a new proof technique that relates stochastic and deterministic recursions.

1 Introduction

Many learning problems may be cast as the optimization of an objective function defined as an expectation of random functions, and which can be accessed only through samples. In this paper, we consider *composite* problems of the form

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_z \ell(z, \theta) + g(\theta), \tag{1}$$

where for any z, $\ell(z, \cdot)$ is a convex quadratic function (plus some linear terms) and g is any extended-value convex function.

In a machine learning context, $\ell(z, \theta)$ is the loss occurred for the observation z and the predictor parameterized by θ, $f(\theta) = \mathbb{E}_z \ell(z, \theta)$ is its generalization error, while the function g represents some additional regularization or constraints on the predictor. Thus in this paper we consider composite least-squares regression problems, noting that solving such problems effectively leads to efficient algorithms for all smooth losses by using an online Newton algorithm (Bach and Moulines, 2013), with the same running-time complexity of $O(d)$ per iteration for linear predictions.

When $g = 0$, averaged stochastic gradient descent with a constant step-size achieves the optimal convergence rate of $O(1/n)$ after n observations, even in ill-conditioned settings without strong

convexity (Dieuleveut et al., 2016; Jain et al., 2016), with precise non-asymptotic results that depend on the statistical noise variance σ^2 of the least-squares problem, as $\sigma^2 d/n$, and on the squared Euclidean distance between the initial predictor θ_0 and the optimal predictor θ_*, as $\|\theta_0 - \theta_*\|_2^2/n$.

In this paper, we extend this $O(1/n)$ convergence result in two different ways:

- **Composite problems**: we provide a new algorithm that deals with composite problems where g is (essentially) any extended-value convex function, such as the indicator function of a convex set for constrained optimization, or a norm or squared norm for additional regularization. This situation is common in many applications in machine learning and signal processing (see, e.g., Rish and Grabarnik, 2014, and references therein). Because we consider large steps-sizes (that allow robustness to ill-conditioning), the new algorithm is *not* simply a proximal extension; for example, in the constrained case, averaged projected stochastic gradient descent with a constant step-size is not convergent, even for quadratic functions.
- **Beyond Euclidean geometry**: Following mirror descent (Nemirovski and Yudin, 1979), our new algorithm can take into account a geometry obtained with a Bregman divergence D_h associated with a convex function h, which can typically be the squared Euclidean norm (leading to regular stochastic gradient descent in the non-composite case), the entropy function, or the squared ℓ_p-norm. This will allow convergence rates proportional to $D_h(\theta_*, \theta_0)/n$, which may be significantly smaller than $\|\theta_0 - \theta_*\|^2/n$ in many situations.

In order to obtain these two extensions, we consider the stochastic dual averaging algorithm of Nesterov (2009) and Xiao (2010) which we present in Section 2, and study under the particular set-up of *constant step-size with averaging*, showing in Section 3 that it also achieves a convergence rate of $O(1/n)$ even without strong-convexity. This is achieved by a new proof technique that relates stochastic and deterministic recursions.

Given that known lower-bounds for this class of problems are proportional to $1/\sqrt{n}$ for function values, we established our $O(1/n)$ results with a different criterion, namely the Mahalanobis distance associated with the Hessian of the least-squares problem. In our simulations in Section 5, the two criteria behave similarly. Finally, in Section 4, we shed additional insights of the relationships between mirror descent and dual averaging, in particular in terms of continuous-time interpretations.

2 Dual averaging algorithm

In this section, we introduce dual averaging as well as related frameworks, together with new results in the deterministic case.

2.1 Assumptions

We consider the Euclidean space $\mathbb{R}^d$ of dimension d endowed with the natural inner product $\langle \cdot, \cdot \rangle$ and an arbitrary norm $\| \cdot \|$ (which may not be the Euclidean norm). We denote by $\| \cdot \|_*$ its dual norm and for any symmetric positive-definite matrix A, by $\| \cdot \|_A = \sqrt{\langle \cdot, A \cdot \rangle}$ the Mahalanobis norm. For a vector $\theta \in \mathbb{R}^d$, we denote by $\theta(i)$ its i-th coordinate and by $\|\theta\|_p = (\sum_{i=1}^d |\theta(i)|^p)^{1/p}$

its ℓ_p-norm. We also denote the convex conjugate of a function f by $f^*(\eta) = \sup_{\theta \in \mathbb{R}^d} \langle \eta, \theta \rangle - f(\theta)$. We remind that a function f is L-smooth with respect to a norm $\|\cdot\|$ if for all $(\alpha, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, $\|\nabla f(\alpha) - \nabla f(\beta)\|_* \leq L\|\alpha - \beta\|$ and is μ-strongly convex if for all $(\alpha, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$ and $g \in \partial f(\beta)$, $f(\alpha) \geq f(\beta) + \langle g, \alpha - \beta \rangle + \frac{\mu}{2}\|\alpha - \beta\|^2$ (see, e.g., Shalev-Shwartz and Singer, 2006).

We consider problems of the form:

$$\min_{\theta \in \mathcal{X}} \psi(\theta) = f(\theta) + g(\theta), \tag{2}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set with non empty interior. Throughout this paper, we make the following general assumptions:

(A1) $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous convex function and is differentiable on $\mathring{\mathcal{X}}$ (the interior of $\mathcal{X}$).

(A2) $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous convex function.

(A3) $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\overline{\text{dom } h} \cap \overline{\text{dom } g} = \mathcal{X}$, $\mathring{\text{dom}}\, h \cap \text{dom } g \neq \emptyset$. Moreover h is a Legendre function (Rockafellar, 1970, chap. 26):

- h is a proper lower semicontinuous strictly convex function, differentiable on $\mathring{\text{dom}}\, h$.
- The gradient of h is diverging on the boundary of dom h (i.e., $\lim_{n \to +\infty} \|\nabla h(\theta_n)\| = \infty$ for any sequence (θ_n) converging to a boundary point of dom h). Note that ∇h is then a bijection from $\mathring{\text{dom}}\, h$ to $\mathring{\text{dom}}\, h^*$ whose inverse is the gradient of the conjugate ∇h^*.

(A4) The function $\psi = f + g$ attains its minimum over $\mathcal{X}$ at a certain $\theta_* \in \mathbb{R}^d$ (which may not be unique).

Note that we adopt the same framework as Bauschke et al. (2016) with the difference that the convex constraint $\mathcal{C}$ can be handled with more flexibility: either by considering a Legendre function h whose domain is $\mathcal{C}$ or by considering the hard constraint $g(\theta) = 1_C(\theta)$ (equal to 0 if $\theta \in \mathcal{C}$ and $+\infty$ otherwise).

2.2 Dual averaging algorithm

In this section we present the dual averaging algorithm (referred to from now on as "DA") for solving composite problems of the form of Eq. (2). It starts from $\theta_0 \in \mathring{\text{dom}}\, h$ and $\eta_0 = \nabla h(\theta_0)$ and iterates for $n \geq 1$ the recursion

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma \nabla f(\theta_{n-1}) \\ \theta_n &= \nabla h_n^*(\eta_n), \end{aligned} \tag{3}$$

with $h_n = h + n\gamma g$ and $\gamma \in (0, \infty)$ (commonly referred to as the step-size in optimization or the learning rate in machine learning). We note that equivalently $\theta_n \in \text{argmax}_{\theta \in \mathbb{R}^d} \{\langle \eta_n, \theta \rangle - h_n(\theta)\}$. When $h = \frac{1}{2}\|\cdot\|_2^2$ and $g = 0$, we recover gradient descent.

Two iterates (η_n, θ_n) are updated in DA. The dual iterate η_n is simply proportional to the sum of the gradients evaluated in the primal iterates (θ_n). The update of the primal iterate θ_n is more

complex and raises two different issues: its existence and its tractability. We discuss the first point in Appendix A and assume, as of now, that the method is generally well defined in practice. The tractability of θ_n is essential and the algorithm is only used in practice if the functions h and g are simple in the sense that the gradient ∇h_n^* may be computed effectively. This is the case if there exists a closed form expression. Usual examples are given in Appendix I.

Euclidean case and proximal operators. In the Euclidean case, Eq. (3) may be written in term of the proximal operator defined by Moreau (1962) as $\text{Prox}_g(\eta) = \text{argmin}_{\theta\in\mathcal{X}}\{\frac{1}{2}\|\theta-\eta\|_2^2 + g(\theta)\}$:

$$\theta_n = \operatorname*{argmin}_{\theta\in\mathcal{X}} \left\{ \langle -\eta_n, \theta\rangle + n\gamma g(\theta) + \frac{1}{2}\|\theta\|_2^2 \right\} = \operatorname*{argmin}_{\theta\in\mathcal{X}} \left\{ \frac{1}{2}\|\theta-\eta_n\|_2^2 + n\gamma g(\theta) \right\} = \text{Prox}_{\gamma n g}(\eta_n).$$

DA is in this sense related to proximal gradient methods, also called forward-backward splitting methods (see, e.g., Beck and Teboulle, 2009; Wright et al., 2009; Combettes and Pesquet, 2011). These methods are tailored to composite optimization problems: at each iteration f is linearized around the current iterate θ_n and they consider the following update

$$\theta_{n+1} = \operatorname*{argmin}_{\theta\in\mathcal{X}} \left\{ \langle \gamma\nabla f(\theta_n), \theta\rangle + \gamma g(\theta) + \frac{1}{2}\|\theta-\theta_n\|_2^2 \right\} = \text{Prox}_{\gamma g}(\theta_n - \gamma\nabla f(\theta_n)).$$

Note the difference with DA which considers a dual iterate and a proximal operator for the function $n\gamma g$ instead of γg (see additional insights in Section 4).

From non-smooth to smooth optimization. DA was initially introduced by Nesterov (2009) to optimize a non-smooth function f with possibly convex constraints ($g = 0$ or $g = 1_{\mathcal{C}}$). It was extended to the general stochastic composite case by Xiao (2010) who defined the iteration as

$$\theta_n = \operatorname*{argmin}_{\theta\in\mathcal{X}} \left\{ \frac{1}{n}\sum_{i=0}^{n-1} \langle z_i, \theta\rangle + g(\theta) + \frac{\beta_n}{n} h(\theta) \right\},$$

where z_i is an unbiased estimate[1] of a subgradient in $\partial f(\theta_i)$ and $(\beta_n)_{n\geq 1}$ a nonnegative and nondecreasing sequence of real numbers. This formulation is equivalent to Eq. (3) for constant sequences $\beta_n = 1/\gamma$. Xiao (2010) proved convergence rates of order $O(1/\sqrt{n})$ for convex problems with decreasing step-size $C/\sqrt{n}$ and $O(1/(\mu n))$ for problems with μ-strongly convex regularization with constant step-size $1/\mu$. DA was also studied with decreasing step-sizes in the distributed case by Duchi et al. (2012); Dekel et al. (2012); Colin et al. (2016) and combined with the alternating direction method of multipliers (ADMM) by Suzuki (2013). It was further shown to be very efficient in manifold identification by Lee and Wright (2012) and Duchi and Ruan (2016).

Relationship with mirror descent. The DA method should be associated with its cousin mirror descent algorithm (referred to from now on as "MD"), introduced by Nemirovski and Yudin (1979) for the constrained case and written under its modern proximal form by Beck and Teboulle (2003)

$$\theta_n = \operatorname*{argmin}_{\theta\in\mathcal{X}} \left\{ \gamma\langle \nabla f(\theta_{n-1}), \theta\rangle + D_h(\theta, \theta_{n-1}) \right\},$$

[1] Their results remain true in the more general setting of online learning.

where we denote by $D_h(\alpha, \beta) = h(\alpha) - h(\beta) - \langle \nabla h(\beta), \alpha - \beta \rangle$ the Bregman divergence associated with h. Moreover it was later extended to the general composite case by Duchi et al. (2010)

$$\theta_n = \operatorname*{argmin}_{\theta \in \mathcal{X}} \left\{ \gamma \langle \nabla f(\theta_{n-1}), \theta \rangle + \gamma g(\theta) + D_h(\theta, \theta_{n-1}) \right\}. \tag{4}$$

DA was initially motivated by Nesterov (2009) to avoid new gradients to be taken into account with less weight than previous ones. However, as an extension of the Euclidean case, DA essentially differs from MD on the way the regularization component is dealt with. See more comparisons in Section 4.

Relationship with online learning. DA was traditionally studied under the online learning setting (Zinkevich, 2003) of regret minimization and is related to the "follow the leader" approach (see, e.g., Kalai and Vempala, 2005) as noted by McMahan (2011). More generally, the DA method may be cast in the primal-dual algorithmic framework of Shalev-Shwartz and Singer (2006) and Shalev-Shwartz and Kakade (2009).

2.3 Deterministic convergence result for dual averaging

In this section we present the convergence properties of the DA method for optimizing deterministic composite problems of the form in Eq. (2), for any smooth function f (see proof in Appendix B).

Proposition 1. *Assume **(A1-4)**. For any step-size γ such that $h - \gamma f$ is convex on $\mathring{\mathcal{X}}$ we have for all $\theta \in \mathcal{X}$*

$$\psi(\theta_n) - \psi(\theta) \leq \frac{D_h(\theta, \theta_0)}{\gamma(n+1)}.$$

Moreover assume $g = 0$, and there exists $\mu \in \mathbb{R}$ such that $f - \mu h$ is also convex on $\mathring{\mathcal{X}}$ then we have for all $\theta \in \mathcal{X}$

$$f(\theta_n) - f(\theta) \leq (1 - \gamma\mu)^n \frac{D_h(\theta, \theta_0)}{\gamma}.$$

We can make the following remarks:

- We adapt the proof of Beck and Teboulle (2003) to the composite case and the DA method by including the regularization component g in the Bregman divergence. If g was differentiable we would simply use $D_{h_n} = D_{h + n\gamma g}$ and prove the following recursion:

$$\begin{aligned} D_{h_n}(\theta_*, \theta_n) - D_{h_{n-1}}(\theta_*, \theta_{n-1}) = &- D_{h_{n-1}}(\theta_n, \theta_{n-1}) + \gamma \langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta_n \rangle \\ &- \gamma(g(\theta_n) - g(\theta)) - \gamma \langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta_* \rangle. \end{aligned}$$

  Since g is not differentiable, we extend instead the notion of Bregman divergence to the non-smooth case in Appendix B.2 and show the proof works in the same way.

- **Related work**: DA was first analyzed for smooth functions in the non-composite case where $g = 0$, by Dekel et al. (2012) in the stochastic setting and by Lu et al. (2016) in the deterministic setting. A result on MD with analogue assumptions is presented by Bauschke et al. (2016) but depends on a symmetry measure of the Bregman divergence D_h which reflects

how different are $D_h(\alpha, \beta)$ and $D_h(\beta, \alpha)$, and the bound is not as simple. The technique to extend the Bregman divergence to analyze the regularization component has its roots in the time-varying potential method in online learning (Cesa-Bianchi and Lugosi, 2006, Chapter 11.6) and the "follow the regularized leader" approach (Abernethy et al., 2008).

- This convergence rate is suboptimal for the class of addressed problems. Indeed accelerated gradient methods achieve the convergence rate of $O(L/n^2)$ in the composite setting (Nesterov, 2013), such a rate being optimal for optimizing smooth functions among first-order techniques that can access only sequences of gradients (Nesterov, 2004).
- Classical results on the convergence of optimization algorithms in non-Euclidean geometries assume on one hand that the function h is strongly convex and on the other hand the function f is Lipschitz or smooth. Following Bauschke et al. (2016); Lu et al. (2016), we consider a different assumption which combines the smoothness of f and the strong convexity of h on the single condition $h - \gamma f$ convex. For the Euclidean geometry where $h(\theta) = \frac{1}{2}\|\theta\|_2^2$, this condition is obviously equivalent to the smoothness of the function f with regards to the ℓ_2-norm. Moreover, under arbitrary norm $\|\cdot\|$, this is also equivalent to assuming h μ-strongly convex and f L-smooth (with respect to this norm). However it is much more general and may hold even when f is non-smooth, which precisely justifies the introduction of this condition (see examples described by Bauschke et al., 2016; Lu et al., 2016).
- The bound adapts to the geometry of the function h through the Bregman divergence between the starting point θ_0 and the solution θ_* and the step-size γ which is controlled by h. Therefore the choice of h influences the constant in the bound. Examples are provided in Appendix I.

3 Stochastic convergence results for quadratic functions

In this section, we consider a symmetric positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ and a convex quadratic function f defined as

(A5) $$f(\theta) = \tfrac{1}{2}\langle \theta, \Sigma\theta\rangle - \langle q, \theta\rangle, \quad \text{with } q \in \mathbb{R}^d \text{ in the column space of } \Sigma,$$

so that f has a global minimizer $\theta_\Sigma \in \mathbb{R}^d$. Without loss of generality[2], Σ is assumed invertible, though its eigenvalues could be arbitrarily small. The global solution is known to be $\theta_\Sigma = \Sigma^{-1}q$, but the inverse of the Hessian is often too expensive to compute when d is large. The function may be simply expressed as $f(\theta_n) = \frac{1}{2}\langle \theta_n - \theta_\Sigma, \Sigma(\theta_n - \theta_\Sigma)\rangle + f(\theta_\Sigma)$ and the excess of the cost function $\psi = f + g$ as

$$\psi(\theta_n) - \psi(\theta_*) = \langle \theta_* - \theta_\Sigma, \Sigma(\theta_n - \theta_*)\rangle + g(\theta_n) - g(\theta_*) \text{ (linear part)} \\ + \frac{1}{2}\langle \theta_n - \theta_*, \Sigma(\theta_n - \theta_*)\rangle \text{ (quadratic part)} .$$

The first-order condition of the optimization problem in Eq. (2) is $0 \in \nabla f(\theta_*) + \partial g(\theta_*)$ and by convexity of g we have $g(\theta_n) - g(\theta_*) \geq \langle z, \theta_n - \theta_*\rangle$ for any $z \in \partial g(\theta_*)$. Therefore this implies

[2] By decomposing θ in $\theta = \theta_\| + \theta_\perp$ with $\theta_\perp \in \text{Null}(\Sigma)$ and $\langle \theta_\perp, \theta_\|\rangle = 0$ and considering $\psi(\theta) = f(\theta_\|) + \tilde{g}(\theta_\|)$ where $\tilde{g}(\theta_\|) = \inf_{\theta_\perp \in \text{Null}(\Sigma)} g(\theta_\perp + \theta_\|)$.

that the linear part $g(\theta_n) - g(\theta_*) + \langle \nabla f(\theta_*), \theta_n - \theta_* \rangle$ is non-negative and we have the bound

$$\frac{1}{2}\|\theta_n - \theta_*\|_\Sigma^2 \leq \psi(\theta_n) - \psi(\theta_*). \tag{5}$$

We derive, in this section, convergence results in terms of the distance $\|\theta_n - \theta_*\|_\Sigma$ which takes into account the ill-conditioning of the matrix Σ and is a lower bound in the excess of function values. Furthermore it directly implies classical results for strongly convex problems.

