Optimized Stochastic Approximation Algorithms for Non-Strongly Convex Problems: Achieving Fast Convergence Rates

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 16, 2024

Abstract

We present novel stochastic approximation algorithms that achieve the convergence rate of $O(1/n)$ for both least-squares and logistic regression, without assuming strong convexity. Our methods leverage the smoothness of the loss functions to improve convergence rates compared to non-smooth approaches. In particular, we propose an averaged stochastic gradient descent (SGD) algorithm with constant step-size that attains optimal rates for both convex and certain non-convex problems. We provide non-asymptotic analysis of the generalization error and demonstrate the effectiveness of our approach through extensive experiments on standard machine learning benchmarks. Our algorithms achieve superior performance compared to existing methods in high-dimensional settings, offering robust theoretical guarantees and practical advantages.

Keywords: Stochastic Approximation, Non-Strongly Convex Optimization, Stochastic Gradient Descent, Convergence Rate, Logistic Regression, Least-Squares Regression

1 Introduction

Large-scale machine learning problems are increasingly prevalent in diverse scientific and engineering fields, where efficient algorithms are critical for processing large datasets. Stochastic gradient descent (SGD) and its variants, though introduced over 60 years ago, remain foundational in this context. Their simplicity, coupled with the ability to process data in an online fashion, makes them attractive for high-dimensional problems. However, their convergence properties, particularly in the absence of strong convexity, are not fully optimized.

Most practical machine learning problems involve functions that are non-strongly convex or even non-convex, making it difficult to achieve fast convergence rates. Traditional SGD methods for such problems often achieve a convergence rate of $O(1/\sqrt{n})$, which is suboptimal for high-dimensional data. In this paper, we address this issue by designing stochastic approximation algorithms that achieve an $O(1/n)$ convergence rate, even in the absence of strong convexity.

We focus on two widely used loss functions: least-squares and logistic regression. For both, we demonstrate that averaging with constant step-size leads to optimal rates without requiring strong convexity. Our theoretical analysis shows that smoothness, rather than strong convexity, plays a key role in achieving these rates.

History. Stochastic approximation algorithms such as stochastic gradient descent (SGD) and its variants, although introduced more than 60 years ago [1], still remain the most widely used and studied method in this context (see, e.g., [2, 3, 4, 5, 6, 7]). We consider minimizing convex functions f, defined on a Euclidean space $\mathcal{H}$, given by $f(\theta) = \mathbb{E}\big[\ell(y, \langle \theta, x \rangle)\big]$, where $(x, y) \in \mathcal{H} \times \mathbb{R}$ denotes the data and ℓ denotes a loss function that is convex with respect to the second variable. This includes logistic and least-squares regression. In the stochastic approximation framework, independent and identically distributed pairs (x_n, y_n) are observed sequentially and the predictor defined by θ is updated after each pair is seen.

We partially understand the properties of f that affect the problem difficulty. *Strong convexity* (i.e., when f is twice differentiable, a uniform strictly positive lower-bound μ on Hessians of f) is a key property. Indeed, after n observations and with the proper step-sizes, averaged SGD achieves the rate of $O(1/\mu n)$ in the strongly-convex case [5, 4], while it achieves only $O(1/\sqrt{n})$ in the non-strongly-convex case [5], with matching lower-bounds [8, 9].

The main issue with strong convexity is that typical machine learning problems are high dimensional and have correlated variables so that the strong convexity constant μ is zero or very close to zero, and in any case smaller than $O(1/\sqrt{n})$. This then makes the non-strongly convex methods better. In this paper, we aim at obtaining algorithms that may deal with arbitrarily small strong-convexity constants, but still achieve a rate of $O(1/n)$.

Smoothness plays a central role in the context of deterministic optimization. The known convergence rates for smooth optimization are better than for non-smooth optimization (e.g., see [10]). However, for stochastic optimization the use of smoothness only leads to improvements on constants (e.g., see [11]) but not on the rate itself, which remains $O(1/\sqrt{n})$ for non-strongly-convex problems.