In many practical situations, the gradient of f is not available for the recursion in Eq. (3), and we have only access to an unbiased estimate $\nabla f_{n+1}(\theta_n)$ of the gradient of f at θ_n. We consider in this case the stochastic dual averaging method (referred to from now on as "SDA") defined the same way as DA as

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma \nabla f_n(\theta_{n-1}) \\ \theta_n &= \nabla h_n^*(\eta_n), \end{aligned} \tag{6}$$

for $\theta_0 \in \overset{\circ}{\operatorname{dom}} h$ and $\eta_0 = \nabla h(\theta_0)$. Here we consider the stochastic approximation framework (Kushner and Yin, 2003). That is, we let $(\mathcal{F}_n)_{n\geq 0}$ be an increasing family of σ-fields such that for each $\theta \in \mathbb{R}^d$ and for all $n \geq 1$ the random variable $\nabla f_n(\theta)$ is square-integrable and $\mathcal{F}_n$-measurable with $\mathbb{E}[\nabla f_n(\theta)|\mathcal{F}_{n-1}] = \nabla f(\theta)$. This includes (but also extends) the usual machine learning situation where ∇f_n is the gradient of the loss associated with the n-th independent observation. We will consider in the following two different gradient oracles.

3.1 Additive noise

We study here the convergence of the SDA recursion defined in Eq. (6) under an additive noise model:

(A6) For all $n \geq 1$, $\nabla f_n(\theta) = \nabla f(\theta) - \xi_n$, where the noise $(\xi_n)_{n\geq 1}$ is a square-integrable martingale difference sequence (i.e., $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$) with bounded covariance $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq C$.

With this oracle and for the quadratic function f, SDA takes the form

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma(\Sigma\theta_{n-1} - q) + \gamma\xi_n \\ \theta_n &= \nabla h_n^*(\eta_n). \end{aligned} \tag{7}$$

We obtain the following convergence result on the average $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$ which is an extension of results from Bach and Moulines (2013) to non-Euclidean geometries and to composite settings (see proof in Appendix C).

Proposition 2. *Assume **(A2-6)**. Consider the recursion in Eq. (7) for any constant step-size γ such that $h - \gamma f$ is convex. Then*

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\min\left\{\frac{D_h(\theta_*, \theta_0)}{\gamma n}; \frac{\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2}{(\gamma n)^2}\right\} + \frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

We can make the following observations:

- The proof in the Euclidean case (Bach and Moulines, 2013) highly uses the equality $\theta_n - \theta_\Sigma = (I - \gamma\Sigma)(\theta_{n-1} - \theta_\Sigma)$ which is no longer available in the non-Euclidean or proximal cases. Instead we adapt the classic proof of convergence of averaged SGD of Polyak and Juditsky (1992) which rests upon the expansion $\sum_{k=0}^n \nabla f_{k+1}(\theta_k) = \sum_{k=0}^n (\eta_k - \eta_{k+1})/\gamma = (\eta_0 - \eta_{n+1})/\gamma$. The crux of the proof is then to consider the difference between the iterations with and without noise, $\eta_n^{\rm sto} - \eta_n^{\rm det}$, which happens to satisfy a similar recursion as Eq. (7) but started from the solution θ_*. The quadratic nature of f is used twice: (a) to bound $\|\eta_n^{\rm sto} - \eta_n^{\rm det}\|_{\Sigma^{-1}} \sim \sqrt{n}$, and (b) to expand $\nabla f(\bar\theta_n) = \overline{\nabla f(\theta_n)} \sim \frac{\eta_n^{\rm sto} - \eta_0}{\gamma n} + 1/\sqrt{n}$.
- As for Proposition 1, the constraint on the step-size γ depends on the function h. Moreover the step-size γ is constant, contrary to previous works on SDA (Xiao, 2010) which prove results for decreasing step-size $\gamma_n = C/\sqrt{n}$ for the convex case (and with a convergence rate of only $O(1/\sqrt{n})$).
- The first term is the "bias" term. It only depends on the "distance" from the initial point θ_0 to the solution θ_* as the minimum of two terms. The first one recovers the deterministic bound of Proposition 1. The second one, specific to quadratic objectives, leads to an accelerated rate of $O(1/n^2)$ for some good starting points such that $\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2 < \infty$, thus extending the result from Flammarion and Bach (2015).
- The second term is the "variance" term which depends on the noise in the gradients. When the noise is structured (such as for least-squares regression), i.e, there exists $\sigma > 0$ such that $C \preccurlyeq \sigma^2 \Sigma$, the variance term becomes $\frac{\sigma^2 d}{n}$ which is optimal over all estimators in $\mathbb{R}^d$ without regularization (Tsybakov, 2008). However the regularization g does not bring statistical improvement as possible, for instance, with ℓ_1-regularization. We believe this is due to our proof technique. Indeed, in the case of linear constraints, Duchi and Ruan (2016) recently showed that the primal iterates (θ_n) follow a central limit theorem (CLT), namely $\sqrt{n}\bar\theta_n$ is asymptotically normal with a covariance precisely restricted to the active constraints. This supports that SDA may leverage the regularization (the active constraints in their case) to get better statistical performance. We leave such non-asymptotic results to future work.

Assumption **(A6)** on the gradient noise is quite general, since the noise (ξ_n) is allowed to be a martingale difference sequence (correct conditional expectation given the past, but not necessarily independence from the past). However it is not verified by the oracle corresponding to regular SDA for least-squares regression, where the noise combines both an additive and a multiplicative part, and its covariance is then no longer bounded in general (it will be for g the indicator function of a bounded set).

3.2 Least-squares regression

We consider now the least-squares regression framework, i.e, risk minimization with the square loss. Following Bach and Moulines (2013), we assume that:

(A7) The observations $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $n \geq 1$, are i.i.d. distributed with finite variances $\mathbb{E}\|x_n\|_2^2 < \infty$ and $\mathbb{E}y_n^2 < \infty$.

(A8) We consider the *least-squares regression* problem which is the minimization of the quadratic function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$.

(A9) We denote by $\Sigma = \mathbb{E}[x_n \otimes x_n]$ the population covariance matrix, which is the Hessian of f at all points. Without loss of generality, we reduce $\mathbb{R}^d$ to the minimal subspace where all x_n, $n \geq 1$, lie almost surely. Therefore Σ is invertible and all the eigenvalues of Σ are strictly positive, even if they may be arbitrarily small.

(A10) We denote the residual by $\xi_n = (y_n - \langle \theta_*, x_n \rangle) x_n$. We have $\mathbb{E}[\xi_n] = 0$ but $\mathbb{E}[\xi_n | x_n] \neq 0$ in general (unless the model is well-specified). There exists $\sigma > 0$ such that $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \sigma^2 \Sigma$.

(A11) There exists $\kappa > 0$ such that for all $z \in \mathbb{R}^d$, $\mathbb{E}\langle z, x_n \rangle^4 \leq \kappa \langle z, \Sigma z \rangle$.

(A12) The function g is lower bounded by some constant which is assumed by sake of simplicity to be 0.

(A13) There exists $L > 0$ such that $Lh - \frac{1}{2}\| \cdot \|_\Sigma^2$ is convex.

Assumptions **(A7-9)** are standard for least-squares regression, while Assumption **(A10)** defines a bounded statistical noise. Assumption **(A11)** is commonly used in the analysis of least-mean-square algorithms (Macchi, 1995) and says the projection of the covariates x_n on any direction $z \in \mathbb{R}^d$ have a bounded *kurtosis*. It is true for Gaussian vectors with $\kappa = 3$. Assumption **(A13)** links up the geometry of the function h and the objective function f; for example for ℓ_p-geometries, L is proportional to $\mathbb{E}\|x\|_q^2$ where $1/p + 1/q = 1$ (see Corollary 2 in Appendix I).

For the least-squares regression problem, the SDA algorithm defined in Eq. (6) takes the form:

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma\big(\langle x_n, \theta_{n-1} \rangle - y_n\big) x_n \\ \theta_n &= \nabla h_n^*(\eta_n). \end{aligned} \tag{8}$$

This corresponds to a stochastic oracle of the form $\nabla f_n(\theta) = (\Sigma + \zeta_n)(\theta - \theta_\Sigma) - \xi_n$ for $\theta \in \mathbb{R}^d$, with $\zeta_n = x_n \otimes x_n - \Sigma$. This oracle combines an additive noise ξ_n satisfying the previous Assumption **(A6)** and a multiplicative noise ζ_n which is harder to analyze.

We obtain a similar result compared to Proposition 2 at the cost of additional corrective terms.

Proposition 3. *Assume **(A2-4)** and **(A7-13)**. Consider the recursion in Eq. (8) for any constant step-size γ such that $\gamma \leq \frac{1}{4\kappa L d}$. Then*

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{32d}{n}\big(\sigma^2 + \kappa \|\theta_* - \theta_\Sigma\|_\Sigma^2\big) + \frac{16\kappa d}{n^2}\left(\frac{5 D_h(\theta_*, \theta_0)}{\gamma} + g(\theta_0)\right).$$

We can make the following remarks:

- The proof technique is similar to the one of Proposition 2. Nevertheless its complexity comes from the extra multiplicative noise ζ_n in the gradient estimate (see Appendix D).
- The result is only proven for $\gamma \leq 1/(4\kappa L d)$ which seems to be a proof artifact. Indeed we empirically observed (see Section 5) that the iterates still converge to the solution for all $\gamma \leqslant 1/(2\mathbb{E}\|x_n\|_2^2)$.
- The global bound leads to a rate of $O(1/n)$ without strong convexity, which is optimal for stochastic approximation, even with strong convexity (Nemirovsky and Yudin, 1983). We recover the terms of Proposition 2 pertubed by: (a) one corrective term of order $O(d/n)$

which depends on the distance between the solution θ_* and the global minimizer θ_Σ of the quadratic function f, which corresponds to the covariance of the multiplicative noise at the optimum, and (b) two residual terms of order $O(d/n^2)$. It would be interesting to study whether these two terms can be removed.

- As in Proposition 2, the bias is also $O\big(\frac{1}{(\gamma n)^2}\|\nabla h(\theta_0) - \nabla h(\theta_*)\|^2_{\Sigma^{-1}}\big)$ for specific starting points (see proof in Appendix D for details).
- It is worth noting that in the constrained case ($g = 1_{\mathcal{C}}$ for a bounded convex set $\mathcal{C}$), the covariance of the noisy oracle is simply bounded by $(\kappa \operatorname{tr}\Sigma r^2 + \sigma^2)\Sigma$ where we denote by $r = \max_{\theta\in\mathcal{C}} \|\theta - \theta_\Sigma\|_2$ (see Appendix D.1 for details). Therefore Proposition 2 already implies $\frac{1}{2}\mathbb{E}\|\bar\theta_n - \theta_*\|^2_\Sigma \leq 2\frac{D_h(\theta_*,\theta_0)}{\gamma n} + \frac{8d}{n}(\sigma^2 + \kappa r^2 \operatorname{tr}\Sigma)$. Moreover the result holds then for any step-size $\gamma \leqslant 1/L$, which is bigger than allowed for $g = 0$ (Bach and Moulines, 2013).

3.3 Convergence results on the objective function

In this section we present the convergence properties of the SDA method on the objective function $\psi = f + g$ rather than on the norm $\|\cdot\|_\Sigma$.

We first start with a disclaimer: it is not possible to obtain general non-asymptotic results on the convergence of the SDA iterates in term of function values without additional assumptions on the regularization g. We indeed show in Appendix E that, even in the simple case of a linear function $f(\theta) = \langle a, \theta\rangle$, for $a \in \mathbb{R}^d$, we can always find, for any finite time horizon N, a quadratic non-strongly convex regularization function g_N such that for any unstructured noise of variance σ^2, the function value $\psi_N(\theta) = f(\theta) + g_N(\theta)$ evaluated in the SDA iterates at time N is lowerbounded by

$$\psi_N(\bar\theta_N) - \psi_N(\theta_*) \geq \frac{\sigma^2}{12}.$$

This lower bound is specific to the SDA algorithm and we underline that the regularization g_N depends on the horizon N. However this result still prevents the possibility of a universal non-asymptotic convergence result on the function value for the SDA iterates for general quadratic and linear functions. We note that this does not apply to the setting of Proposition 2 and Proposition 3 since $\Sigma = 0$ for a linear function and the vector q defining the linear term $\langle q, \theta\rangle$ cannot be in the column space of Σ, thus violating Assumption **(A5)**. We conjecture that in the setting of Assumption **(A5)**, the lower bound is $O(1/\sqrt{n})$ as well.

We now provide some specific examples for which we can prove convergence in function values.

Quadratic objectives with smooth regularization. When there exists a constant $L_g \geq 0$ such that $L_g f - g$ is convex on $\mathring{\mathcal{X}}$ then results from Propositions 2 and 3 directly imply convergence of the composite objective to the optimum through

$$\psi(\bar\theta_n) - \psi(\theta_*) \leq \frac{(L_g + 1)}{2}\|\bar\theta_n - \theta_*\|^2_\Sigma = O(1/n),$$

with precise constants from Propositions 2 and 3. Indeed we have in that case $(L_g+1)f - \psi$ convex and this would be directly implied by Proposition 4 in Appendix B.

An easy but still interesting application is the non-regularized case ($g = 0$) when the optimum θ_* is the global optimum θ_Σ of f, because then $\psi(\theta) - \psi(\theta_*) = \frac{1}{2}\|\theta - \theta_*\|_\Sigma^2$. Thus this extends previous results on function values (Dieuleveut et al., 2016) to non-Euclidean geometries.

Constrained problems. When g is the indicator function of a convex set $\mathcal{C}$ then by definition the primal iterate $\theta_n \in \mathcal{C}$ and by convexity $\bar{\theta}_n \in \mathcal{C}$. Therefore $\psi(\bar{\theta}_n) = f(\bar{\theta}_n) + 1_\mathcal{C}(\bar{\theta}_n) = f(\bar{\theta}_n)$ and we obtain with the Cauchy-Schwarz inequality:

$$\begin{aligned} f(\bar{\theta}_n) - f(\theta_*) &= \langle \nabla f(\theta_*), \bar{\theta}_n - \theta_* \rangle + \frac{1}{2}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \\ &\leq \|\theta_* - \theta_\Sigma\|_2 \|\bar{\theta}_n - \theta_*\|_\Sigma + \frac{1}{2}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 = O\Big(\frac{\|\theta_* - \theta_\Sigma\|_2}{\sqrt{n}}\Big), \end{aligned}$$

with precise constants from Propositions 2 and 3. Hence we obtain a global rate of order $O(1/\sqrt{n})$ for the convergence of the function value in the constrained case.

These rates may be accelerated to $O(1/n)$ for certain specific convex constraints or when the global optimum $\theta_\Sigma \in \mathcal{C}$; Duchi and Ruan (2016) recently obtained asymptotic convergence results for the iterates in the cases of linear and ℓ_2-ball constraints for linear objective functions. Their results can be directly extended to asymptotic convergence of function values and very probably to all strongly convex sets (see, e.g., Vial, 1983). However, even for the simple ℓ_2-ball constrained problem, we were not able to derive non-asymptotic convergence rates for function values.

However the global rate of order $O(1/\sqrt{n})$ is statically non-improvable in general. In Appendix F, we relate the stochastic convex optimization problem (Agarwal et al., 2012) to the statistical problem of convex aggregation of estimators (Tsybakov, 2003; Lecué, 2006). These authors showed lower bounds on the performance of such estimators which provide us lower bounds on the performance of any stochastic algorithm to solve constrained problems. In Proposition 7 and Proposition 9 of Appendix F, we derive more precisely lower bound results for linear and quadratic functions for certain ranges of n and d confirming the optimality of the convergence rate $O(1/\sqrt{n})$. This being said, in our experiments in Section 5, we observed that the convergence of function values follows closely the convergence in the Mahalanobis distance.

4 Parallel between dual averaging and mirror descent

In this section we compare the behaviors of DA and MD algorithms, by highlighting their similarities and differences, in particular in terms of continuous-time interpretation.

4.1 Lazy versus greedy projection methods

DA and MD are often described in the online-learning literature as "lazy" and "greedy" projection methods (Zinkevich, 2003). Indeed, the difference between these two methods is more apparent in the Euclidean projection case (when $g = 1_\mathcal{C}$ and $h = \frac{1}{2}\|\cdot\|_2^2$). MD is then projected gradient descent and may be written under its primal-dual form as:

$$\eta_n^{\rm md} = \theta_{n-1}^{\rm md} - g_n^{\rm md} \quad \text{with} \quad g_n^{\rm md} \in \partial f(\theta_{n-1}^{\rm md}) \quad \text{and} \quad \theta_n^{\rm md} = \operatorname*{argmin}_{\theta \in \mathcal{C}} \|\eta_n^{\rm md} - \theta\|_2.$$

Whereas DA takes the form

$$\eta_n^{\text{da}} = \eta_{n-1}^{\text{da}} - g_n^{\text{da}} \quad \text{with} \quad g_n^{\text{da}} \in \partial f(\theta_{n-1}^{\text{da}}) \quad \text{and} \quad \theta_n^{\text{da}} = \operatorname*{argmin}_{\theta \in \mathcal{C}} \|\eta_n^{\text{da}} - \theta\|_2.$$

Therefore, imagining the subgradients g_n are provided by an adversary without the need to compute the primal sequence (θ_n), no projections are needed to update the dual sequence (η_n^{da}), and this one moves far away in the asymptotic direction of the gradient at the optimum $\nabla f(\theta_*)$. Furthermore the primal iterate θ_n^{da} is simply obtained, when required, by projecting back the dual iterate in the constraint set. Conversely, the MD dual iterate η_n^{md} update calls for θ_{n-1}^{md}, and therefore a projection step is unavoidable. Thereby MD iterates $(\eta_n^{\text{md}}, \theta_n^{\text{md}})$ are going, at each iteration, back-and-forth between the boundary and the outside of the convex set $\mathcal{C}$.

4.2 Strongly convex cases

MD converges linearly for smooth and strongly convex functions f, in the absence of a regularization component (Lu et al., 2016) or for Euclidean geometries (Nesterov, 2013). However we were not able to derive faster convergence rates for DA when the function f or the regularization g are strongly convex. Moreover the only results we found in the literature are about (a) an alteration of the dual gradient method (Devolder et al., 2013, Section 4) which is itself a modification of DA with an additional projection step proposed by Nesterov (2013) for smooth optimization, (b) the strongly convex regularization g which enables Xiao (2010) to obtain a $O(1/\mu n)$ convergence rate in the stochastic case.

At the simplest level, for $h = \frac{1}{2}\|\cdot\|_2^2$ and $f = 0$, MD is equivalent to the proximal point algorithm (Martinet, 1970) $\theta_n^{\text{md}} = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left\{ g(\theta) + \frac{1}{\gamma}\|\theta - \theta_{n-1}^{\text{md}}\|_2^2 \right\}$, whereas DA, which is not anymore iterative, is such that $\theta_n^{\text{da}} = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left\{ g(\theta) + \frac{1}{\gamma n}\|\theta\|_2^2 \right\}$. For the squared ℓ_2-regularization $g(\theta) = \frac{\nu}{2}\|\theta - \theta_*\|_2^2$, we compute exactly (see Appendix G)

$$g(\theta_n^{\text{md}}) - g(\theta_*) = \left(\frac{1}{\gamma\nu}\right)^n [g(\theta_0^{\text{md}}) - g(\theta_*)] \quad \text{and} \quad g(\theta_n^{\text{da}}) - g(\theta_*) = \frac{g(\theta_*)}{(\gamma n)^2}.$$

Therefore the convergence of DA can be dramatically slower than MD. However when noise is present, its special structure may be leveraged to get interesting results.