We show that for the square loss and for the logistic loss, we may use the smoothness of the loss and obtain algorithms that have a convergence rate of $O(1/n)$ without any strong convexity assumptions. More precisely, for least-squares regression, we show in Section 2 that *averaged* stochastic gradient descent *with constant step-size* achieves the desired rate. For logistic regression this is achieved by a novel stochastic gradient algorithm that (a) constructs successive local quadratic approximations of the loss functions, while (b) preserving the same running time complexity as stochastic gradient descent (see Section 3). For these algorithms, we provide a non-asymptotic analysis of their generalization error (in expectation, and also in high probability for least-squares), and run extensive experiments on standard machine learning benchmarks showing in Section 4 that they often outperform existing approaches.

Contribution. The main contributions of this paper are threefold. First, we present a new analysis of stochastic approximation algorithms that attain the optimal $O(1/n)$ convergence rate without assuming strong convexity. Second, we introduce a novel SGD variant for logistic regression that leverages smoothness to improve both theoretical guarantees and empirical performance. Third, we validate our findings through extensive experiments on benchmark datasets, showing superior performance over existing methods.

Organization. The remainder of the paper is structured as follows. In Section 2, we introduce the constant-step-size least-mean-square algorithm and analyze its convergence properties. Section 3 extends our approach to M-estimation, with a focus on logistic regression, and provides theoretical guarantees. In Section 4, we present experimental results demonstrating the efficacy of our methods. Finally, Section 5 concludes with a discussion of future directions.

2 Constant-step-size least-mean-square algorithm

In this section, we consider stochastic approximation for least-squares regression, where SGD is often referred to as the least-mean-square (LMS) algorithm. The novelty of our convergence result is the use of the constant step-size with averaging, leading to $O(1/n)$ rate without strong convexity.

2.1 Convergence in expectation

We make the following assumptions:

(A1) $\mathcal{H}$ is a d-dimensional Euclidean space, with $d \geqslant 1$.

(A2) The observations $(x_n, z_n) \in \mathcal{H} \times \mathcal{H}$ are independent and identically distributed.

(A3) $\mathbb{E}\|x_n\|^2$ and $\mathbb{E}\|z_n\|^2$ are finite. Denote by $H = \mathbb{E}(x_n \otimes x_n)$ the covariance operator from $\mathcal{H}$ to $\mathcal{H}$. Without loss of generality, H is assumed invertible (by projecting onto the minimal subspace where x_n lies almost surely). However, its eigenvalues may be arbitrarily small.

(A4) The global minimum of $f(\theta) = (1/2)\mathbb{E}\big[\langle\theta, x_n\rangle^2 - 2\langle\theta, z_n\rangle\big]$ is attained at a certain $\theta_* \in \mathcal{H}$. We denote by $\xi_n = z_n - \langle\theta_*, x_n\rangle x_n$ the residual. We have $\mathbb{E}\big[\xi_n\big] = 0$, but in general, it is not true that $\mathbb{E}^{\xi_n}_{1=,2=}\big[x_n \,\big|\!=\big]\, 0$ (unless the model is well-specified).

(A5) We study the stochastic gradient (a.k.a. least mean square) recursion defined as

$$\theta_n = \theta_{n-1} - \gamma(\langle\theta_{n-1}, x_n\rangle x_n - z_n) = (I - \gamma x_n \otimes x_n)\theta_{n-1} + \gamma z_n \tag{1}$$

started from $\theta_0 \in \mathcal{H}$. We also consider the averaged iterates $\bar{\theta}_n = (n+1)^{-1}\sum_{k=0}^{n}\theta_k$.

(A6) There exists $R > 0$ and $\sigma > 0$ such that $\mathbb{E}\big[\xi_n \otimes \xi_n\big] \preccurlyeq \sigma^2 H$ and $\mathbb{E}\big(\|x_n\|^2 x_n \otimes x_n\big) \preccurlyeq R^2 H$, where $\preccurlyeq$ denotes the the order between self-adjoint operators, i.e., $A \preccurlyeq B$ if and only if $B - A$ is positive semi-definite.

Discussion of assumptions. Assumptions **(A1-5)** are standard in stochastic approximation (see, e.g., [12, 6]). Note that for least-squares problems, z_n is of the form $y_n x_n$, where $y_n \in \mathbb{R}$ is the response to be predicted as a linear function of x_n. We consider a slightly more general case than least-squares because we will need it for the quadratic approximation of the logistic loss in Section 3.1. Note that in assumption **(A4)**, we do not assume that the model is well-specified.

Assumption **(A6)** is true for least-square regression with almost surely bounded data, since, if $\|x_n\|^2 \leqslant R^2$ almost surely, then $\mathbb{E}\big(\|x_n\|^2 x_n \otimes x_n\big) \preccurlyeq \mathbb{E}\big(R^2 x_n \otimes x_n\big) = R^2 H$; a similar inequality holds for the output variables y_n. Moreover, it also holds for data with infinite supports, such as Gaussians or mixtures of Gaussians (where all covariance matrices of the mixture components are lower and upper bounded by a constant times the same matrix). Note that the finite-dimensionality assumption could be relaxed, but this would require notions similar to degrees of freedom [13], which is outside of the scope of this paper.

The goal of this section is to provide a bound on the expectation $\mathbb{E}\big[f(\bar{\theta}_n) - f(\theta_*)\big]$, that (a) does not depend on the smallest non-zero eigenvalue of H (which could be arbitrarily small) and (b) still scales as $O(1/n)$.

Theorem 1. *Assume **(A1-6)**. For any constant step-size $\gamma < \frac{1}{R^2}$, we have*

$$\mathbb{E}\big[f(\overline{\theta}_{n-1}) - f(\theta_*)\big] \leqslant \frac{1}{2n}\left[\frac{\sigma\sqrt{d}}{1-\sqrt{\gamma R^2}} + R\|\theta_0 - \theta_*\|\frac{1}{\sqrt{\gamma R^2}}\right]^2 \tag{2}$$

When $\gamma = 1/(4R^2)$, we obtain $\mathbb{E}\big[f(\overline{\theta}_{n-1}) - f(\theta_*)\big] \leqslant \frac{2}{n}\Big[\sigma\sqrt{d} + R\|\theta_0 - \theta_*\|\Big]^2$.

Proof technique. We adapt and extend a proof technique from [14] which is based on non-asymptotic expansions in powers of γ. We also use a result from [2] which studied the recursion in Eq. (1), with $x_n \otimes x_n$ replaced by its expectation H. See the appendix for details.

Optimality of bounds. Our bound in Eq. (2) leads to a rate of $O(1/n)$, which is known to be optimal for least-squares regression (i.e., under reasonable assumptions, no algorithm, even more complex than averaged SGD can have a better dependence in n) [15]. The term $\sigma^2 d/n$ is also unimprovable.

Initial conditions. If γ is small, then the initial condition is forgotten more slowly. Note that with additional strong convexity assumptions, the initial condition would be forgotten faster (exponentially fast without averaging), which is one of the traditional uses of constant-step-size LMS [16].

Specificity of constant step-sizes. The non-averaged iterate sequence (θ_n) is a homogeneous Markov chain; under appropriate technical conditions, this Markov chain has a unique stationary (invariant) distribution and the sequence of iterates (θ_n) converges in distribution to this invariant distribution; see [17, Chapter 17]. Denote by π_γ the invariant distribution. Assuming that the Markov Chain is Harris recurrent, the ergodic theorem for Harris Markov chain shows that $\overline{\theta}_{n-1} = n^{-1}\sum_{k=0}^{n-1}\theta_k$ converges almost-surely to $\overline{\theta}_\gamma \overset{\text{def}}{=} \int \theta \pi_\gamma(\mathrm{d}\theta)$, which is the mean of the stationary distribution. Taking the expectation on both side of Eq. (1), we get $\mathbb{E}[\theta_n] - \theta_* = (I - \gamma H)(\mathbb{E}[\theta_{n-1}] - \theta_*)$, which shows, using that $\lim_{n\to\infty}\mathbb{E}[\theta_n] = \overline{\theta}_\gamma$ that $H\overline{\theta}_\gamma = H\theta_*$ and therefore $\overline{\theta}_\gamma = \theta_*$ since H is invertible. Under slightly stronger assumptions, it can be shown that