4.3 Continuous time interpretation of DA et MD

Following Nemirovsky and Yudin (1983); Krichene et al. (2015); Wibisono et al. (2016) we propose a continuous interpretation of these methods for g twice differentiable. Precise computations are derived in Appendix H.

The MD iteration in Eq. (4) may be viewed as a forward-backward Euler discretization of the MD ODE

$$\dot{\theta} = -\nabla^2 h(\theta)^{-1}[\nabla f(\theta) + \nabla g(\theta)]. \tag{9}$$

On the other hand, the ODE associated to DA takes the form

$$\dot{\theta} = -\nabla^2 (h(\theta) + t g(\theta))^{-1}(\nabla f(\theta) + \nabla g(\theta)). \tag{10}$$

It is worth noting that these ODEs are very similar, with an additional term $tg(\theta)$ in the inverse mapping $\nabla^2(h(\theta) + tg(\theta))^{-1}$ which may slow down the DA dynamics.

In analogy with the discrete case, the Bregman divergences D_h and D_{h+tg} are respectively Lyapunov functions for the MD and the DA ODEs (see, e.g., Krichene et al., 2015) and we notice in Appendix H the continuous time argument really mimics the proof of Proposition 1 without the technicalities associated with discrete time. Moreover we recover the variational interpretation of Krichene et al. (2015); Wibisono et al. (2016); Wilson et al. (2016): the Lyapunov function generates the dynamic in the sense that a function L is first chosen and secondly a dynamics, for which L is a Lyapunov function, is then designed. In this way MD and DA are the two different dynamics associated to the two different Lyapunov functions D_h and D_{h+tg}. We also provide in Appendix H a slight extension to the noisy-gradient case.

5 Experiments

In this section, we illustrate our theoretical results on synthetic examples. We provide additional experiments on a standard machine learning benchmark in Appendix K.

Simplex-constrained least-squares regression with synthetic data. We consider normally distributed inputs $x_n \in \mathbb{R}^d$ with a covariance matrix Σ that has random eigenvectors and eigenvalues $1/k$, for $k = 1, \dots, d$ and a random global optimum $\theta_\Sigma \in [0, +\infty)^d$. The outputs y_n are generated from a linear function with homoscedatic noise with unit signal to noise-ratio ($\sigma^2 = 1$). We denote by $R^2 = \operatorname{tr} \Sigma$ the average radius of the data and we show results averaged over 10 replications.

We consider the problem of least-squares regression constrained on the simplex Δ_d of radius $r = \|\theta_\Sigma\|_1/2$, i.e., $\min_{\theta \in r\Delta_d} \mathbb{E}(\langle x_n, \theta \rangle - y_n)^2$, for $d = 100$. We compare the performance of SDA and SGD algorithms with different settings of the step-size γ_n, constant or proportional to $1/\sqrt{n}$. In the left plot of Figure 1 we show the performance on the objective function and on the right plot, we show the performance on the squared Mahalanobis norm $\| \cdot \|_\Sigma^2$. All costs are shown in log-scale, normalized so that the first iteration leads to $f(\theta_0) - f(\theta_*) = 1$. We can make the following observations (we only show results on Euclidean geometry since results under the negative entropy geometry were very similar):

- With constant step-size, SDA converges to the solution at rate $O(1/n)$ whereas the SGD algorithm does not converge to the optimal solution.
- With decaying step-size $\gamma_n = 1/(2R^2\sqrt{n})$, SDA and SGD converge first at rate $O(1/\sqrt{n})$, then at rate $O(1/n)$, taking finally advantage of the strong-convexity of the problem.
- We note (a) there is no empirical difference between the performance on the objective function and the squared distance $\| \cdot \|_\Sigma^2$, (b) with decreasing step-size, SGD and SDA behave very similarly.

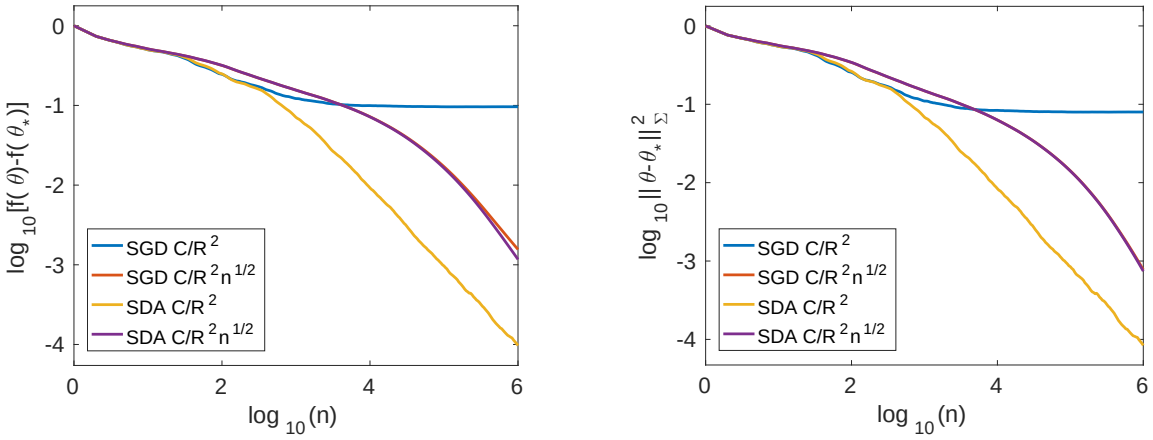

Figure 1: Simplex-constrained least-squares regression with synthetic data. Left: Performance on the objective function. Right: Performance on the Mahalanobis norm $\|\cdot\|_\Sigma^2$.

6 Conclusion

In this paper, we proposed and analyzed the first algorithm to achieve a convergence rate of $O(1/n)$ for stochastic composite objectives, without the need for strong convexity. This was achieved by considering a constant step-size and averaging of the primal iterates in the dual averaging method.

Our results only apply to expectations of quadratic functions (but to any additional potentially non-smooth terms). In fact, constant step-size stochastic dual averaging is not convergent for general smooth objectives; however, as done in the non-composite case by Bach and Moulines (2013), one could iteratively solved quadratic approximations of the smooth problems with the algorithm we proposed in this paper to achieve the same rate of $O(1/n)$, still with robustness to ill-conditioning and efficient iterations. Finally, it would be worth considering accelerated extensions to achieve a forgetting of initial conditions in $O(1/n^2)$.

Acknowledgements

The authors would like to thank Aymeric Dieuleveut and Damien Garreau for helpful discussions.

References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, pages 263–274, 2008.

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

F. Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1): 115–129, 2015.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, December 2013.

H. H. Bauschke and J. M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.

H. H. Bauschke and P. L. and Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 2016.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

L. M. Bregman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

I. Colin, A. Bellet, J. Salmon, and S. Clémençon. Gossip dual averaging for decentralized optimization of pairwise functions. In *Proceedings of the conference on machine learning (ICML)*, 2016.

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

A. Defazio, F. Bach, and S. 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.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 13:165–202, 2012.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers*, 2013016, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *arXiv preprint arXiv:1602.05419v2*, 2016.

J. Duchi and F. Ruan. Local asymptotics for some stochastic optimization problems: Optimality, constraint identification, and dual averaging. *arXiv preprint arXiv:1612.05612*, 2016.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the International Conference on Learning Theory (COLT)*, pages 14–26, 2010.

J. Duchi, A. Agarwal, and M. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Trans. Automat. Control*, 57(3):592–606, 2012.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2015.

C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the International Conference on Learning Theory (COLT)*, pages 1–11, 1999.

O. Hanner. On the uniform convexity of L^p and l^p. *Ark. Mat.*, 3:239–244, 1956.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *arXiv preprint arXiv:1610.03774*, 2016.

A. Juditsky and A. S. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *J. Comput. System Sci.*, 71(3):291–307, 2005.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inform. and Comput.*, 132(1):1–63, 1997.

K. C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM J. Control Optim.*, 35(4):1142–1168, 1997.

W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2845–2853, 2015.

H. Kushner and G G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer, 2003.

G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 364–378. Springer, Berlin, 2006.

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13 (4):1000–1022, 2007.

S. Lee and S. J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *J. Mach. Learn. Res.*, 13:1705–1744, 2012.

H. Lu, R. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *arXiv preprint arXiv:1610.05708*, 2016.

O. Macchi. *Adaptive processing: The least mean squares approach with applications in transmission*. Wiley West Sussex, 1995.

B. Martinet. Breve communication. régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis - Modlisation Mathmatique et Analyse Numrique*, 4:154–158, 1970.

H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *AISTATS*, pages 525–533, 2011.

J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris*, 255:2897–2899, 1962.

A. S. Nemirovski and D. B. Yudin. Effective methods for the solution of convex programming problems of large dimensions. *Èkonom. i Mat. Metody*, 15(1):135–152, 1979.

A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.*, 120(1, Ser. B):221–259, 2009.

Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140:125–161, 2013.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Trans. Inform. Theory*, 57(10):7036–7056, 2011.

I. Rish and G. Grabarnik. *Sparse modeling: theory, algorithms, and applications*. CRC press, 2014.

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

S. Shalev-Shwartz and S. m. Kakade. Mind the duality gap: Logarithmic regret algorithms for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1457–1464, 2009.

S. Shalev-Shwartz and Y. Singer. Online learning meets optimization in the dual. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 423–437. Springer, Berlin, 2006.

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the conference on machine learning (ICML)*, pages 392–400, 2013.

A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

J.-P. Vial. Strong and weak convexity of sets and functions. *Math. Oper. Res.*, 8(2):231–259, 1983.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47), 2016.

A. I. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635v3*, 2016.

S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the conference on machine learning (ICML)*, 2003.

A Unambiguity of the primal iterate

We describe here conditions under which the primal iterate θ_n in Eq. (3) is correctly defined. Since h is strictly convex, h_n^* is continuously differentiable on $\overset{\circ}{\text{dom}}\, h_n^*$ (see Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.1.1). Therefore the primal iterate θ_n is well defined if the dual iterate $\eta_n \in \overset{\circ}{\text{dom}}\, h_n^*$. It is, for example, the case under two natural assumptions as shown by the next lemma which is an adaption of Lemma 2 by Bauschke et al. (2016).

Lemma 1. *We make the following assumptions:*

(B1) *h or g is supercoercive.*

(B2) $\operatorname{argmin}_{\theta\in\mathcal{X}} \psi(\theta)$ *is compact and h bounded below.*

*Under **(B1)** or **(B2)** the primal iterates (θ_n) defined in Eq. (3) are well defined.*

Proof. Since h is strictly convex, h_n^* is continuously differentiable on $\overset{\circ}{\text{dom}}\, h_n^*$ (see Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.1.1). Therefore the primal iterate θ_n is well defined if the dual iterate $\eta_n \in \overset{\circ}{\text{dom}}\, h_n^*$.

- If h or g is supercoercive then h_n is supercoercive (see Bauschke and Combettes, 2011, Proposition 11.13) and it follows from Hiriart-Urruty and Lemaréchal (2001, Chapter E, Proposition 1.3.8) that $\text{dom}\, h_n^* = \mathbb{R}^d$.
- If $\operatorname{argmin}_{\theta\in\mathcal{X}}\{\psi(\theta)\}$ is compact then $\psi + 1_{\mathcal{X}}$ is coercive. Moreover

$$\begin{aligned}
h_n^*(\eta_n) &= \sup_{\theta\in\mathcal{X}} \left\{\langle \eta_n, \theta\rangle - h_n(\theta)\right\} \text{ since} \mathcal{X} \subset \overline{\text{dom}\, h} \\
&= -\inf_{\theta\in\mathcal{X}} \Big\{ h(\theta) + \gamma \sum_{i=1}^{n} (g(\theta) + f(\theta_{i-1}) + \langle \nabla f(\theta_{i-1}), \theta - \theta_{i-1}\rangle) \Big\} \\
&\quad + \gamma \sum_{i=1}^{n} (f(\theta_{i-1}) - \langle \nabla f(\theta_{i-1}), \theta_{i-1}\rangle) \\
&\leq -\inf_{\theta\in\mathcal{X}} \left\{ h(\theta) + n\gamma(g(\theta) + f(\theta)) \right\} \text{ by convexity of } f \\
&\quad + \gamma \sum_{i=1}^{n} (f(\theta_{i-1}) - \langle \nabla f(\theta_{i-1}), \theta_{i-1}\rangle).
\end{aligned}$$

  Therefore $\eta_n \in \overset{\circ}{\text{dom}}\, h_n^*$ since $\psi + 1_{\mathcal{X}}$ is coercive and h bounded below (see Bauschke and Combettes, 2011, Proposition 11.15).

□

B Proof of convergence of deterministic DA

We first describe a new notion of smoothness defined by Bauschke et al. (2016). Then we present our extension of the Bregman divergence to the non-smooth function g to finally prove Proposition 1.

B.1 A *Lipschitz-like/convexity condition*

Classical results on the convergence of optimization algorithms in non-Euclidean geometry assume on one hand that the function h is strongly convex and on the other hand the function f is Lipschitz or smooth. Following Bauschke et al. (2016); Lu et al. (2016), we consider a different assumption which combines the smoothness of f and the strong convexity of h on a single condition called *Lipschitz-like/Convexity Condition* by Bauschke et al. (2016) and denoted by **(LC)**:

(LC) There exists a constant $L \in \mathbb{R}$ such that $Lh - f$ is convex on $\overset{\circ}{\mathcal{X}}$.

For Euclidean geometry, this condition is obviously equivalent to the smoothness of the function f with regards to the ℓ_2-norm. Moreover, under an arbitrary norm $\|\cdot\|$, assuming h μ-strongly convex and f L-smooth clearly implies, by simple convex computation, **(LC)** with constant L/μ. However **(LC)** is much more general and may hold even when f is non-smooth what precisely justifies the introduction of this condition. Many examples are described by Bauschke et al. (2016); Lu et al. (2016). Furthermore this notion has the elegance of pairing well with Bregman divergences and leading to more refined proofs as shown in the following proposition which summarizes equivalent properties of **(LC)**.

Proposition 4 (Bauschke et al. (2016))**.** *Assume **(A1-4)**. For $L > 0$ the following conditions are equivalent:*

- *$Lh - f$ is convex on $\overset{\circ}{\mathcal{X}}$, i.e., **(LC)** holds,*
- *$D_f(\alpha, \beta) \leq L D_h(\alpha, \beta)$ for all $(\alpha, \beta) \in \mathcal{X} \times \overset{\circ}{\mathcal{X}}$.*

Furthermore, when f and h are assumed twice differentiable, then the above is equivalent to

$$\nabla^2 f(\theta) \preccurlyeq L \nabla^2 h(\theta) \quad \text{for all } \theta \in \overset{\circ}{\mathcal{X}}.$$

B.2 Generalized Bregman divergence

The Bregman divergence was defined by Bregman (1967) for a differentiable convex function h as

$$D_h(\alpha, \beta) = h(\alpha) - h(\beta) - \langle \nabla h(\beta), \alpha - \beta \rangle, \text{ for } (\alpha, \beta) \in \operatorname{dom} h \times \overset{\circ}{\operatorname{dom}} h. \tag{11}$$

It behaves as a squared distance depending on the function h and extends the computational properties of the squared ℓ_2-norm to non-Euclidean spaces. Indeed most proofs in Euclidean space rest upon the expansion $\|\theta_n - \theta_* - \gamma \nabla f(\theta_n)\|_2^2 = \|\theta_n - \theta_*\|_2^2 + \gamma^2 \|\nabla f(\theta_n)\|_2^2 - 2\gamma \langle \nabla f(\theta_n), \theta_n - \theta_* \rangle$

which is not available in non-Euclidean geometry. Therefore the Bregman divergence comes to rescue and is used to compute a deviation between the current iterate of the algorithm and the solution of the problem and, seemingly, used as an non-Euclidean Lyapunov function. It has been widely used in optimization (see, e.g., Bauschke and Borwein, 1997, for a review).

We follow this path and include the regularization component g of the objective function $\psi = f + g$ in the Bregman divergence for the sake of the analysis. If g was differentiable we would simply use $D_{h+n\gamma g}$. Since g is not differentiable, D_{h_n} is not well defined. However for $(\alpha, \eta) \in \operatorname{dom} h \times \overset{\circ}{\operatorname{dom}} h_n^*$, we denote by extension for $\theta = \nabla h_n^*(\eta)$:

$$\tilde{D}_n(\alpha, \eta) = h_n(\alpha) - h_n(\theta) - \langle \eta, \alpha - \theta \rangle. \tag{12}$$

This extension is different from the one defined by Kiwiel (1997). It is worth noting that if there exists μ such that $\alpha = \nabla h_n^*(\mu)$, we recover the classical formula $\tilde{D}_n(\alpha, \eta) = D_{h_n^*}(\eta, \mu)$ which is well defined since h_n^* is differentiable. Yet $\tilde{D}_n$ is defined more generally since such a μ does not always exist. The next lemma relates $\tilde{D}_n$ to D_h and is obvious if g is differentiable since $D_{h_n} = D_h + \gamma n D_g$.