$$\lim_{n\to\infty} n\mathbb{E}[(\overline{\theta}_n - \theta_*)^2] = \mathrm{Var}_{\pi_\gamma}(\theta_0) + 2\sum_{k=1}^{\infty}\mathrm{Cov}_{\pi_\gamma}(\theta_0, \theta_k)\ ,$$

where $\mathrm{Cov}_{\pi_\gamma}(\theta_0, \theta_k)$ denotes the covariance of θ_0 and θ_k when the Markov chain is started from stationarity. This implies that $\lim_{n\to\infty} n\mathbb{E}[f(\overline{\theta}_n) - f(\theta_*)]$ has a finite limit. Therefore, this interpretation explains why the averaging produces a sequence of estimators which converges to the solution θ_* pointwise, and that the rate of convergence of $\mathbb{E}[f(\theta_n) - f(\theta_*)]$ is of order $O(1/n)$. Note that for other losses than quadratic, the same properties hold *except* that the mean under the stationary distribution does not coincide with θ_* and its distance to θ_* is typically of order γ^2 (see Section 3).

2.2 Convergence in higher orders

We are now going to consider an extra assumption in order to bound the p-th moment of the excess risk and then get a high-probability bound. Let p be a real number greater than 1.

(A7) There exists $R > 0$, $\kappa > 0$ and $\tau \geqslant \sigma > 0$ such that, for all $n \geqslant 1$, $\|x_n\|^2 \leqslant R^2$ a.s., and

$$\mathbb{E}\|\xi_n\|^p \leqslant \tau^p R^p \quad \text{and} \quad \mathbb{E}\big[\xi_n \otimes \xi_n\big] \preccurlyeq \sigma^2 H, \tag{3}$$

$$\forall z \in \mathcal{H}, \quad \mathbb{E}\langle z, x_n\rangle^4 \leqslant \kappa \langle z, Hz\rangle^2 \tag{4}$$

The last condition in Eq. (4) says that the *kurtosis* of the projection of the covariates x_n on any direction $z \in \mathcal{H}$ is bounded. Note that computing the constant κ happens to be equivalent to the optimization problem solved by the FastICA algorithm [18], which thus provides an estimate of κ. In Table 1, we provide such an estimate for the non-sparse datasets which we have used in experiments, while we consider only directions z along the axes for high-dimensional sparse datasets. For these datasets where a given variable is equal to zero except for a few observations, κ is typically quite large. Adapting and analyzing normalized LMS techniques [19] to this set-up is likely to improve the theoretical robustness of the algorithm (but note that results in expectation from Theorem 1 do not use κ). The next theorem provides a bound for the p-th moment of the excess risk.

Theorem 2. *Assume **(A1-7)**. For any real $p \geqslant 1$, and for a step-size $\gamma \leqslant 1/(12p\kappa R^2)$, we have:*

$$\big(\mathbb{E}\big|f(\overline{\theta}_{n-1}) - f(\theta_*)\big|^p\big)^{1/p} \leqslant \frac{p}{2n}\left(7\tau\sqrt{d} + R\|\theta_0 - \theta_*\|\sqrt{3 + \frac{2}{\gamma p R^2}}\right)^2 \tag{5}$$

For $\gamma = 1/(12p\kappa R^2)$, we get: $\big(\mathbb{E}\big|f(\overline{\theta}_{n-1}) - f(\theta_*)\big|^p\big)^{1/p} \leqslant \frac{p}{2n}\big(7\tau\sqrt{d} + 6\sqrt{\kappa}R\|\theta_0 - \theta_*\|\big)^2$.

Note that to control the p-th order moment, a smaller step-size is needed, which scales as $1/p$. We can now provide a high-probability bound; the tails decay polynomially as $1/(n\delta^{12\gamma\kappa R^2})$ and the smaller the step-size γ, the lighter the tails.

Corollary 1. *For any step-size such that $\gamma \leqslant 1/(12\kappa R^2)$, any $\delta \in (0,1)$,*

$$\mathbb{P}\left(f(\overline{\theta}_{n-1}) - f(\theta_*) \geqslant \frac{1}{n\delta^{12\gamma\kappa R^2}} \frac{\big[7\tau\sqrt{d} + R\|\theta_0 - \theta_*\|(\sqrt{3} + \sqrt{24\kappa})\big]^2}{24\gamma\kappa R^2}\right) \leqslant \delta \tag{6}$$

3 Beyond least-squares: M-estimation

In Section 2, we have shown that for least-squares regression, averaged SGD achieves a convergence rate of $O(1/n)$ with no assumption regarding strong convexity. For all losses, with a constant step-size γ, the stationary distribution π_γ corresponding to the homogeneous Markov chain (θ_n) does always satisfy $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$, where f is the generalization error. When the gradient f' is linear (i.e., f is quadratic), then this implies that $f'(\int \theta\pi_\gamma(\mathrm{d}\theta)) = 0$, i.e., the averaged recursion converges pathwise to $\overline{\theta}_\gamma = \int \theta\pi_\gamma(\mathrm{d}\theta)$ which coincides with the optimal value θ_* (defined through $f'(\theta_*) = 0$). When the gradient f' is no longer linear, then $\int f'(\theta)\pi_\gamma(\mathrm{d}\theta) \neq f'(\int \theta\pi_\gamma(\mathrm{d}\theta))$. Therefore, for general M-estimation problems we should expect that the averaged sequence still converges at rate $O(1/n)$ to the mean of the stationary distribution $\overline{\theta}_\gamma$, but not to the optimal predictor θ_*. Typically, the average distance between θ_n and θ_* is of order γ (see Section 4 and [20]), while for the averaged iterates that converge pointwise to $\overline{\theta}_\gamma$, it is of order γ^2 for strongly convex problems under some additional smoothness conditions on the loss functions (these are satisfied, for example, by the logistic loss [21]).

Since quadratic functions may be optimized with rate $O(1/n)$ under weak conditions, we are going to use a quadratic approximation around a well chosen support point, which shares some similarity with the Newton procedure (however, with a non trivial adaptation to the stochastic approximation framework). The Newton step for f around a certain point $\widetilde{\theta}$ is equivalent to minimizing a quadratic surrogate g of f around $\widetilde{\theta}$, i.e., $g(\theta) = f(\widetilde{\theta}) + \langle f'(\widetilde{\theta}), \theta - \widetilde{\theta} \rangle + \frac{1}{2}\langle \theta - \widetilde{\theta}, f''(\widetilde{\theta})(\theta - \widetilde{\theta})\rangle$. If $f_n(\theta) \overset{\text{def}}{=} \ell(y_n, \langle \theta, x_n \rangle)$, then $g(\theta) = \mathbb{E} g_n(\theta)$, with $g_n(\theta) = f(\widetilde{\theta}) + \langle f_n'(\widetilde{\theta}), \theta - \widetilde{\theta}\rangle + \frac{1}{2}\langle \theta - \widetilde{\theta}, f_n''(\widetilde{\theta})(\theta - \widetilde{\theta})\rangle$; the Newton step may thus be solved approximately with stochastic approximation (here constant-step size LMS), with the following recursion:

$$\theta_n = \theta_{n-1} - \gamma g_n'(\theta_{n-1}) = \theta_{n-1} - \gamma \big[f_n'(\widetilde{\theta}) + f_n''(\widetilde{\theta})(\theta_{n-1} - \widetilde{\theta}) \big] \tag{7}$$

This is equivalent to replacing the gradient $f_n'(\theta_{n-1})$ by its first-order approximation around $\widetilde{\theta}$. A crucial point is that for machine learning scenarios where f_n is a loss associated to a single data point, its complexity is only twice the complexity of a regular stochastic approximation step, since, with $f_n(\theta) = \ell(y_n, \langle x_n, \theta \rangle)$, $f_n''(\theta)$ is a rank-one matrix.