Lemma 2. *Let $n \geq 0$, $\alpha \in$ dom h and $\eta \in \overset{\circ}{\operatorname{dom}} h_n^*$, then with $\theta = \nabla h_n^*(\eta)$,*

$$\tilde{D}_n(\alpha, \eta) \geq D_h(\alpha, \theta). \tag{13}$$

Proof. $\theta = \nabla h_n^*(\eta)$, thus $\eta \in \partial h_n(\theta)$ and by elementary calculus rule $\partial h_n(\theta) = \nabla h(\theta) + n\gamma \partial g(\theta)$. Consequently $\eta - \nabla h(\theta) \in n\gamma \partial g(\theta)$ and by convexity of g

$$\tilde{D}_n(\alpha, \eta) - D_h(\alpha, \theta) = n\gamma \left[g(\alpha) - g(\theta) - \left\langle \frac{\eta - \nabla h(\theta)}{\gamma n}, \alpha - \theta \right\rangle \right] \geq 0.$$

□

B.3 Proof of Proposition 1

We assume their exists a constant $L > 0$ such that $Lh - f$ is convex on $\mathring{\mathcal{X}}$ and we assume the step-size $\gamma \leq 1/L$. We first show that the Bregman divergence decreases along the iterates (see, e.g., Beck and Teboulle, 2003; Bach, 2015). For all $\theta \in \mathcal{X}$,

$$\begin{aligned}
\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) &= h_{n-1}(\theta_{n-1}) - h_n(\theta_n) + h_n(\theta) - h_{n-1}(\theta) \\
&\quad - \langle \eta_n, \theta - \theta_n \rangle + \langle \eta_{n-1}, \theta - \theta_{n-1} \rangle \\
&= h_{n-1}(\theta_{n-1}) - h_{n-1}(\theta_n) - \gamma(g(\theta_n) - g(\theta)) \\
&\quad + \langle \eta_{n-1}, \theta_n - \theta_{n-1} \rangle + \langle \eta_n - \eta_{n-1}, \theta_n - \theta \rangle \\
&= -\tilde{D}_{n-1}(\theta_n, \eta_{n-1}) - \gamma(g(\theta_n) - g(\theta)) - \gamma \langle \nabla f(\theta_{n-1}), \theta_n - \theta \rangle.
\end{aligned}$$

Therefore for all $\theta \in \mathcal{X}$,

$$\begin{aligned}
\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) = -\tilde{D}_{n-1}(\theta_n, \eta_{n-1}) + \gamma \langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta_n \rangle \\
- \gamma(g(\theta_n) - g(\theta)) - \gamma \langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta \rangle.
\end{aligned} \tag{14}$$

It follows from Proposition 4 and Lemma 2

$$f(\theta_n) - f(\theta_{n-1}) + \langle \nabla f(\theta_{n-1}), \theta_n - \theta_{n-1} \rangle \leq L D_h(\theta_n, \theta_{n-1}) \leq L D_{n-1}(\theta_n, \theta_{n-1}),$$

and from the convexity of f,

$$-\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta \rangle \leq f(\theta) - f(\theta_{n-1}).$$

And Eq. (14) is bounded by

$$\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) \leq \gamma(\psi(\theta) - \psi(\theta_n)) + (\gamma L - 1) D_h(\theta_n, \theta_{n-1}).$$

Thus for $\gamma \leq 1/L$,

$$\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) \leq \gamma(\psi(\theta) - \psi(\theta_n)).$$

Taking $\theta = \theta_{n-1}$ we note that the sequence $\{\psi(\theta_n)\}_{n \geq 0}$ is decreasing and we obtain for $\gamma \leq 1/L$,

$$\psi(\theta_n) - \psi(\theta) \leq \frac{1}{n+1} \sum_{k=0}^{n} [\psi(\theta_i) - \psi(\theta)] \leq \frac{D_h(\theta, \theta_0) - \tilde{D}_n(\theta, \eta_n)}{\gamma(n+1)}. \tag{15}$$

We assume now that the non-smooth part $g = 0$ and there exists $\mu \geq 0$ such that $f - \mu h$ is convex. So Proposition 4 implies

$$-\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta \rangle \leq f(\theta) - f(\theta_{n-1}) - \mu D_h(\theta, \theta_{n-1}),$$

which gives with Eq. (14) the better bound

$$D_h(\theta, \theta_n) - D_h(\theta, \theta_{n-1}) \leq \gamma(f(\theta) - f(\theta_n)) - \gamma\mu D_h(\theta, \theta_{n-1}) + (\gamma L - 1) D_h(\theta_n, \theta_{n-1}).$$

And for $\gamma \leq 1/L$, this can be simplified as

$$D_h(\theta, \theta_n) \leq (1 - \gamma\mu) D_h(\theta, \theta_{n-1}) + \gamma(f(\theta) - f(\theta_n)).$$

The sequence $\{f(\theta_n)\}_{n \geq 0}$ is still decreasing and we obtain by expanding the recursion

$$\begin{aligned}
D_h(\theta, \theta_n) &\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \sum_{k=1}^{n} (1 - \gamma\mu)^{n-k} \gamma(f(\theta) - f(\theta_k)) \\
&\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \sum_{k=1}^{n} (1 - \gamma\mu)^{n-k} \gamma(f(\theta) - f(\theta_k)) \\
&\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \sum_{k=1}^{n} (1 - \gamma\mu)^{n-k} \gamma(f(\theta) - f(\theta_n)) \\
&\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \gamma \frac{1 - (1 - \gamma\mu)^n}{\gamma\mu} (f(\theta) - f(\theta_n)).
\end{aligned}$$

Thus for all $\theta \in \mathcal{X}$,

$$\frac{1 - (1 - \gamma\mu)^n}{\mu} (f(\theta_n) - f(\theta)) + D_h(\theta, \theta_n) \leq (1 - \gamma\mu)^n D_h(\theta, \theta_0),$$

and

$$f(\theta_n) - f(\theta) \leq \frac{\gamma\mu(1 - \gamma\mu)^n}{1 - (1 - \gamma\mu)^n} \frac{D_h(\theta, \theta_0)}{\gamma} \leq (1 - \gamma\mu)^n \frac{D_h(\theta, \theta_0)}{\gamma},$$

since $(1 - \gamma\mu)^2 \leq 1 - \gamma\mu$ implies $\gamma\mu / (1 - (1 - \gamma\mu)^n) \leq 1$.

C Proof of Proposition 2

In this section, we will prove Proposition 2. The proof relies on considering the difference between the iteration with noise we denote by (η_n, θ_n) and without noise we denote by (ω_n, ϕ_n), which happens to verify a similar recursion as the SDA recursion.

- We first show in Lemma 3 that the distance $\mathbb{E}\|\eta_n - \omega_n\|^2_{\Sigma^{-1}}$ is of order n.
- Then in Lemma 4 we show that $\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|^2_{\Sigma}$ is of order $O(1/n)$, by: (a) noticing that $\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|^2_{\Sigma}$ is of order $\frac{\mathbb{E}\|\eta_n - \omega_n\|^2_{\Sigma^{-1}}}{n^2} + \frac{\text{variance}}{n}$, (b) combining this with the result of Lemma 3.

C.1 Two technical lemmas

We first present and prove two technical lemmas.

C.1.1 Bound on the difference of two dual iterates

In the following lemma we show that the difference between two dual iterates that follow the same recursion is of order n. This will be used with the iteration with noise (η_n, θ_n) and without noise (ω_n, ϕ_n).

Lemma 3. *Let us consider two sequences of iterates (μ_k, α_k) and (ν_k, β_k) which satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma\Sigma(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h_n^*(\mu_n)$ and $\beta_n = \nabla h_n^*(\nu_n)$ and assume that γ is such that $2h - \gamma f$ is convex then for all $n \geq 0$*

$$\mathbb{E}\|\mu_n - \nu_n\|^2_{\Sigma^{-1}} \leq \|\mu_0 - \nu_0\|^2_{\Sigma^{-1}} + n\gamma^2 \operatorname{tr}\Sigma^{-1}C.$$

Proof. We first expand the square.

$$\begin{aligned}
\|\mu_{n+1} - \nu_{n+1}\|^2_{\Sigma^{-1}} &= \|\mu_n - \nu_n\|^2_{\Sigma^{-1}} + \gamma^2\|[\Sigma(\alpha_n - \beta_n) - \xi_{n+1}]\|^2_{\Sigma^{-1}} \\
&\quad -2\gamma\langle\Sigma(\alpha_n - \beta_n) - \xi_{n+1}, \Sigma^{-1}(\mu_n - \nu_n)\rangle \\
&= \|\mu_n - \nu_n\|^2_{\Sigma^{-1}} + \gamma^2\|\alpha_n - \beta_n\|^2_{\Sigma} + \gamma^2\|\xi_{n+1}\|^2_{\Sigma^{-1}} \\
&\quad -2\gamma^2\langle\alpha_n - \beta_n, \xi_{n+1}\rangle - 2\gamma\langle\alpha_n - \beta_n - \Sigma^{-1}\xi_{n+1}, \mu_n - \nu_n\rangle.
\end{aligned}$$

And taking the expectation

$$\begin{aligned}
\mathbb{E}[\|\mu_{n+1} - \nu_{n+1}\|^2_{\Sigma^{-1}}|\mathcal{F}_n] &= \|\mu_n - \nu_n\|^2_{\Sigma^{-1}} + \gamma^2\mathbb{E}[\|\xi_{n+1}\|^2_{\Sigma^{-1}}|\mathcal{F}_n] \\
&\quad +\gamma^2\|\alpha_n - \beta_n\|^2_{\Sigma} - 2\gamma^2\mathbb{E}[\langle\alpha_n - \beta_n, \xi_{n+1}\rangle|\mathcal{F}_n] \\
&\quad -2\gamma\mathbb{E}[\langle\alpha_n - \beta_n) - \Sigma^{-1}\xi_{n+1}, \mu_n - \nu_n\rangle|\mathcal{F}_n] \\
&= \|\mu_n - \nu_n\|^2_{\Sigma^{-1}} + \gamma^2 \operatorname{tr}\Sigma^{-1}\mathbb{E}[\xi_{n+1} \otimes \xi_{n+1}|\mathcal{F}_n] \\
&\quad +\gamma^2\|\alpha_n - \beta_n\|^2_{\Sigma} - 2\gamma^2\langle\alpha_n - \beta_n, \mathbb{E}[\xi_{n+1}|\mathcal{F}_n]\rangle \\
&\quad -2\gamma\langle\alpha_n - \beta_n) - \Sigma^{-1}\mathbb{E}[\xi_{n+1}|\mathcal{F}_n], \mu_n - \nu_n\rangle \\
&= \|\mu_n - \nu_n\|^2_{\Sigma^{-1}} + \gamma^2 \operatorname{tr}\Sigma^{-1}C \\
&\quad +\gamma^2\|\alpha_n - \beta_n\|^2_{\Sigma} - 2\gamma\langle\alpha_n - \beta_n, \mu_n - \nu_n\rangle.
\end{aligned}$$

Moreover, using the definition of α_n and β_n,

$$\begin{aligned}\gamma\|\alpha_n-\beta_n\|_\Sigma^2-2\langle\alpha_n-\beta_n,\mu_n-\nu_n\rangle &= \langle\gamma\Sigma(\alpha_n-\beta_n)-2(\mu_n-\nu_n),\alpha_n-\beta_n\rangle\\ &= \langle\gamma\nabla f(\alpha_n)-\nabla f(\beta_n)-2(\nabla h(\alpha_n)-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle\\ &\quad -2\langle(\mu_n-\nabla h(\alpha_n))-(\nu_n-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle\\ &= \langle\nabla(\gamma f-2h)(\alpha_n)-\nabla(\gamma f-2h)(\beta_n),\alpha_n-\beta_n\rangle\\ &\quad -2\langle(\mu_n-\nabla h(\alpha_n))-(\nu_n-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle.\end{aligned}$$

Using the h-smoothness of f and assuming that γ is such $2h-\gamma f$ is convex,

$$\langle\nabla(\gamma f-2h)(\alpha_n)-\nabla(\gamma f-2h)(\beta_n),\alpha_n-\beta_n\rangle\leq 0,$$

and as explained in the proof of Lemma 2, $\mu_n-\nabla h(\alpha_n)\in\partial n\gamma g(\alpha_n)$ and $\nu_n-\nabla h(\beta_n)\in\partial n\gamma g(\beta_n)$ and consequently

$$\langle(\mu_n-\nabla h(\alpha_n))-(\nu_n-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle\leq 0,$$

by convexity of g. This explains that

$$\gamma\|\alpha_n-\beta_n\|_\Sigma^2-2\langle\alpha_n-\beta_n,\mu_n-\nu_n\rangle\leq 0.$$

Then, taking the global expectation, we have shown that

$$\mathbb{E}\|\mu_{n+1}-\nu_{n+1}\|_{\Sigma^{-1}}^2\leq\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2+\gamma^2\operatorname{tr}\Sigma^{-1}C,$$

which concludes the proof. □

C.1.2 Bound on the difference of the average of two primal iterates

In the following lemma we adapt the classic proof of averaged SGD by Polyak and Juditsky (1992) to show that the difference between two averaged primal iterates, which follow the same recursion, is of order $O(1/n)$.

Lemma 4. *Let us consider two sequences of iterates (μ_k,α_k) and (ν_k,β_k) which satisfy the recursion $\mu_n-\nu_n=\mu_{n-1}-\nu_{n-1}-\gamma\Sigma(\alpha_{n-1}-\beta_{n-1})+\gamma\xi_n$, $\alpha_n=\nabla h_n^*(\mu_n)$ and $\beta_n=\nabla h_n^*(\nu_n)$ and assume that γ is such that $2h-\gamma f$ is convex then for all $n\geq 0$*

$$\mathbb{E}\|\bar\alpha_n-\bar\beta_n\|_\Sigma^2\leq 4\frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2}+\frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

Proof. Let us consider two sequences of iterates (μ_k,α_k) and (ν_k,β_k) which satisfy the recursion $\mu_n-\nu_n=\mu_{n-1}-\nu_{n-1}-\gamma\Sigma(\alpha_{n-1}-\beta_{n-1})+\gamma\xi_n$, $\alpha_n=\nabla h_n^*(\mu_n)$ and $\beta_n=\nabla h_n^*(\nu_n)$. This can be written as

$$\Sigma(\alpha_n-\beta_n)=\frac{\mu_n-\nu_n-\mu_{n+1}+\nu_{n+1}}{\gamma}+\xi_{n+1}.$$

Thus we obtain

$$\Sigma^{1/2}\sum_{i=0}^{n-1}(\alpha_i-\beta_i) = \frac{\Sigma^{-1/2}(\mu_0-\nu_0-\mu_n+\nu_n)}{\gamma}+\sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1}.$$

Finally, using that by convexity $(a+b)^2 \leq 2(a^2+b^2)$, this leads to

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 2\mathbb{E}\Big\|\frac{\Sigma^{-1/2}(\mu_0-\nu_0)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1}\Big\|_2^2 + 2\mathbb{E}\Big\|\frac{\Sigma^{-1/2}(\mu_n-\nu_n)}{\gamma}\Big\|_2^2.$$

Using martingale second moment expansions, we obtain

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 2\mathbb{E}\frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + 2\frac{\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{2}{n^2}\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}).$$

We compute $\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}) = \sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}C = n\operatorname{tr}\Sigma^{-1}C$ and, using Lemma 3, we bound $\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2$ as

$$\frac{\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \leq \frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{1}{n}\operatorname{tr}\Sigma^{-1}C.$$

This implies the final bound

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 4\frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

□

C.2 Application of Lemma 4 to prove Proposition 2

First of all we define the sequence

$$\eta_n^* = \nabla h(\theta_*) - n\gamma\nabla f(\theta_*). \tag{16}$$

By definition of θ_*, $-\nabla f(\theta_*) \in \partial g(\theta_*)$ then $\eta_n^* \in \partial(h+n\gamma g)\theta_*$ and $\theta_* = \nabla h_n^*(\eta_n^*)$. Therefore the sequence η_n^* is obtained by iterating DA started from the solution of the problem θ_*.

We note then than Lemma 4 applied to $(\mu_n = \eta_n, \alpha_n = \theta_n)$ and $(\nu_n = \eta_n^*, \beta_n = \theta_*)$ gives the first bound of Proposition 2.

On the other hand, when considering the noiseless iterates (ω_n, ϕ_n) defined by $\omega_n = \omega_{n-1} - \gamma\Sigma(\phi_{n-1}-\theta_\Sigma)$ and $\phi_n = \nabla h_n^*(\omega_n)$, started from the same point $\phi_0 = \theta_0$, we obtain, following Proposition 1, for gamma such that $h - \gamma f$ is convex, the bound

$$\frac{1}{2}\|\bar{\phi}_n - \theta_*\|_\Sigma^2 \leq \psi(\bar{\phi}_n) - \psi(\theta_*) \leq \frac{D_h(\theta_*,\theta_0)}{\gamma n}.$$

Therefore, considering the difference between the semi-stochastic and the noiseless iterate $(\eta_n - \omega_n)$ which verifies the same equation $\eta_n - \omega_n = \eta_{n-1} - \omega_{n-1} - \gamma\Sigma(\theta_{n-1}-\phi_{n-1}) + \gamma\xi_n$ with $\theta_0 - \phi_0 = 0$ as initial value, we may apply Lemma 4 to show

$$\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 \leq \frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

And by the Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 &\leq 2\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 + 2\mathbb{E}\|\bar{\phi}_n - \theta_*\|_\Sigma^2 \\ &\leq \frac{8}{n}\operatorname{tr}\Sigma^{-1}C + 4\frac{D_h(\theta_*, \theta_0)}{\gamma n},\end{aligned}$$

which proves the second bound of Proposition 2.

It is worth noting that the condition on the step-size of Lemma 3 is less restrictive than in Proposition 2. Indeed for all γ such that $2h - \gamma f$ is convex, the difference between the dual iterates of the stochastic and deterministic recursions stay close but the deterministic iterates only converge to the solution for γ such that $h - \gamma f$ is convex.

D Proof of Proposition 3

In this section, we prove Proposition 3. The proof technique is similar to Proposition 2 but with the additional difficulty of the multiplicative noise.

We first note that Assumption **(A11)** is equivalent by the Cauchy-Schwarz inequality to

$$\mathbb{E}\langle x_n, Mx_n\rangle\langle x_n, Nx_n\rangle \leq \kappa \operatorname{tr}(M\Sigma)\operatorname{tr}(N\Sigma), \tag{17}$$

for all positive semi-definite symmetric matrices M and N (see, e.,g., proof in Dieuleveut et al., 2016). We will often use, in the following demonstrations, Eq. (17) and its direct corollary

$$\langle x_n, Mx_n\rangle x_n \otimes x_n \preccurlyeq \kappa \operatorname{tr}(M\Sigma)\Sigma, \tag{18}$$

without always referring to it.

D.1 A simple proof for the bounded constrained case

We first prove Proposition 3 for the constrained case. It is then a simple corollary of Proposition 2.

Let us denote by $\mathcal{C}$ a bounded convex set and consider the constrained problem ($g = \mathbf{1}_\mathcal{C}$). We remind that the general stochastic oracle for SDA in least-squares regression is

$$\nabla f_n(\theta) = (\Sigma + \zeta_n)(\theta - \theta_\Sigma) - \xi_n, \text{ for } \theta \in \mathbb{R}^d,$$

with $\zeta_n = x_n \otimes x_n - \Sigma$. We denote by $r = \max_{\theta\in\mathcal{C}} \|\theta - \theta_\Sigma\|_2$ and we show that the noise covariance is directly bounded, despite the multiplicative noise:

$$\mathbb{E}\Big[\big(\nabla f_n(\theta) - \nabla f(\theta)\big) \otimes \big(\nabla f_n(\theta) - \nabla f(\theta)\big)\Big] \preccurlyeq 2\mathbb{E}\big[\zeta_n(\theta - \theta_\Sigma) \otimes (\theta - \theta_\Sigma)\zeta_n\big] + 2\mathbb{E}\xi_n \otimes \xi_n,$$

and using Assumption **(A11)**

$$\mathbb{E}\big[\zeta_n(\theta - \theta_\Sigma) \otimes (\theta - \theta_\Sigma)\zeta_n\big] \preccurlyeq r^2\mathbb{E}\zeta_n\zeta_n \preccurlyeq r^2\kappa(\operatorname{tr}\Sigma)\Sigma.$$

Therefore

$$\mathbb{E}\Big[\big(\nabla f_n(\theta) - \nabla f(\theta)\big) \otimes \big(\nabla f_n(\theta) - \nabla f(\theta)\big)\Big] \preccurlyeq 2\big(\sigma^2 + r^2\kappa(\operatorname{tr}\Sigma)\big)\Sigma.$$

Hence Proposition 2 already implies for all step-size such that $h - \gamma f$ is convex

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{8d}{n}(\sigma^2 + \kappa r^2 \operatorname{tr}\Sigma).$$

D.2 A general result

We prove in this section a more general result than Proposition 3 under the additional assumption

(A14) There exists $b \in [0,1]$ and $\mu_b > 0$ such that $h - \frac{\mu_b}{2}\|\cdot\|_{\Sigma^b}^2$ is convex.

Proposition 5. *Assume **(A2-4)** and **(A7-14)**. Consider the recursion in Eq. (8). For any constant step-size γ such that $\gamma \leq \min\{\frac{\mu_b}{4\kappa \operatorname{tr}\Sigma^{1-b}}, \frac{1}{\kappa L d}\}$. Then*

$$\begin{aligned}\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{24}{n}\operatorname{tr}\Sigma^{-1}C + \frac{16\kappa d\gamma}{n\mu_b}\operatorname{tr} C\Sigma^{-b} \\ + \frac{8\kappa d}{n}\left(\frac{4\kappa\gamma \operatorname{tr}\Sigma^{1-b}}{\mu_b} + 3\right)\|\theta_* - \theta_\Sigma\|_\Sigma^2 + 80\frac{\kappa d}{\gamma n^2}D_h(\theta_*, \theta_0) + \frac{16\kappa d}{n^2}g(\theta_0).\end{aligned}$$

We note Assumption **(14)** is always satisfied for $b = 1$, for which it is Assumption **(13)**. Therefore Proposition 5 directly implies Proposition 3 as a corollary. We prove now two auxiliary lemmas which will be used in the proof of the Proposition 5.

D.3 Two auxiliary results for least-squares objectives

For $b \in [0,1]$, we denote by T_b the operator $T_b = \mathbb{E}[\langle x, \Sigma^{-b}x\rangle x \otimes x]$. We first prove that, for least-square objectives, the sum of the function evaluated along the primal iterates remains bounded.

Lemma 5. *Let us consider the recursion $\eta_n = \eta_{n-1} - \gamma x_n \otimes x_n(\theta_{n-1} - \theta_*) + \gamma\xi_n$ and assume g is positive and there exist μ_b such that $h - \frac{\mu_b}{2}\|\cdot\|_{\Sigma^b}^2$ is convex and κ such that $T_b \preccurlyeq \kappa \operatorname{tr}(\Sigma^{1-b})\Sigma$, then for $\gamma \leq \mu_b/(4\kappa \operatorname{tr}\Sigma^{1-b})$ and $\theta \in \mathcal{X}$ we have*

$$\begin{aligned}&\mathbb{E}\sum_{i=0}^{n}[\psi(\theta_i) - \psi(\theta)] + \Big(1 - 4\gamma\kappa \operatorname{tr}(\Sigma^{1-b})/\mu_b\Big)\sum_{i=0}^{n}\frac{1}{2}\mathbb{E}\|\theta_i - \theta\|_\Sigma^2 \\ &\leq \frac{D_h(\theta, \theta_0) - \mathbb{E}D_h(\theta, \theta_{n+1})}{\gamma} + (n+1)\gamma/\mu_b \operatorname{tr}\Sigma^{-b}C + 4(n+1)\kappa \operatorname{tr}(\Sigma^{1-b})/\mu_b f(\theta) + g(\theta_0).\end{aligned}$$

We note that we can also obtain a bound depending on $2\psi(\theta)$ rather than $4f(\theta)$ with a similar proof.

Proof. Let denote by $f_n(\theta) = x_n \otimes x_n(\theta - \theta_\Sigma) + \xi_n$. Then following the proof of Proposition 1 (see Eq. (14)) we have the expansion

$$\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) \leq -\gamma(g(\theta_n) - g(\theta)) - \gamma\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta\rangle \\ - D_h(\theta_n, \theta_{n-1}) + \gamma\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta_n\rangle. \quad (19)$$

Since $h - \frac{\mu_b}{2}\|\cdot\|^2_{\Sigma^b}$ is convex, using Proposition 4, we get that $D_h(\theta_n, \theta_{n-1}) \geq \frac{\mu_b}{2}\|\theta_n - \theta_{n-1}\|^2_{\Sigma^b}$. Let denote by $A = -D_h(\theta_n, \theta_{n-1}), +\gamma\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta_n\rangle$,

$$\begin{aligned} A &\leq -\frac{\mu_b}{2}\|\theta_n - \theta_{n-1}\|^2_{\Sigma^b} + \gamma\langle x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n, \theta_{n-1} - \theta_n\rangle \\ &\leq -\frac{\mu_b}{2}\|\theta_n - \theta_{n-1}\|^2_{\Sigma^b} \\ &\quad + \Big\langle \frac{\gamma\Sigma^{-b/2}}{\sqrt{\mu_b}}[x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n], \Sigma^{b/2}\sqrt{\mu_b}\theta_{n-1} - \theta_n \Big\rangle \\ &\leq \frac{\gamma^2\mu_b}{2}\|x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n\|^2_{\Sigma^{-b}} \\ &\quad -\tfrac{1}{2}\|\gamma\Sigma^{b/2}\sqrt{\mu_b}(\theta_n - \theta_{n-1}) - \tfrac{\gamma\Sigma^{-b/2}}{\sqrt{\mu_b}}[x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n]\|^2_2 \\ &\leq \frac{\gamma^2}{2\mu_b}\|x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n\|^2_{\Sigma^{-1}} \\ &\leq \frac{\gamma^2}{\mu_b}\|\theta_{n-1} - \theta_\Sigma\|^2_{T_b} + \frac{\gamma^2}{\mu_b}\|\xi_n\|^2_{\Sigma^{-b}}. \end{aligned}$$

Thus, taking the conditional expectation and assuming that κ is such that $T_b \preccurlyeq \kappa \operatorname{tr}(\Sigma^{1-b})\Sigma$ we obtain

$$-\mathbb{E}[D_h(\theta_n, \theta_{n-1})\mathcal{F}_{n-1}] + \gamma\mathbb{E}\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta_n\rangle\mathcal{F}_{n-1}] \\ \leq \frac{\gamma^2\kappa \operatorname{tr}(\Sigma^{1-b})}{\mu_b}\|\theta_{n-1} - \theta_\Sigma\|^2_\Sigma + \frac{\gamma^2}{\mu_b}\operatorname{tr}\Sigma^{-b}C.$$

Taking again the conditional expectation in Eq. (19), we have for $\theta \in \mathcal{X}$

$$\begin{aligned} \mathbb{E}[\tilde{D}_n(\theta, \eta_n)|\mathcal{F}_{n-1}] - \tilde{D}_{n-1}(\theta, \eta_{n-1}) &\leq \frac{\gamma^2\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\|\theta_{n-1} - \theta_\Sigma\|^2_\Sigma + \frac{\gamma^2}{\mu_b}\operatorname{tr}\Sigma^{-b}C \\ &\quad -\gamma\mathbb{E}[\langle x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n, \theta_{n-1} - \theta\rangle|\mathcal{F}_{n-1}] \\ &\quad -\gamma(\mathbb{E}[g(\theta_n)|\mathcal{F}_{n-1}] - g(\theta)) \\ &\leq \frac{\gamma^2\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\|\theta_{n-1} - \theta_\Sigma\|^2_\Sigma - \gamma\langle \theta_{n-1} - \theta_\Sigma, \Sigma(\theta_{n-1} - \theta)\rangle \\ &\quad +\frac{\gamma^2}{\mu_b}\operatorname{tr}\Sigma^{-b}C - \gamma(\mathbb{E}[g(\theta_n)|\mathcal{F}_{n-1}] - g(\theta)). \end{aligned}$$

And we note that

$$-\gamma\langle \theta_{n-1} - \theta_\Sigma, \Sigma(\theta_{n-1} - \theta_*)\rangle = -\gamma[f(\theta_{n-1}) - f(\theta_*)] - \frac{\gamma}{2}\|\theta_{n-1} - \theta\|^2_\Sigma.$$

Therefore

$$\begin{aligned}\mathbb{E}[\tilde{D}_n(\theta_*,\eta_n)|\mathcal{F}_{n-1}] - \tilde{D}_{n-1}(\theta_*,\eta_{n-1}) &\leq -\gamma[f(\theta_{n-1}) - f(\theta_*) + \mathbb{E}[g(\theta_n)|\mathcal{F}_{n-1}] - g(\theta_*)] \\ &\quad -\frac{\gamma}{2}\Big(1 - 4\frac{\gamma\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\Big)\|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 \\ &\quad +2\frac{\gamma^2\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\|\theta - \theta_\Sigma\|_\Sigma^2 + \frac{\gamma^2}{\mu_b}\operatorname{tr}\Sigma^{-b}C.\end{aligned}$$

Taking the total expectation we obtain

$$\begin{aligned}&\mathbb{E}f(\theta_{n-1}) - f(\theta_*) + \mathbb{E}g(\theta_n) - g(\theta_*)] + \frac{1}{2}\Big(1 - 4\frac{\gamma\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\Big)\|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 \\ &\qquad \leq \frac{\mathbb{E}\tilde{D}_{n-1}(\theta_*,\eta_{n-1}) - \mathbb{E}\tilde{D}_n(\theta_*,\eta_n)}{\gamma} + 2\frac{\gamma\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\|\theta - \theta_\Sigma\|_\Sigma^2 + \frac{\gamma}{\mu_b}\operatorname{tr}\Sigma^{-b}C,\end{aligned}$$

which, summing from $i = 0$ to $i = n$, leads to

$$\begin{aligned}&\sum_{i=0}^n[\mathbb{E}f(\theta_i) - f(\theta_*) + \mathbb{E}g(\theta_i) - g(\theta_*)]] + \Big(1 - 4\frac{\gamma\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})\Big)\sum_{i=0}^n\frac{1}{2}\|\theta_i - \theta_\Sigma\|_\Sigma^2 \leq \\ &\qquad \frac{D_h(\theta_*,\theta_0) - \mathbb{E}\tilde{D}_{n+1}(\theta_*,\eta_{n+1})}{\gamma} + 4\frac{\gamma\kappa}{\mu_b}\operatorname{tr}(\Sigma^{1-b})(n+1)\|\theta - \theta_\Sigma\|_\Sigma^2 \\ &\qquad\qquad + (n+1)\frac{\gamma}{\mu_b}\operatorname{tr}\Sigma^{-b}C - \mathbb{E}g(\theta_{n+1}) + g(\theta_0).\end{aligned}$$

The result follows if g is non negative. □

We now present an extension of Lemma 3 to least-squares objectives.

Lemma 6. *Let us consider two sequences of iterates (μ_k, α_k) and (ν_k, β_k) which satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h_n^*(\mu_n)$ and $\beta_n = \nabla h_n^*(\nu_n)$ and denote by $C = \mathbb{E}[x_n \otimes x_n]$ for $n \geq 0$. Assume that γ is such that $h - \gamma T$ is convex. Then*

$$\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2 \leq \mathbb{E}\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2 + 2\gamma^2 n\operatorname{tr}\Sigma^{-1}C.$$

We note that the condition $h - \gamma T$ is rather restrictive since bounds on T are often of the form d times a matrix. For instance Eq. (17) directly implies $T \preccurlyeq \kappa d\Sigma$. Even for independent normal data x_n with diagonal covariance matrix Σ we are able to derive the equality $T = (d+2)\Sigma$.

Proof. We expand

$$\begin{aligned}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2 &= \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + \gamma^2\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2 \\ &\quad -2\gamma\langle x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n, \Sigma^{-1}(\mu_{n-1} - \nu_{n-1})\rangle.\end{aligned}$$

Taking conditional expectations, we get

$$\begin{aligned}\mathbb{E}[\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] &= \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + \gamma^2\mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] \\ &\quad -2\gamma\mathbb{E}[\langle x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n, \Sigma^{-1}(\mu_{n-1} - \nu_{n-1})\rangle|\mathcal{F}_{n-1}] \\ &= \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + \gamma^2\mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] \\ &\quad -2\gamma\langle\alpha_{n-1} - \beta_{n-1}, \mu_{n-1} - \nu_{n-1}\rangle.\end{aligned}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ and denoting by $B = \mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|^2_{\Sigma^{-1}}|\mathcal{F}_{n-1}]$, this leads to

$$\begin{aligned} B &\leq 2\mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1})\|^2_{\Sigma^{-1}}|\mathcal{F}_{n-1}] + 2\mathbb{E}[\|\xi_n\|^2_{\Sigma^{-1}}|\mathcal{F}_{n-1}] \\ &\leq 2\|\alpha_{n-1} - \beta_{n-1}\|^2_{\mathbb{E}[x_n \otimes x_n \Sigma^{-1} x_n \otimes x_n|\mathcal{F}_{n-1}]} + 2\operatorname{tr}\Sigma^{-1}\mathbb{E}[\epsilon_n \otimes \epsilon_n|\mathcal{F}_{n-1}] \\ &\leq 2\|\alpha_{n-1} - \beta_{n-1}\|^2_T + 2\operatorname{tr}\Sigma^{-1}C, \end{aligned}$$

with $T = \mathbb{E}[x \otimes x\Sigma^{-1}x \otimes x]$. Thus we obtain

$$\begin{aligned} \mathbb{E}[\|\mu_n - \nu_n\|^2_{\Sigma^{-1}}|\mathcal{F}_{n-1}] &\leq \|\mu_{n-1} - \nu_{n-1}\|^2_{\Sigma^{-1}} + 2\gamma^2\operatorname{tr}\Sigma^{-1}C \\ &\quad -2\gamma\langle \mu_{n-1} - \nu_{n-1} - \gamma T(\alpha_{n-1} - \beta_{n-1}), \alpha_{n-1} - \beta_{n-1}\rangle \\ &\leq \|\mu_{n-1} - \nu_{n-1}\|^2_{\Sigma^{-1}} + 2\gamma^2\operatorname{tr}\Sigma^{-1}C, \end{aligned}$$

assuming that γ is such $h - \gamma\frac{1}{2}\|\cdot\|^2_T$ is convex (as in the proof of Lemma 3). Taking global expectations, we have shown that

$$\mathbb{E}\|\mu_n - \nu_n\|^2_{\Sigma^{-1}} \leq \mathbb{E}\|\mu_{n-1} - \nu_{n-1}\|^2_{\Sigma^{-1}} + 2\gamma^2\operatorname{tr}\Sigma^{-1}C.$$

□

D.4 Bound on the difference between two averages of primal variables

We present now the following lemma with is an analogue of Lemma 4 for the least-squares problem. It shows that the difference between the average of two sequences of primal iterates which follow the same recursion is $O(1/n)$.

Lemma 7. *Let us consider two sequences of iterates (μ_k, α_k) and (ν_k, β_k) which satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h^*_n(\mu_n)$ and $\beta_n = \nabla h^*_n(\nu_n)$. For $n \geq 0$ denote by $C = \mathbb{E}[x_n \otimes x_n]$. Assume that γ is such that $h - \gamma T$ is convex and there exists κ such that $T \preccurlyeq \kappa d\Sigma$. Then*

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|^2_\Sigma \leq 4\frac{\kappa d - 1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|\alpha_i - \beta_i\|^2_\Sigma + 4\frac{\|\eta_0 - \mu_0\|^2_{\Sigma^{-1}}}{(\gamma n)^2} + \frac{8}{n}\operatorname{tr}\Sigma^{-1}C.$$

Proof. Using the expansion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, we derive

$$\begin{aligned} \Sigma(\alpha_n - \beta_n) &= (\Sigma - x_{n+1} \otimes x_{n+1})(\alpha_n - \beta_n) + x_{n+1} \otimes x_{n+1}(\alpha_n - \beta_n) \\ &= (\Sigma - x_{n+1} \otimes x_{n+1})(\alpha_n - \beta_n) + \frac{\mu_n - \nu_n - \mu_{n+1} + \nu_{n+1}}{\gamma} + \xi_{n+1}. \end{aligned}$$

We obtain by summing n times

$$\Sigma^{1/2}\sum_{i=0}^{n-1}(\alpha_i - \beta_i) = \sum_{i=0}^{n-1}\Sigma^{-1/2}X_{i+1} + \frac{\Sigma^{-1/2}(\mu_0 - \nu_0 - \mu_n + \nu_n)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1},$$

where we denote by $X_i = (\Sigma - x_i \otimes x_i)(\alpha_{i-1} - \beta_{i-1})$ which is a square-integrable martingale difference sequence. We use $(a+b)^2 \leq 2(a^2+b^2)$ to obtain

$$\|(\bar{\alpha}_n - \bar{\beta}_n)\|_\Sigma^2 \leq \frac{2}{n^2}\Big\|\sum_{i=0}^{n-1}\Sigma^{-1/2}X_{i+1} + \frac{\Sigma^{-1/2}(\mu_0-\nu_0)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1}\Big\|_2^2 + 2\frac{\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2}.$$

Therefore using martingale square moment inequalities which here amount to considering the variance of the sum as the sum of the variance, we have

$$\mathbb{E}\|(\bar{\alpha}_n - \bar{\beta}_n)\|_\Sigma^2 \leq \frac{4}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|X_{i+1}\|_{\Sigma^{-1}}^2 + 2\frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + 2\frac{\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{4}{n^2}\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}). \quad (20)$$

- The variance term may be bounded as

$$\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}) \leq \sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}C \leq n\operatorname{tr}\Sigma^{-1}C.$$

- Following Lemma 6 we bound the dual iterates $\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2$ as

$$\frac{\mathbb{E}\|\mu_n-\nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \leq \frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{2}{n}\operatorname{tr}\Sigma^{-1}C.$$

- The martingale difference sequence (X_i) satisfies

$$\begin{aligned}
\mathbb{E}\|\Sigma^{-1/2}X_{i+1}\|_2^2 &\leq \mathbb{E}\langle(\Sigma - x_{i+1}\otimes x_{i+1})(\alpha_i-\beta_i), \Sigma^{-1}(\Sigma - x_{i+1}\otimes x_{i+1})(\alpha_i-\beta_i)\rangle \\
&\leq \langle\alpha_i-\beta_i, \mathbb{E}[(\Sigma - x_{i+1}\otimes x_{i+1})^\top\Sigma^{-1}(\Sigma - x_{i+1}\otimes x_{i+1})](\alpha_i-\beta_i)\rangle \\
&\leq \langle\alpha_i-\beta_i, [\mathbb{E}(x_{i+1}\otimes x_{i+1})^\top\Sigma^{-1}x_{i+1}\otimes x_{i+1} - \Sigma](\alpha_i-\beta_i)\rangle \\
&\leq \langle\alpha_i-\beta_i, [T-\Sigma](\alpha_i-\beta_i)\rangle \\
&\leq (\kappa d - 1)\|\alpha_i-\beta_i\|_\Sigma^2.
\end{aligned}$$

Consequently we obtain in Eq. (20)

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\alpha}_n - \bar{\beta}_n)\|_2^2 \leq 4\frac{\kappa d - 1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|\alpha_i-\beta_i\|_\Sigma^2 + 4\frac{\|\mu_0-\nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{8}{n}\operatorname{tr}\Sigma^{-1}C.$$

□

D.5 Application of Lemma 7 to the proof of Proposition 5

We are now able to prove Proposition 5 using Lemma 7.