Choice of support points for quadratic approximation. An important aspect is the choice of the support point $\widetilde{\theta}$. In this paper, we consider two strategies:

- **Two-step procedure**: for convex losses, averaged SGD with a step-size decaying at $O(1/\sqrt{n})$ achieves a rate (up to logarithmic terms) of $O(1/\sqrt{n})$ [5, 6]. We may thus use it to obtain a first decent estimate. The two-stage procedure is as follows (and uses $2n$ observations): n steps of averaged SGD with constant step size $\gamma \propto 1/\sqrt{n}$ to obtain $\widetilde{\theta}$, and then averaged LMS for the Newton step around $\widetilde{\theta}$. As shown below, this algorithm achieves the rate $O(1/n)$ for logistic regression. However, it is not the most efficient in practice.

- **Support point = current average iterate**: we simply consider the current averaged iterate $\overline{\theta}_{n-1}$ as the support point $\widetilde{\theta}$, leading to the recursion:

$$\theta_n = \theta_{n-1} - \gamma \big[f_n'(\overline{\theta}_{n-1}) + f_n''(\overline{\theta}_{n-1})(\theta_{n-1} - \overline{\theta}_{n-1}) \big] \tag{8}$$

  Although this algorithm has shown to be the most efficient in practice (see Section 4) we currently have no proof of convergence. Given that the behavior of the algorithms does not change much when the support point is updated less frequently than each iteration, there may be some connections to two-time-scale algorithms (see, e.g., [22]). In Section 4, we also consider several other strategies based on doubling tricks.

Interestingly, for non-quadratic functions, our algorithm imposes a new bias (by replacing the true gradient by an approximation which is only valid close to $\overline{\theta}_{n-1}$) in order to reach faster convergence (due to the linearity of the underlying gradients).

Relationship with one-step-estimators. One-step estimators (see, e.g., [23]) typically take any estimator with $O(1/n)$-convergence rate, and make a full Newton step to obtain an efficient estimator (i.e., one that achieves the Cramer-Rao lower bound). Although our novel algorithm is largely inspired by one-step estimators, our situation is slightly different since our first estimator has only convergence rate $O(n^{-1/2})$ and is estimated on different observations.

3.1 Self-concordance and logistic regression

We make the following assumptions:

(B1) $\mathcal{H}$ is a d-dimensional Euclidean space, with $d \geqslant 1$.

(B2) The observations $(x_n, y_n) \in \mathcal{H} \times \{-1, 1\}$ are independent and identically distributed.

(B3) We consider $f(\theta) = \mathbb{E}\big[\ell(y_n, \langle x_n, \theta \rangle)\big]$, with the following assumption on the loss function ℓ (whenever we take derivatives of ℓ, this will be with respect to the second variable):

$$\forall (y, \widehat{y}) \in \{-1, 1\} \times \mathbb{R}, \quad \ell'(y, \widehat{y}) \leqslant 1, \quad \ell''(y, \widehat{y}) \leqslant 1/4, \quad |\ell'''(y, \widehat{y})| \leqslant \ell''(y, \widehat{y})$$

We denote by θ_* a global minimizer of f, which we thus assume to exist, and we denote by $H = f''(\theta_*)$ the Hessian operator at a global optimum θ_*.

(B4) We assume that there exists $R > 0$, $\kappa > 0$ and $\rho > 0$ such that $\|x_n\|^2 \leqslant R^2$ almost surely, and

$$\mathbb{E}\big[x_n \otimes x_n\big] \preccurlyeq \rho \mathbb{E}\big[\ell''(y_n, \langle \theta_*, x_n \rangle) x_n \otimes x_n\big] = \rho H \tag{9}$$

$$\forall z \in \mathcal{H}, \theta \in \mathcal{H}, \ \mathbb{E}\big[\ell''(y_n, \langle \theta, x_n \rangle)^2 \langle z, x_n \rangle^4\big] \leqslant \kappa \big(\mathbb{E}\big[\ell''(y_n, \langle \theta, x_n \rangle) \langle z, x_n \rangle^2\big]\big)^2. \tag{10}$$

Assumption **(B3)** is satisfied for the logistic loss and extends to all generalized linear models (see more details in [21]), and the relationship between the third derivative and second derivative of the loss ℓ is often referred to as *self-concordance* (see [24, 25] and references therein). Note moreover that we must have $\rho \geqslant 4$ and $\kappa \geqslant 1$.