Firstly we can directly apply Lemma 7 to $(\mu_n = \eta_n, \alpha_n = \theta_n)$ and $(\nu_n = \eta_n^*, \beta_n = \theta_*)$ where (η_n^*, θ_*) are defined in Eq. (16). This implies

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|_2^2 \leq 4\frac{\kappa d - 1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i - \theta_*\|_\Sigma^2 + 4\frac{\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{8}{n}\operatorname{tr}\Sigma^{-1}C.$$

Following Lemma 5, the primal variables (θ_i) satisfy

$$\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma} + 2\frac{n\gamma}{\mu_b}\operatorname{tr}\Sigma^{-b}C + \frac{8n\gamma\kappa\operatorname{tr}\Sigma^{1-b}}{\mu_b}f(\theta_*) + 2g(\theta_0).$$

This leads to the final bound

$$\begin{aligned}\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|_2^2 \leq\ & 8\frac{\kappa d - 1}{\gamma n^2}D_h(\theta_*, \theta_0) + 4\frac{\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \\ & + 8\frac{1}{n}\operatorname{tr}\Sigma^{-1}C + 8\frac{\kappa d - 1}{n}\frac{\gamma}{\mu_b}[\operatorname{tr}\Sigma^{-b}C + 4\kappa\operatorname{tr}\Sigma^{1-b}f(\theta_*)] + 8\frac{\kappa d - 1}{n^2}g(\theta_0). \quad (21)\end{aligned}$$

This bound depends on $\|\cdot\|_{\Sigma^{-1}}$ which may be infinite. For this reason we compare again the noisy iterate θ_n to the noiseless iterate we still denote by (ϕ_n). We remind these iterates verify the recursion

$$\omega_n = \omega_{n-1} - \gamma\Sigma(\phi_{n-1} - \theta_\Sigma).$$

Therefore the difference $(\eta_n - \omega_n)$ satisfies the same form of recursion as (η_n):

$$\eta_n - \omega_n = \nabla\eta_{n-1} - \omega_{n-1} - \gamma x_n \otimes x_n(\theta_{n-1} - \phi_{n-1}) + \gamma\epsilon_n,$$

with a different noise $\epsilon_n = \xi_n - [x_n \otimes x_n - \Sigma](\phi_{n-1} - \theta_\Sigma)$ and 0 for initial value. Although the noise ϵ_n is different from ξ_n, its covariance is still bounded by

$$\begin{aligned}\frac{1}{3}\mathbb{E}[\epsilon_n \otimes \epsilon_n] \preccurlyeq\ & \mathbb{E}[\xi_n \otimes \xi_n] + \mathbb{E}[[x_n \otimes x_n - \Sigma](\phi_{n-1} - \theta_*) \otimes (\phi_{n-1} - \theta_*)[x_n \otimes x_n - \Sigma]] \\ & + \mathbb{E}[[x_n \otimes x_n - \Sigma](\theta_* - \theta_\Sigma) \otimes (\theta_* - \theta_\Sigma)[x_n \otimes x_n - \Sigma]] \\ \preccurlyeq\ & \mathbb{E}[\xi_n \otimes \xi_n] - \mathbb{E}[\Sigma(\phi_{n-1} - \theta_*)^{\otimes 2}\Sigma] - \mathbb{E}[\Sigma(\theta_* - \theta_\Sigma)^{\otimes 2}\Sigma] \\ & + \mathbb{E}[x_n \otimes x_n(\phi_{n-1} - \theta_*)^{\otimes 2}x_n \otimes x_n] + \mathbb{E}[x_n \otimes x_n(\theta_* - \theta_\Sigma)^{\otimes 2}x_n \otimes x_n] \\ \preccurlyeq\ & \mathbb{E}[\xi_n \otimes \xi_n] + (\kappa - 1)\big(\|\phi_{n-1} - \theta_*\|_\Sigma^2 + \|\theta_* - \theta_\Sigma\|_\Sigma^2\big)\Sigma,\end{aligned}$$

where we have use that for $z \in \mathbb{R}^d$, $\mathbb{E}\langle z, x_n\rangle^4 \leq \kappa\langle z, \Sigma z\rangle$. We may apply Proposition 1 and obtain

$$\mathbb{E}[\epsilon_n \otimes \epsilon_n] \preccurlyeq 3C + \frac{6(\kappa - 1)}{\gamma n}D_h(\theta_*, \theta_0)\Sigma + 6(\kappa - 1)f(\theta_*).$$

Thereby Lemma 7 can be applied with $\theta_0 = \alpha_0$ and we get

$$\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 \leq 4\frac{\kappa d - 1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i - \phi_i\|_\Sigma^2 + \frac{8}{n}\operatorname{tr}\Sigma^{-1}\mathbb{E}[\epsilon_n \otimes \epsilon_n].$$

As before we apply Lemma 5 to have

$$\begin{aligned}\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i-\phi_i\|_\Sigma^2 &\leq \left[2\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i-\theta_*\|_\Sigma^2+2\sum_{i=0}^{n-1}\|\phi_i-\theta_*\|_\Sigma^2\right]\\ &\leq \left[\frac{8D_h(\theta_*,\theta_0)}{\gamma}+\frac{n\gamma}{\mu_b}4\operatorname{tr}\Sigma^{-b}C+\frac{16n\gamma\kappa\operatorname{tr}\Sigma^{1-b}}{\mu_b}f(\theta_*)+4g(\theta_0)\right].\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\|\bar\theta_n-\bar\phi_n\|_\Sigma^2 \leq 4\frac{\kappa d-1}{n^2}&\left[8\frac{D_h(\theta_*,\theta_0)}{\gamma}+\frac{n\gamma}{\mu_b}4\operatorname{tr}\Sigma^{-b}C+\frac{16n\gamma\kappa\operatorname{tr}\Sigma^{1-b}}{\mu_b}f(\theta_*)+4g(\theta_0)\right]\\ &+\frac{8}{n}\left[3\operatorname{tr}\Sigma^{-1}C+\frac{6(\kappa-1)}{\gamma n}D_h(\theta_*,\theta_0)d+6(\kappa-1)f(\theta_*)d\right].\end{aligned}$$

And rearranging terms we obtain

$$\begin{aligned}\mathbb{E}\|\bar\theta_n-\bar\phi_n\|_\Sigma^2 \leq 80\frac{\kappa d}{\gamma n^2}D_h(\theta_*,\theta_0)+\frac{64\kappa^2 d\gamma}{n\mu_b}\operatorname{tr}\Sigma^{1-b}f(\theta_*)+\frac{16\kappa d\gamma}{n\mu_b}\operatorname{tr}C\Sigma^{-b}\\ +\frac{16\kappa d}{n^2}g(\theta_0)+\frac{24}{n}\operatorname{tr}\Sigma^{-1}C+\frac{48\kappa d}{n}f(\theta_*).\end{aligned}$$

And by the Cauchy-Schwarz inequality ($\mathbb{E}\|\bar\theta_n-\theta_*\|_\Sigma^2\leq 2\mathbb{E}\|\bar\theta_n-\bar\phi_n\|_\Sigma^2+2\mathbb{E}\|\bar\phi_n-\theta_*\|_\Sigma^2$)

$$\begin{aligned}\frac{1}{2}\mathbb{E}\|\bar\theta_n-\theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*,\theta_0)}{\gamma n}+\frac{24}{n}\operatorname{tr}\Sigma^{-1}C+\frac{16\kappa d\gamma}{n\mu_b}\operatorname{tr}C\Sigma^{-b}+\frac{16\kappa d}{n}\Big(\frac{4\kappa\gamma\operatorname{tr}\Sigma^{1-b}}{\mu_b}+3\Big)f(\theta_*)\\ +80\frac{\kappa d}{\gamma n^2}D_h(\theta_*,\theta_0)+\frac{16\kappa d}{n^2}g(\theta_0),\end{aligned}$$

which proves the second bound of Proposition 5.

D.6 A corollary of Proposition 5 for h with an Euclidean behavior

When h rather behaves as an Euclidean norm, we may replace Assumptions **(A12-13)** by the following:

(A12') There exists $\mu_h>0$ such that $h-\frac{\mu_h}{2}\|\cdot\|_2^2$ is convex.

(A13 ') There exists R^2 such that $\mathbb{E}[\|x_n\|_2^2 x_n\otimes x_n]\preccurlyeq R^2\Sigma$.

And Proposition 5 implies the following corollary.

Corollary 1. *Assume For any constant step-size γ such that $\gamma\leq\min\{\frac{\mu_h}{4\kappa R^2},\frac{R^2}{4\kappa d}\}$. Then*

$$\frac{1}{2}\mathbb{E}\|\bar\theta_n-\theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*,\theta_0)}{\gamma n}+\frac{8}{n}\Big(3+\frac{4\gamma\kappa R^2}{\mu_h}\Big)\Big(\sigma^2 d+\kappa d\|\theta_*-\theta_\Sigma\|_\Sigma^2\Big)+\frac{16\kappa d}{n^2}\Big(\frac{5D_h(\theta_*,\theta_0)}{\gamma}+g(\theta_0)\Big).$$

This corollary would pave the way for a general result for larger step-size γ without the condition $\gamma\leq\frac{R^2}{4\kappa d}$. Unfortunately the latter seems not improvable, as noted after Lemma 6.

E Lower bound for non-strongly convex quadratic regularization

We derive, in this section, a lower bound on the performance of SDA when f is the linear form $f(\theta) = \langle a, \theta \rangle$ with $a \in \mathbb{R}^d$ and g is a non-strongly convex quadratic function. We assume that the vector a is not available and we only have access to estimates of the gradient

$$\nabla f_n(\theta) = a + \xi_n \text{ for } n \geq 1, \tag{22}$$

where (ξ_n) is an uncorrelated zero-mean noise sequence with bounded covariance.

Proposition 6. *For any $d \geq 2$, $L > 0$; $\gamma > 0$ and finite time horizon $N \geq 1$, there exists a quadratic function g L-smooth such that for any uncorrelated zero-mean noise sequence (ξ_n) with bounded covariance $\mathbb{E}[\xi_n \otimes \xi_n] = \sigma^2 L I_d$, SDA with constant step-size γ applied with the oracle Eq. (22) satisfies*

$$\psi(\bar{\theta}_N) - \psi(\theta_*) \geq \frac{\sigma^2}{12} \min\{(L\gamma)^2, 1\}.$$

Proof. For sake of clarity, we consider $d = 2$ and $a = 0$. Thus $f(\theta) = \mathbb{E}\langle \xi_n, \theta \rangle = 0$. Let $g(\theta) = \frac{1}{2}\langle \theta, A\theta \rangle$ be a quadratic form with $A = \begin{pmatrix} L & 0 \\ 0 & \mu \end{pmatrix}$ for $L \geq \mu > 0$ with μ possibly arbitrary small. The noise (ξ_n) is assumed to be uncorrelated zero-mean with bounded covariance $\mathbb{E}[\xi_n \otimes \xi_n] = \sigma^2 L I_2$. The stochastic dual algorithm with step-size γ takes the form:

$$\begin{aligned}
\theta_n &= \nabla h_n^*(-n\gamma \bar{\xi}_n) \\
&= \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \Big\{ \langle \bar{\xi}_n, \theta \rangle + \frac{1}{2}\langle \theta, A\theta \rangle + \frac{1}{2n\gamma} \|\theta\|_2^2 \Big\} \\
&= \gamma n (I + \gamma n A)^{-1} \bar{\xi}_n.
\end{aligned}$$

And

$$\begin{aligned}
\bar{\theta}_n &= \frac{\gamma}{n} \sum_{k=1}^{n-1} \sum_{j=1}^{k} k (I + \gamma k A)^{-1} \frac{1}{k} \xi_j \\
&= \frac{\gamma}{n} \sum_{j=1}^{n-1} \Big(\sum_{k=j}^{n-1} (I + \gamma k A)^{-1} \Big) \xi_j.
\end{aligned}$$

Therefore using standard martingale square moment inequalities

$$\begin{aligned}
\mathbb{E}\langle \bar{\theta}_n, A\bar{\theta}_n \rangle &= \frac{\gamma^2}{n^2} \sum_{j=1}^{n} \mathbb{E}\Big\langle \xi_j \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big), A \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big) \xi_j \Big\rangle \\
&= \frac{\gamma^2 \sigma^2 L}{n^2} \operatorname{tr} \sum_{j=1}^{n} \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big) A \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big) I_2 \\
&= \frac{\gamma^2 \sigma^2 L}{n^2} \sum_{j=1}^{n} \Big[L \Big(\sum_{k=j}^{n} \frac{1}{1 + \gamma L k} \Big)^2 + \mu \Big(\sum_{k=j}^{n} \frac{1}{1 + \gamma \mu k} \Big)^2 \Big].
\end{aligned}$$

And

$$\begin{aligned}\mathbb{E}\langle\bar{\theta}_n, A\bar{\theta}_n\rangle &\geq \frac{\gamma^2\sigma^2 L}{n^2}\Big[\frac{L}{(1+\gamma L n)^2}+\frac{\mu}{(1+\gamma\mu n)^2}\Big]\sum_{j=1}^{n}(n-j)^2 \\ &\geq \frac{n\sigma^2\gamma^2 L}{3}\Big[\frac{L}{(1+\gamma L n)^2}+\frac{\mu}{(1+\gamma\mu n)^2}\Big] \geq \frac{n\sigma^2\gamma^2}{3}\frac{\mu}{(1+\gamma\mu n)^2} \\ &\geq \frac{\sigma^2 L}{12}\min\Big(n\mu\gamma^2, \frac{1}{\mu n}\Big).\end{aligned}$$

Conclude by taking $\mu = L/N$.

The proof is the same for $d \geq 2$ by considering $A = \mathrm{diag}(L, \ldots, L, L\mu)$ with $d-1$ L. □

F Lower bound for stochastic approximation problems

In this section we relate the problem of aggregation of estimators to the stochastic convex optimization problem, i.e., minimizing a convex function, given only unbiased estimates of its gradients. We will consider the regression and the classification with hinge loss problems which will individually provide lower bounds for quadratic and linear functions. We follow here Tsybakov (2003); Lecué (2006); Agarwal et al. (2012).

F.1 Oracle complexity of stochastic convex optimization

Beforehand we describe the stochastic oracle model formalism as done by Nemirovsky and Yudin (1983); Agarwal et al. (2012); Raginsky and Rakhlin (2011). For a given class of problems we aim to determine lower bounds on the number of queries to a stochastic first-order oracle needed to optimize to a certain precision any function in this class. To this end we have the following definition.

Definition 1 (Agarwal et al. (2012))**.** *For a given constraint convex set $\mathcal{C}$, and a function class $\mathcal{S}$, a first-order stochastic oracle is a random mapping $\pi : \mathcal{C} \times \mathcal{S} \to \mathbb{R} \times \mathbb{R}^d$ of the form*

$$\phi(\theta, f) = (\tilde{f}(\theta), g(\theta)),$$

such that

$$\mathbb{E}\tilde{f}(\theta) = f(\theta); \qquad \mathbb{E}g(\theta) = \nabla f(\theta),$$

and there exists a constant $C < \infty$ such that for every $\theta \in \mathbb{R}^d$

$$\mathbb{E}[\|g(\theta) - \nabla f(\theta)\|^2] \leq C(1 + \|\theta\|^2).$$

The class of first-order stochastic oracle is denoted by Φ. A stochastic approximation algorithm M is a method which approximately minimizes a function f by querying, at each iteration i, the oracle at the point θ_i. The oracle answers with the information $\phi(\theta_i, f)$ and the method uses all the information $\{\phi(\theta_0, f), \ldots, \phi(\theta_i, f)\}$ to build a new point θ_{i+1}. For $n \in \mathbb{N}$ we denote by $\mathcal{M}_n$ the

class of all such methods that are allowed to make n queries. As done by Agarwal et al. (2012), we denote the error of the method M on the function f after n steps as

$$\epsilon_n(M, f, \mathcal{C}, \phi) = f(\theta_n) - \min_{\theta \in \mathcal{C}} f(\theta).$$

Given a class of functions $\mathcal{S}$, an oracle ϕ and a convex constraint set $\mathcal{C}$, Agarwal et al. (2012) also defines the minimax error as

$$\epsilon_n^*(\mathcal{S}, \mathcal{C}, \phi) = \inf_{M \in \mathcal{M}_n} \sup_{f \in \mathcal{S}} \mathbb{E}_\phi \epsilon_n(M, f, \mathcal{C}, \phi).$$

We will lower bound this minimax error by relating convex stochastic approximation with convex aggregation of estimators (Juditsky and Nemirovski, 2000; Tsybakov, 2003).

F.2 Convex aggregation of estimators

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and $\mathcal{Y} \subset \mathbb{R}$. We consider random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with probability distribution denoted by π. We observe n i.i.d. pairs $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ which follow the law π and we want to predict the output Y for any feature $X \in \mathcal{X}$ by a prediction $f(X)$ for a measurable function f from $\mathcal{X}$ to $\mathbb{R}$. For this purpose we want to minimize the risk defined by

$$A(f) = \mathbb{E}[\ell(f(X), Y)],$$

for any measurable function f from $\mathcal{X}$ to $\mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function.

We consider we have access to d different arbitrary estimators $\mathcal{F} = \{f_1, \dots, f_d\}$ with values in $\mathcal{Y}$. We denote their convex hull by $\mathcal{C} = conv(f_1, \dots, f_d)$. The aim of convex aggregation is to build a new estimator which is a convex combination of the different f_i and behaves as the best among the estimators f_i. The aggregation problem is equivalent to a minimization problem over the simplex Δ_d since for $f \in \mathcal{C}$ there is $\theta \in \Delta_d$ such that $f = \sum_{i=1}^d \theta(i) f_i$. Therefore, defining $B(\theta) = A\Big(\sum_{i=1}^d \theta(i) f_i\Big)$, we have

$$\min_{f \in \mathcal{C}} A(f) = \min_{\theta \in \Delta_d} B(\theta).$$

We denote by $F : \mathcal{X} \to \mathbb{R}^d, x \mapsto (f_1(x), \dots, f_d(x))$ the function whose the ith coordinate is the function f_i, and we have

$$B(\theta) = \mathbb{E}[\ell(\langle F(X), \theta \rangle, Y)].$$

Therefore the convex aggregation problem of minimizing $A(f)$ over the convex hull of $\mathcal{F}$ is formally equivalent to the stochastic approximation problem of minimizing, over the simplex Δ_d, the function $B(\theta) = \mathbb{E}[\ell(\langle F(X), \theta \rangle, Y)]$, given only unbiased estimates of its gradient $\nabla B_n(\theta) = \nabla \ell(\langle F(x_n), \theta \rangle, y_n)$. Hence lower bounds on convex aggregation problems provide lower bounds on stochastic approximation problems studied in this paper.

F.3 Aggregation in regression and application to oracle complexity of stochastic quadratic optimization

We first consider the regression problem for which $\mathcal{Y} = \mathbb{R}$. We rely substantially on Tsybakov (2003). The regression model is

$$Y_i = f_*(X_i) + \xi_i, \text{ for } i = 1, \dots, n,$$

where $X_1, \dots, X_n$ are i.i.d. random vectors of $\mathcal{X}$ of law P^X and ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables such that $(\xi_1, \dots, \xi_n)$ is independent of $(X_1, \dots, X_n)$ and $f_* : \mathcal{X} \to \mathbb{R}$ is the regression function. Regression problem aims to estimate the unknown regression function f_* based on the data D_n by minimizing the risk

$$A_{\text{reg}}(f, f_*) = \mathbb{E}(f(X) - f_*(X))^2.$$

The problem of the optimal rate of convex aggregation has been studied by Tsybakov (2003). We reintroduce his notations and assumptions for sake of completeness.

Let denote by $\mathcal{F}_0 = \{f : \|f\|_\infty \leq L\}$ for $L > 0$ and assume that

(B1) There exists a cube $S \subset \mathcal{X}$ such that P^X admits a bounded density μ on S w.r.t. the Lebesgue measure and $\mu(x) \geq \mu_0 > 0$ for all $x \in S$.

(B2) There exists a constant c_0 such that $d \leq c_0 \exp(n)$.

We have the following result

Theorem 1 (Theorem 2, Tsybakov (2003))**.** *Under assumptions **(B1-2)** we have*

$$\sup_{f_1, \dots, f_d \in \mathcal{F}_0} \inf_{T_n} \sup_{f_* \in \mathcal{F}_0} \left[\mathbb{E}_{D_n} A_{reg}(T_n, f_*) - \min_{f \in \mathcal{C}} A_{reg}(f, f_*)\right] \geq c\zeta_n(d),$$

for some constant $c > 0$ and any integer n, where $\inf_{T_n}$ denotes the infimum over all estimators, $\mathbb{E}_{D_n}$ denotes the expectation with regard to the probability distribution of the data D_n and

$$\zeta_n(d) = \begin{cases} d/n & \text{if } d \leq \sqrt{n} \\ \sqrt{\frac{1}{n} \log\left(\frac{d}{\sqrt{n}} + 1\right)} & \text{if } d > \sqrt{n}. \end{cases}$$

We relate now the problem of convex aggregation of regression functions to the problem of stochastic quadratic functions optimization. Consider $\mathcal{F} = \{f_1, \dots, f_d\}$ the set of estimators given by Proposition 1 and denote by $F : \mathcal{X} \to \mathbb{R}^d, x \mapsto (f_1(x), \dots, f_d(x))$. For $f \in \mathcal{C}$, there is $\theta \in \Delta_d$ such that $f = \sum_{i=1}^d \theta(i) f_i$ and we obtain

$$A_{\text{reg}}(f, f_*) = \mathbb{E}[(\langle \theta, F(X) \rangle - f_*(X))^2] = B(\theta),$$

where $B(\theta) = \langle \theta, \mathbb{E}[F(X) \otimes F(X)], \theta \rangle - 2\langle \theta \mathbb{E}[f_*(X) F(X)] \rangle + \mathbb{E}[f_*(X)^2]$ is a quadratic function. This set enables us to construct a difficult subclass of quadratic functions:

$$\mathcal{G}_{\text{quad}} = \left\{ B(\theta) = \frac{1}{2} \mathbb{E}[(\langle \theta, F(X) \rangle - f_*(X))^2]; f_* \in \mathcal{F}_0 \right\}.$$

We also define the first-order stochastic oracle ϕ_{quad} on $\mathcal{G}_{\text{quad}}$ as follows

$$\phi_{\text{quad}}(\theta, f) = \left(\frac{1}{2}(\langle\theta, F(x)\rangle - f_*(x))^2, (\langle\theta, F(x)\rangle - f_*(x))F(x)\right), \text{ for } x \sim P^X.$$

We can optimize B with a stochastic approximation algorithm $M \in \mathcal{M}_n$ to obtain $\theta_n \in \Delta_d$ and therefore build a estimator $T_n = \sum_{i=1}^d \theta_n(i) f_i$ which belongs to $\mathcal{C}$. Moreover we have

$$A_{\text{reg}}(T_n, f_*) = B(\theta_n) \text{ and } \min_{f\in\mathcal{C}} A_{\text{reg}}(f, f_*) = \min_{\theta\in\Delta_d} B(\theta).$$

Consequently, for the oracle ϕ_{quad} and the class $\mathcal{G}_{\text{quad}}$ Proposition 1 implies that

$$\epsilon_n^*(\mathcal{G}_{\text{quad}}, \Delta_d, \phi_{\text{quad}}) \geq c\zeta_n(d). \tag{23}$$

And we have proven the following minimax oracle complexity.

Proposition 7. *Let Δ_d be the simplex. Then there exists universal constants $c_0 > 0$ and $c > 0$ such that the minimax oracle complexity over the class $\mathcal{S}_{quad}$ of quadratic functions satisfies the following lower bounds:*

- *For $d \leq \sqrt{n}$*
$$\sup_{\phi\in\Phi} \epsilon_n^*(\mathcal{S}_{quad}, \Delta_d, \phi) \geq c\frac{d}{n}.$$
- *For $\sqrt{n} \leq d \leq c_0 \exp(n)$*
$$\sup_{\phi\in\Phi} \epsilon_n^*(\mathcal{S}_{quad}, \Delta_d, \phi) \geq c\sqrt{\frac{1}{n}\log\left(\frac{d}{\sqrt{n}} + 1\right)}.$$

We note that without assumption on d the lower-bound for the class of quadratic functions is of order $O(1/n)$ but in high-dimensional settings it becomes of order $(1/\sqrt{n})$. Nevertheless we will see in the next section this lower-bound is always of order $(1/\sqrt{n})$ for the class of linear functions.

F.4 Aggregation in classification and application to oracle complexity of stochastic linear optimization

We consider now the classification problem with the hinge loss for which $\mathcal{Y} = \{-1, 1\}$. We follow very closely the framework of Lecué (2006, 2007) and use their notations. We still consider random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with probability distribution denoted by π. We observe n i.i.d. pairs $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ which follow the law π and we want to predict the label Y for any feature $X \in \mathcal{X}$ by minimizing the hinge risk defined by

$$A_{\text{cla}}(f) = \mathbb{E}\max(1 - Yf(X), 0),$$

for any measurable function f from $\mathcal{X}$ to $\mathbb{R}$. We consider we have access to d different estimators $\mathcal{F} = \{f_1, \ldots, f_d\}$ with values in $[-1, 1]$. We denote their convex hull by $\mathcal{C} = conv(f_1, \ldots, f_d)$. Lecué (2006, Theorem 1) and Lecué (2007, Theorem 2) provide a lower bound on this aggregation problem for classification we adapt to our specific case.

Proposition 8 (Adaptation of Theorem 2 of Lecué (2007) for $\kappa = \infty$)**.** *Let d, n be two integers such that $2\log_2 d \leq n$. We assume that the input space $\mathcal{X}$ is infinite. There exists an absolute constant $c > 0$, and a set of prediction rules $\mathcal{F} = \{f_1, \dots, f_n\}$ such that for any real-valued procedure T_n, there exists a probability measure π, for which*

$$\mathbb{E}_{D_n}[A_{cla}(T_n)] - \min_{f \in \mathcal{C}}(A_{cla}(f)) \geq c\sqrt{\frac{\log d}{n}}.$$

Proof. Theorem 2 of Lecué (2007) is stated under an additional Margin assumption MAH(κ) (see definition and notation below Eq. (9) in Lecué (2007)) on the probability distribution π, i.e., there exists a constant c_0 such that

$$\mathbb{E}[|f(X) - f^*(X)|] \leq c_0(A(f) - A^*)^{1/\kappa},$$

for any function f on $\mathcal{X}$ with values in $[-1, 1]$. Therefore taking $\kappa \to \infty$, we can always consider $c_0 = 2$. And the constant $c(\kappa)$ in Theorem 2 of Lecué (2007) is

$$c(\kappa) = c_0^{\kappa}(4e)^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)},$$

which goes when $\kappa \to \infty$ to $c_\infty = \sqrt{2}/(4e\sqrt{\log 2})$. Hence taking $\kappa \to \infty$ in Theorem 2 of Lecué (2007) implies Proposition 8. We could also have plugged arguments of the proof of Theorem 14.5 of Devroye et al. (1996) to directly prove this result. □

We relate now the problem of convex aggregation of classifiers to the problem of optimizing a linear function on the simplex. Consider the set of prediction rules $\mathcal{F} = \{f_1, \dots, f_n\}$ given by Proposition 8 and denote by $F : \mathcal{X} \to \mathbb{R}^d, x \mapsto (f_1(x), \dots, f_d(x))$. For $f \in \mathcal{C}$, there is $\theta \in \Delta_d$ such that $f = \sum_{i=1}^d \theta(i) f_i$ and we obtain

$$A_{\text{cla}}(f) = \mathbb{E}\max(1 - Y\langle F(X), \theta\rangle, 0).$$

On the other hand, when the f_i are valued in $[-1, 1]$, the classification problem becomes equivalent to maximize the expectation $\mathbb{E}Yf(X)$ since the hinge loss is linear on $[-1, 1]$:

$$Y \in \{-1, 1\}, f(X) \in [-1, 1] \implies Yf(X) \in [-1, 1] \implies \mathbb{E}\max(1 - Yf(X), 0) = 1 - \mathbb{E}Yf(X).$$

Combining both, we obtain that

$$A_{\text{cla}}(f) = 1 - \langle \mathbb{E}[YF(X)], \theta\rangle = 1 + C(\theta),$$

where $C(\theta) = -\langle \mathbb{E}[YF(X)], \theta\rangle$ is a linear function. This set enables us to construct a difficult subclass of linear functions

$$\mathcal{G}_{\text{lin}} = \{C(\theta) = -\langle \mathbb{E}[YF(X)], \theta\rangle; (X, Y) \sim \pi\}.$$

We also define the first-order stochastic oracle ϕ_{lin} on $\mathcal{G}_{\text{lin}}$ as follows

$$\phi_{\text{lin}}(\theta, f) = \Big(\langle yF(x), \theta\rangle, yF(x)\Big), \text{ for } (x, y) \sim \pi.$$

As before we may optimize C with a stochastic approximation algorithm $M \in \mathcal{M}_n$ to obtain $\theta_n \in \Delta_d$ and therefore build a estimator $T_n = \sum_{i=1}^d \theta_n(i) f_i$ which belongs to $\mathcal{C}$. Moreover we have

$$A_{\text{cla}}(T_n) = C(\theta_n) \text{ and } \min_{f \in \mathcal{C}} A_{\text{cla}}(f) = \min_{\theta \in \Delta_d} C(\theta).$$

Consequently, for the oracle ϕ_{lin} and the class $\mathcal{G}_{\text{lin}}$ Proposition 1 implies that

$$\epsilon_n^*(\mathcal{G}_{\text{lin}}, \Delta_d, \phi_{\text{lin}}) \geq c\sqrt{\frac{\log d}{n}}. \tag{24}$$

And we have proven the following minimax oracle complexity.

Proposition 9. *Let Δ_d be the simplex. Then there exists universal constant $c > 0$ such that the minimax oracle complexity over the class $\mathcal{S}_{lin}$ of linear functions satisfies the following lower bound for $2\log_2 d \leq n$*

$$\sup_{\phi \in \Phi} \epsilon_n^*(\mathcal{S}_{lin}, \Delta_d, \phi) \geq c\sqrt{\frac{\log(d)}{n}}.$$

G Lower-bound on the rates of convergence of DA and MD algorithms

Let us consider in this section that $f = 0$, $g(\theta) = \frac{1}{2\nu}\|\theta - \theta_*\|_2^2$ and $h = \frac{1}{2}\|\theta\|_2^2$. In this case, for $n \geq 1$, MD iterates (θ_n^{md}) verify

$$\theta_n^{\text{md}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2\nu}\|\theta - \theta_*\|_2^2 + \frac{1}{2\gamma}\|\theta - \theta_{n-1}^{\text{md}}\|_2^2 \right\}.$$

Therefore $\theta_n^{\text{md}} = \theta_* + \frac{1}{\gamma\nu}(\theta_{n-1}^{\text{md}} - \theta_*)$, $\theta_n^{\text{md}} - \theta_* = \frac{1}{(\gamma\nu)^n}(\theta_0^{\text{md}} - \theta_*)$ and

$$g(\theta_n^{\text{md}}) - g(\theta_*) = \frac{g(\theta_0^{\text{md}}) - g(\theta_*)}{(\gamma\nu)^{2n}}.$$

Whereas DA iterates (θ_n^{da}) satisfy

$$\theta_n^{\text{da}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2\nu}\|\theta - \theta_*\|_2^2 + \frac{1}{2\gamma n}\|\theta\|_2^2 \right\}.$$

We compute $\theta_n^{\text{da}} = \frac{\gamma\nu n}{\gamma\nu n + 1}\theta_*$ and

$$g(\theta_n^{\text{da}}) - g(\theta_*) = \frac{g(\theta_0^{\text{da}}) - g(\theta_*)}{(\gamma n)^2}.$$

H Continuous time interpretation of DA et MD

Following Nemirovsky and Yudin (1983); Krichene et al. (2015) we propose a continuous interpretation of these methods for g twice differentiable. We note this could be extended for g non-smooth with differential inclusions.

Derivation of the ordinary differential equation (ODE). The first-order optimality condition of the MD iteration in Eq. (4) $\gamma\nabla f(\theta_n) + \gamma\nabla g(\theta_{n+1}) + \nabla h(\theta_{n+1}) - \nabla h(\theta_n)$ can be rearranged as

$$\frac{\nabla h(\theta_{n+1}) - \nabla h(\theta_n)}{\gamma} = -\nabla f(\theta_n) - \nabla g(\theta_{n+1}).$$

Noting $\partial_t \nabla h(\theta) = \nabla^2 h(\theta)\dot{\theta}$, this is exactly a forward-backward Euler discretization of the MD ODE

$$\dot{\theta} = -\nabla^2 h(\theta)^{-1}[\nabla f(\theta) + \nabla g(\theta)]. \tag{25}$$

On the other hand, considering the DA iteration in Eq. (3) we obtain

$$\frac{\eta_n - \eta_{n-1}}{\gamma} = -\nabla f(\theta_{n-1}) \qquad \text{and} \qquad \eta_n = n\gamma\nabla g(\theta_n) + \nabla h(\theta_n). \tag{26}$$

Combining both parts in Eq. (26) leads to the single equation

$$n\gamma\frac{\nabla g(\theta_n) - \nabla g(\theta_{n-1})}{\gamma} + \nabla g(\theta_{n-1}) + \frac{\nabla h(\theta_n) - \nabla h(\theta_{n-1})}{\gamma} = -\nabla f(\theta_{n-1}),$$

which is the explicit Euler discretization of the ODE $\partial_t(t\nabla g(\theta) + \nabla h(\theta)) = -\nabla f(\theta)$. Therefore the ODE associated to DA takes the form

$$\dot{\theta} = -\nabla^2(h(\theta) + tg(\theta))^{-1}(\nabla f(\theta) + \nabla g(\theta)). \tag{27}$$

It is worth noting that this ODE is very similar to the MD ODE in Eq. (25), with an additional term $tg(\theta)$ in the inverse mapping $\nabla^2(h(\theta) + tg(\theta))^{-1}$ which may thus slow down the DA dynamic.

Lyapunov analyzes. Lyapunov functions are used to prove convergence of the solutions of ODEs. In analogy with the discrete case, the Bregman divergence is a Lyapunov function for these ODEs (see, e.g., Krichene et al., 2015) since

$$\begin{aligned}
\partial_t D_h(\theta_*, \theta(t)) &= \partial_t[h(\theta_*) - h(\theta(t)) - \langle\nabla h(\theta(t)), \theta_* - \theta(t)\rangle] \\
&= -\langle\nabla h(\theta(t)), \dot{\theta}(t)\rangle + \langle\nabla^2 h(\theta(t))\dot{\theta}(t), \theta(t) - \theta_*\rangle + \langle\nabla h(\theta(t)), \dot{\theta}(t)\rangle \\
&= \langle\nabla^2 h(\theta(t))\dot{\theta}(t), \theta(t) - \theta_*\rangle.
\end{aligned}$$

For the MD ODE in Eq. (25) we obtain

$$\begin{aligned}
\partial_t D_h(\theta_*, \theta(t)) &= -\langle\nabla f(\theta(t)) + \nabla g(\theta(t)), \theta(t) - \theta_*\rangle \\
&\leq \psi(\theta_*) - \psi(\theta(t)) \qquad \text{(by convexity of } \psi\text{)}.
\end{aligned}$$

Integrating, this yields with Jensen inequality

$$\psi(\bar{\theta}(t)) - \psi(\theta_*) \leq \frac{1}{t}\int_0^t \big(\psi(\theta(s)) - \psi(\theta_*)\big)ds \leq \frac{D_h(\theta_*, \theta(0)) - D_h(\theta_*, \theta(t))}{t},$$

for $\bar{\theta}(t) = \frac{1}{t}\int_0^t \theta(s)ds$. This is the same convergence result as in the discrete time. For the DA ODE in Eq. (27) we obtain

$$\begin{aligned}\partial_t D_{h+tg}(\theta_*, \theta(t)) &= \partial_t[(h+tg)(\theta_*) - (h+tg)(\theta(t)) - \langle \nabla(h+tg)(\theta(t)), \theta_* - \theta(t)\rangle] \\ &= g(\theta_*) - \langle (\nabla h(\theta(t)) + t\nabla g(\theta(t))), \dot{\theta}(t)\rangle + g(\theta(t)) \\ &\quad + \langle \partial_t(\nabla h + t\nabla g)(\theta(t)), \theta(t) - \theta_*\rangle + \langle (\nabla + t\nabla g)h(\theta(t)), \dot{\theta}(t)\rangle \\ &= g(\theta_*) - g(\theta_t) - \langle \nabla f(\theta(t)), \theta(t) - \theta_*\rangle.\end{aligned}$$

Therefore by convexity of f, $\partial_t D_{h+tg}(\theta_*, \theta(t)) \leq \psi(\theta_*) - \psi(\theta(t))$ and we obtain

$$\psi(\bar{\theta}(t)) - \psi(\theta_*) \leq \frac{D_h(\theta_*, \theta(0)) - D_{h+tg}(\theta_*, \theta(t))}{t}.$$

The continuous time argument really mimics the proof of Proposition 1 without the technicalities associated with the discrete time. We remind that we recover the variational interpretation of Krichene et al. (2015); Wibisono et al. (2016); Wilson et al. (2016): the Lyapunov function generates the dynamic in the sense that a function L is first chosen and secondly a dynamics, for which L is a Lyapunov function, is then designed. In this way MD and DA are the two different dynamics associated to the two different Lyapunov functions D_h and D_{h+tg}.

Extension to the noisy-gradient case. We consider now we only have access to noisy estimates of the gradient as in Section 3 and propose a continuous-time interpretation of these stochastic methods. Stochastic MD and SDA may be viewed, in their primal-dual forms, as discretizations of the following stochastic differential equations (SDE). For stochastic MD

$$\mathrm{d}\eta(t) = -[\nabla f(\theta(t)) + \nabla g(\theta(t))]\mathrm{d}t + \sigma \mathrm{d}W(t)\mathrm{d}t \qquad \text{and} \qquad \eta(t) = \nabla h(\theta(t)),$$

and for SDA

$$\mathrm{d}\eta(t) = -\nabla f(\theta(t))\mathrm{d}t + \sigma \mathrm{d}W(t)\mathrm{d}t \qquad \text{and} \qquad \eta(t) = \nabla(h+tg)(\theta(t)),$$

where W_t is a Wiener process and $\sigma > 0$. We note that the regularization g does not take part in the SDA SDE which explains this dynamic is efficient in presence of noise. In contrast, the stochastic MD SDE is corrupted by the presence of the gradient ∇g which may not behaves well for non-smooth g. This continuous-time interpretation of stochastic algorithms could lead to further insights but is outside the scope of this paper.