A loose upper bound for ρ is $1/\inf_n \ell''(y_n, \langle \theta_*, x_n \rangle)$ but in practice, it is typically much smaller (see Table 1). The condition in Eq. (10) is hard to check because it is uniform in θ. With a slightly more complex proof, we could restrict θ to be close to θ_*; with such constraints, the value of κ we have found is close to the one from Section 2.2 (i.e., without the terms in $\ell''(y_n, \langle \theta, x_n \rangle)$).

Theorem 3. *Assume **(B1-4)**, and consider the vector ζ_n obtained as follows: (a) perform n steps of averaged stochastic gradient descent with constant step size $1/2R^2\sqrt{n}$, to get $\widetilde{\theta}_n$, and (b) perform n step of averaged LMS with constant step-size $1/R^2$ for the quadratic approximation of f around $\widetilde{\theta}_n$. If $n \geqslant (19 + 9R\|\theta_0 - \theta_*\|)^4$, then*

$$\mathbb{E} f(\zeta_n) - f(\theta_*) \leqslant \frac{\kappa^{3/2} \rho^3 d}{n} (16 R \|\theta_0 - \theta_*\| + 19)^4 \tag{11}$$

We get an $O(1/n)$ convergence rate without assuming strong convexity, even locally, thus improving on results from [21] where the the rate is proportional to $1/(n \lambda_{\min}(H))$. The proof relies on self-concordance properties and the sharp analysis of the Newton step (see appendix).

4 Experiments

4.1 Synthetic data

Least-mean-square algorithm. We consider normally distributed inputs, with covariance matrix H that has random eigenvectors and eigenvalues $1/k$, $k = 1, \dots, d$. The outputs are generated from a linear function with homoscedastic noise with unit signal to noise-ratio. We consider $d = 20$

and the least-mean-square algorithm with several settings of the step size γ_n, constant or proportional to $1/\sqrt{n}$. Here R^2 denotes the *average radius of the data*, i.e., $R^2 = \operatorname{tr} H$. In the left plot of Figure 1, we show the results, averaged over 10 replications.

Without averaging, the algorithm with constant step-size does not converge pointwise (it oscillates), and its average excess risk decays as a linear function of γ (indeed, the gap between each values of the constant step-size is close to $\log_{10}(4)$, which corresponds to a linear function in γ).

With averaging, the algorithm with constant step-size does converge at rate $O(1/n)$, and for all values of the constant γ, the rate is actually the same. Moreover (although it is not shown in the plots), the standard deviation is much lower.

With decaying step-size $\gamma_n = 1/(2R^2\sqrt{n})$ and without averaging, the convergence rate is $O(1/\sqrt{n})$, and improves to $O(1/n)$ with averaging.

Logistic regression. We consider the same input data as for least-squares, but now generates outputs from the logistic probabilistic model. We compare several algorithms and display the results in Figure 1 (middle and right plots).

On the middle plot, we consider SGD. Without averaging, the algorithm with constant step-size does not converge and its average excess risk reaches a constant value which is a linear function of γ (indeed, the gap between each values of the constant step-size is close to $\log_{10}(4)$). With averaging, the algorithm does converge, but as opposed to least-squares, to a point which is not the optimal solution, with an error proportional to γ^2 (the gap between curves is twice as large).

On the right plot, we consider various variations of our Newton-approximation scheme. The "2-step" algorithm is the one for which our convergence rate holds (n being the total number of examples, we perform $n/2$ steps of averaged SGD, then $n/2$ steps of LMS). Not surprisingly, it is not the best in practice (in particular at $n/2$, when starting the constant-size LMS, the performance worsens temporarily). It is classical to use doubling tricks to remedy this problem while preserving convergence rates [26], this is done in "2-step-dbl.", which avoids the previous erratic behavior.