I Examples of different geometries

We describe now different examples of concrete geometries and how SDA is then implemented for well known regularizations g.

Euclidean distance. The simplest geometry is obtained by taking the function $h(\theta) = \frac{1}{2}\|\theta\|_2^2$, which is a Legendre function on $\mathrm{dom}\, h = \mathbb{R}^d$. Its associated Bregman divergence is also the squared Euclidean distance $D_h(\alpha, \beta) = \frac{1}{2}\|\alpha - \beta\|_2^2$. Therefore **(LC)** is equivalent to the smoothness of the function f and we return to classic results on proximal gradient descent.

- Projection: Let $g = 1_{\mathcal{C}}$ be the indicator of a convex set $\mathcal{C}$. The SDA method yields to the projected method
$$\theta_n = \min_{\theta \in \mathcal{C}} \Big\| \theta + \gamma \sum_{k=0}^{n-1} \nabla f_{k+1}(\theta_k) \Big\|_2^2.$$

- ℓ_2-regularization: Let $g = \frac{1}{2}\|\cdot\|_Q^2$ where $Q \succcurlyeq 0$, we directly have $\nabla h_n^*(\eta) = (I + n\gamma Q)^{-1}\eta$ and the SDA method comes back to
$$\theta_n = \theta_{n-1} - (\gamma^{-1} I + nQ)^{-1}(Q\theta_{n-1} + \nabla f_n(\theta_{n-1})), \text{ for } n \geq 1,$$
which is a standard gradient descent on $f + g$ with a structured decreasing step-size $\gamma_n = (\gamma^{-1} I + nQ)^{-1}$.

- ℓ_1-regularization: Let $g = \lambda\|\cdot\|_1$, we can compute the primal iterate with, for $i = 1, \dots, d$, $\nabla_i h_n^*(\eta) = \operatorname{sign}(\eta(i)) \max(|\eta(i)| - n\gamma\lambda, 0)$. Therefore the SDA method is equivalent to the iteration:
$$\theta_n(i) = -\operatorname{sign}\Big(\sum_{k=0}^{n-1} \nabla_i f_{k+1}(\theta_k)\Big) \max\Big(\Big|\sum_{k=0}^{n-1} \nabla_i f_{k+1}(\theta_k)\Big| - n\gamma\lambda, 0\Big) \text{ for } i = 1, \dots, d.$$
Yet since convergence results hold on the average of the iterates $\bar{\theta}_n$, SDA provides less sparse solutions than other methods which rather consider final iterates as outputs.

Kullback-Leibler divergence. The negative entropy $h(\theta) = \sum_{i=1}^n \theta(i)\log(\theta(i))$ is a Legendre function on $\operatorname{dom} h = (0, \infty)^n$ whose associated Bregman divergence is the Kullback-Leibler divergence
$$D_h(\alpha, \beta) = \sum_{i=1}^n \alpha(i) \log\Big(\frac{\alpha(i)}{\beta(i)}\Big) + \sum_{i=1}^n (\beta(i) - \alpha(i)),$$
and its conjugate gradient mapping is $\nabla_i h^*(\eta) = \exp(\eta_i)$ for $i = 1, \dots, d$.

Since h is 1-strongly convex with respect to the ℓ_1-norm (see, e.g., Beck and Teboulle, 2003, Proposition 5.1), **(LC)** holds, for example, if f is smooth with regards to the ℓ_1-norm. This illustrates one of the non-Euclidean benefit since Lipschitz constants under the ℓ_∞-norm are smaller than under the ℓ_2-norm.

This geometry is particularly appropriated to constrained minimization on the simplex Δ_d. With $g(\theta) = 1_{\Delta_d}$, SDA update is the dual averaging analogue of the exponentiated gradient algorithm (Kivinen and Warmuth, 1997):
$$\theta_n(i) = \frac{\exp(\eta_n(i))}{\sum_{j=1}^d \exp(\eta_n(j))} \text{ for } i = 1, \dots, d.$$

ℓ_p-norm. The choice $h = \frac{1}{2(p-1)}\|\cdot\|_p^2$ for $p \in (1, 2]$ is believed to adapt to the geometry of learning problem and is often used with $p = 1 + 1/\log(d)$ in association with ℓ_1-regularization (see, e.g., Duchi et al., 2010). Its Fenchel conjugate is the squared conjugate norm $h^* = \frac{1}{2(q-1)}\|\cdot\|_q^2$

for $1/p + 1/q = 1$ and its conjugate gradient mapping is $\nabla_i h^*(\eta) = \frac{\text{sign}(\eta(i))|\eta(i)|^{q-1}}{(q-1)\|\eta\|_q^{q-2}}$ (see, e.g., Gentile and Littlestone, 1999). For ℓ_1-regularization, this yields to:

$$\nabla_i h_n^*(\eta) = \nabla_i h^*\big(\text{sign}(\eta(i))\max(|\eta(i)| - n\gamma\lambda, 0)\big) \text{ for } i = 1, \dots, d.$$

The function h is 1-strongly convex with respect to the ℓ_p-norm (see, e.g., Hanner, 1956). Therefore **(LC)** holds if f is smooth with respect to the the ℓ_p-norm. However when the function f considered is quadratic as in Section 3, we can directly show that **(LC)** holds under tighter conditions on the Hessian matrix Σ (see proof in Appendix J).

Proposition 10. *Assume that $f(\theta) = \frac{1}{2}\langle\theta, \Sigma\theta\rangle$ and $h(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$. Then $h - \gamma f$ is convex for any constant step-size γ such that*

$$\gamma \leq \min_{\alpha} \frac{\|\alpha\|_p^2}{\langle\alpha, \Sigma\alpha\rangle}.$$

When $\Sigma = \mathbb{E}(x \otimes x)$ is a covariance matrix as in Section 3.2, $\langle\alpha, \Sigma\alpha\rangle = \mathbb{E}\langle x, \alpha\rangle^2 \leq \mathbb{E}\|x\|_q^2\|\alpha\|_p^2$ by Hölder inequality, and Proposition 10 admits the following corollary.

Corollary 2. *Assume that $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x, \theta\rangle - y)^2$, $h(\theta) = \frac{1}{2}\|\theta\|_p^2$ and q such that $1/p + 1/q = 1$. Then $h - \gamma f$ is convex for any constant step-size γ such that*

$$\gamma \leq 1/\mathbb{E}\|x\|_q^2.$$

Therefore we may use the algorithm with bigger step-size than in the Euclidean case. Moreover when the algorithm is started from $\theta_0 = 0$, the Bregman divergence is $D_h(\theta_*, \theta_0) = \frac{1}{2(p-1)}\|\theta_*\|_p^2$ and the bias in Proposition 1 would be bounded by $\frac{\mathbb{E}\|x\|_q^2\|\theta_*\|_p^2}{2(p-1)}$.

For high-dimension problems, taking $q = 1 + \log(d)$ (with $p \sim 1$ and $q \sim +\infty$) yields to bounds depending on the ℓ_1-norm of the optimal predictor and the ℓ_∞-norm of the features which is advisable for sparse problems.

J Proof of Proposition 10

We consider here $h(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$. For $\theta \in \mathbb{R}^d$, h is twice differentiable. Its gradient is

$$\nabla_i h(\theta) = \frac{\text{sign}(\theta(i))|\theta(i)|^{p-1}}{(p-1)\|\theta\|_p^{p-2}},$$

and its Hessian may be written for $\alpha = \frac{2-p}{(p-1)}\|\theta\|_p^{-2(p-1)}$, $u(i) = \|\theta\|_p^{2-p}\theta(i)^{p-2}$ and $v(i) = \theta(i)^{p-1}$ for $i = 1, \dots, d$, as

$$\nabla^2 h(\theta) = \text{Diag}(u) + \alpha vv^\top,$$

The function $h - \gamma f$ is convex if and only if $\nabla^2 h(\theta) \preccurlyeq \gamma\Sigma$ for all $\theta \in \mathbb{R}^d$. This condition is equivalent to

$$\min_{\theta} \min_{\alpha} \frac{\langle\alpha, \nabla^2 h(\theta)\alpha\rangle}{\langle\alpha, \Sigma\alpha\rangle} \geq \gamma.$$

A sufficient condition is that $\operatorname{Diag} u \succcurlyeq \gamma\Sigma$. After a change of variables, u may be written as $u(i) = \eta(i)^{p-2}$ where $\eta(i) = |\theta(i)|/\|\theta\|_p$ satisfies $\sum_{i=1}^d \eta(i)^p = 1$ and $\eta(i) \geq 0$. Hence for all $\theta, \alpha \in \mathbb{R}^d$

$$\langle \alpha, \nabla^2 h(\theta)\alpha\rangle \geq \sum_{i=1}^d \alpha(i)^2 u(i) = \sum_{i=1}^d \alpha(i)^2 \eta(i)^{p-2},$$

which implies

$$\min_{\theta\in\mathbb{R}^d} \langle \alpha, \nabla^2 h(\theta)\alpha\rangle \geq \min_{\eta\in\mathbb{R}^d} \sum_{i=1}^d \alpha(i)^2\eta(i)^{p-2} \text{ such that } \sum_{i=1}^d \eta(i)^p = 1 \text{ and } \eta(i) \geq 0.$$

This optimization problem is equivalent with $v(i) = \eta(i)^p$ to the one the simplex Δ_d

$$\min_{v\in\mathbb{R}^d} \sum_{i=1}^d \alpha(i)^2 v(i)^{1-2/p} \text{ such that } \sum_{i=1}^d v(i) = 1 \text{ and } \nu(i) \geq 0,$$

for which we define the Lagrangian $\mathcal{L}(v, \lambda, \mu) = \sum_{i=1}^d \alpha(i)^2 v(i)^{1-2/p} - \langle \lambda, v\rangle + \nu(1 - \sum_{i=1}^d v(i))$ for $\lambda \in \mathbb{R}_+^d$ and $\mu \in \mathbb{R}$. Its gradient is $\nabla_{v(i)}\mathcal{L}(v, \lambda, \mu) = (1 - 2/p)\alpha(i)^2/v(i)^{2/p} - \lambda(i) - \nu$. Writing the KKT condition for this problem (see, e.g., Boyd and Vandenberghe, 2004), we have that (v, λ, ν) is optimal if and only if $(1 - 2/p)\alpha(i)^2/v(i)^{2/p} - \lambda(i) - \nu = 0$, $\sum_{i=1}^d v(i) = 1$ and for all i; $\lambda(i) \geq 0$, $v(i) \geq 0$ and $\lambda(i)v(i) = 0$. These conditions are satisfied by $v(i) = \frac{\alpha(i)^p}{\sum_{i=1}^d a(j)^p}$, $\alpha(i) = 0$ and $\nu = (1 - 2p)(\sum_{i=1}^d a(j)^p)^{2/p}$. Hence the minimum value is

$$\sum_{i=1}^d \alpha(i)^2 v(i)^{1-2/p} = \sum_{i=1}^d \alpha(i)^2 \frac{\alpha(i)^{p-2}}{(\sum_{i=1}^d a(j)^p)^{1-2/p}} = \frac{\sum_{i=1}^d a(j)}{(\sum_{i=1}^d a(j)^p)^{1-2/p}} = \|\alpha\|_p^2.$$

Consequently

$$\langle \alpha, \nabla^2 h(\theta)\alpha\rangle \geq \|\alpha\|_p^2,$$

and $h - \gamma f$ is convex for $\gamma \leq \min_{\alpha\in\mathbb{R}^d} \frac{\|\alpha\|_p^2}{\langle \alpha, \Sigma\alpha\rangle}$.

K Standard benchmarks

We have considered the *sido* dataset which is often used for comparing large-scale optimization algorithms. This is a *finite* binary classification dataset with finite number of observations with outputs in $\{-1, 1\}$. We have followed the following experimental protocol: (1) remove all outliers, i.e., sample points x_n whose norms is greater than 5 times the average norm. (2) divide the dataset in two equal parts, one for training, one for testing, (3) start the algorithms from $\theta_0 = 0$, (4) sample within the training dataset with replacement, for 100 times the number of observations in the training set; a dashed line marks the first effective pass in all plots, (5) compute averaged cost on training and testing data based on 10 replications. All cost are shown in log-scale, normalized to that the first iteration leads to $\psi(\theta_0) - \psi(\theta_*) = 1$.

We solved a ℓ_1-regularized least-squares regression for three different values of ℓ_1-regularization: (1) one with the λ_* which corresponds to the best generalization error after 500 effective passes through the train set, (2) one with $\lambda_*/8$ and (3) one with $256\lambda_*$.

We compare five algorithms: averaged SGD with constant step-size, average SGD with decreasing step-size $C/(R^2\sqrt{n})$, SDA with constant step-size, SDA with decreasing step-size $C/(R^2\sqrt{n})$ and SAGA with constant step-size (Defazio et al., 2014), which showed state-of-the-art performance in the set-up of finite data sets. We consider the theoretical value of step-size which ensures convergence. We note the behaviors are comparable to the situation where step-sizes with the best testing error after one effective pass through the data (testing powers of 4 times the theoretical step-size) are used.

We can make the following observations:

- We show results for $\lambda = \lambda_*$ in Figure 2. SAGA, constant-step-size SDA and constant-step-size SGD exhibit the best behavior for both settings of step-size. However the training error of SGD does not converge to 0. On the other hand, SGD and SDA with step-size decaying as $C/R^2\sqrt{n}$ are slower. SAGA and constant-step-size SDA exhibit some overfitting after more than 10 passes on the regularized objective ψ.
- We show results for $\lambda = \lambda_*/8$ in Figure 3. The problem is then very little regularized and the behavior of constant-step-size SGD gets closer to constant-step-size SDA. There is here still overfitting for the regularized objective ψ.
- We show results for $\lambda = 256\lambda_*$ in Figure 2. The problem is then much more regularized. In this case the regularization has an important weight and the stochasticity of the quadratic objective plays a minor role. Therefore SAGA exhibits the best behavior, despite strong early oscillations, with a linear convergence but reaches a saturation point after few passes over the data. On the other hand, constant-step-size SDA exhibits a sublinear convergence which is faster at the beginning and catches up with SAGA at the end. Constant-step-size SGD is not converging to the solution.

To conclude, constant-step-size SDA behaves similarly to SAGA which is specially dedicated to the set-up of finite data sets. For larger datasets, where only a single pass is possible, SAGA could not be run. Moreover SAGA does not come with generalization guarantees while SDA does (if a single pass is made).

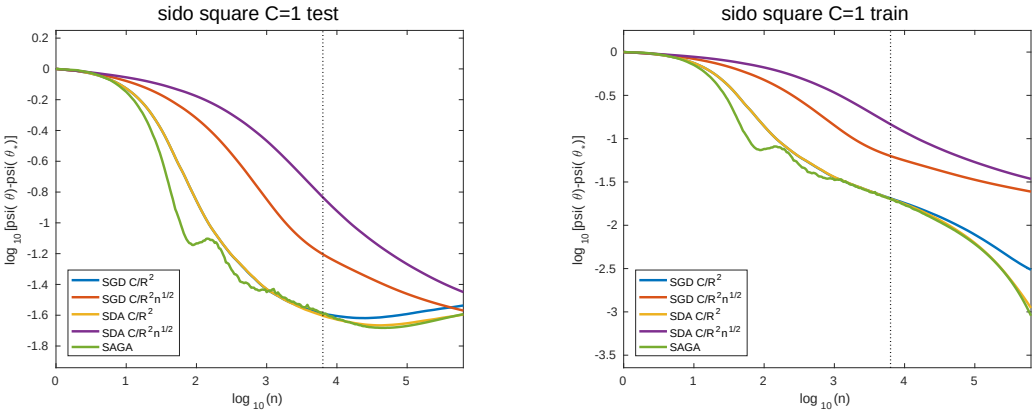

Figure 2: Test and train performances for ℓ_1-regularized least-squares regression on the *sido* dataset with $\lambda = \lambda_{\text{opt}}$. Left: test performance. Right: train performance.

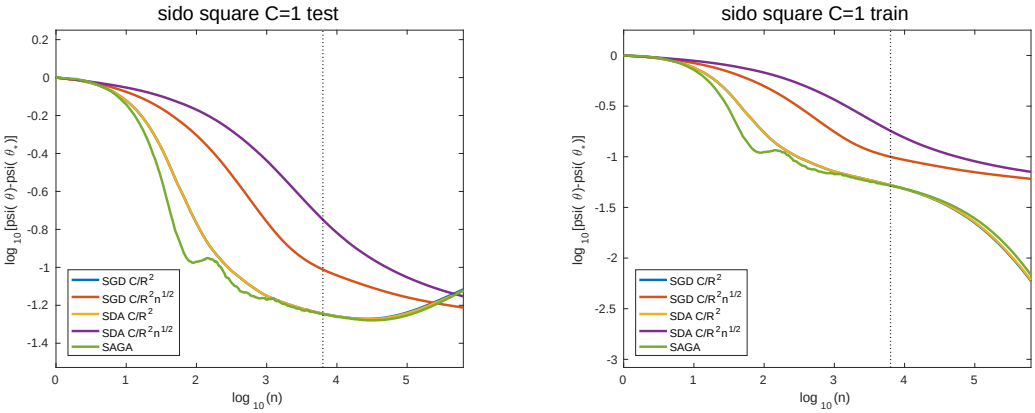

Figure 3: Test and train performances for ℓ_1-regularized least-squares regression on the *sido* dataset with $\lambda = \frac{\lambda_{\text{opt}}}{8}$. Left: test performance. Right: train performance.

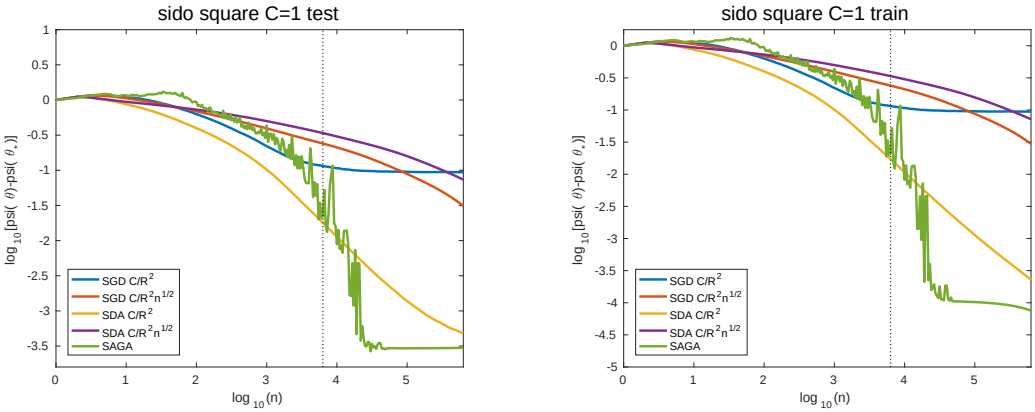

Figure 4: Test and train performances for ℓ_1-regularized least-squares regression on the *sido* dataset with $\lambda = 256\lambda_{\text{opt}}$. Left: test performance. Right: train performance.