We have also considered getting rid of the first stage where plain averaged stochastic gradient is used to obtain a support point for the quadratic approximation. We now consider only Newton-steps but change only these support points. We consider updating the support point at every iteration, i.e., the recursion from Eq. (8), while we also consider updating it every dyadic point ("dbl.-approx"). The last two algorithms perform very similarly and achieve the $O(1/n)$ early. In all experiments on real data, we have considered the simplest variant (which corresponds to Eq. (8)).

4.2 Standard benchmarks

We have considered 6 benchmark datasets which are often used in comparing large-scale optimization methods. The datasets are described in Table 1 and vary in values of d, n and sparsity levels. These are all *finite* binary classification datasets with outputs in $\{-1, 1\}$. For least-squares and logistic regression, we have followed the following experimental protocol: (1) remove all outliers (i.e., sample points x_n whose norm is greater than 5 times the average norm), (2) divide the dataset in two equal parts, one for training, one for testing, (3) sample within the training dataset with replacement, for 100 times the number of observations in the training set (this corresponds to 100 effective passes; in all plots, a black dashed line marks the first effective pass), (4) compute averaged cost on training and testing data (based on 10 replications). All the costs are shown in log-scale, normalized to that the first iteration leads to $f(\theta_0) - f(\theta_*) = 1$.

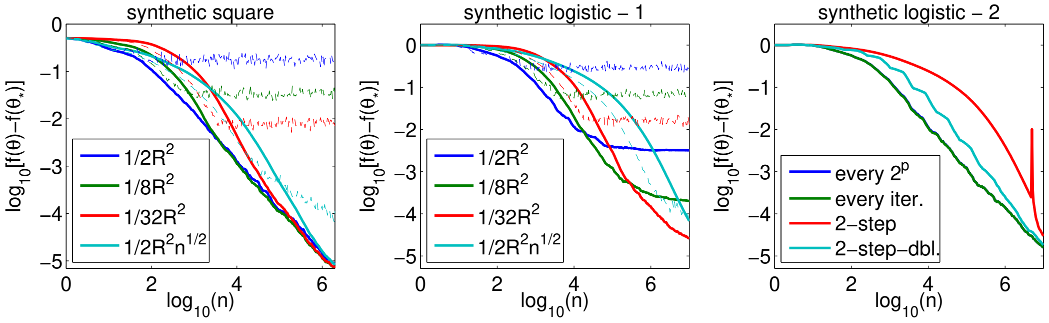

Figure 1. Synthetic data. Left: least-squares regression. Middle: logistic regression with averaged SGD with various step-sizes, averaged (plain) and non-averaged (dashed). Right: various Newton-based schemes for the same logistic regression problem. Best seen in color. See text for details

All algorithms that we consider (ours and others) have a step-size, and typically a theoretical value that ensures convergence. We consider two settings: (1) one when this theoretical value is used, (2) one with the best testing error after one effective pass through the data (testing powers of 4 times the theoretical step-size).

Here, we only consider *covertype*, *alpha*, *sido* and *news*, as well as test errors. For all training errors and the two other datasets (*quantum*, *rcv1*), see the appendix.

Least-squares regression. We compare three algorithms: averaged SGD with constant step-size, averaged SGD with step-size decaying as $C/R^2\sqrt{n}$, and the stochastic averaged gradient (SAG) method which is dedicated to finite training data sets [27], which has shown state-of-the-art performance in this set-up[1]. We show the results in the two left plots of Figure 2 and Figure 3.

Averaged SGD with decaying step-size equal to $C/R^2\sqrt{n}$ is slowest (except for *sido*). In particular, when the best constant C is used (right columns), the performance typically starts to increase significantly. With that step size, even after 100 passes, there is no sign of overfitting, even for the high-dimensional sparse datasets.

SAG and constant-step-size averaged SGD exhibit the best behavior, for the theoretical step-sizes and the best constants, with a significant advantage for constant-step-size SGD. The non-sparse datasets do not lead to overfitting, even close to the global optimum of the (unregularized) training objectives, while the sparse datasets do exhibit some overfitting after more than 10 passes.

Logistic regression. We also compare two additional algorithms: our Newton-based technique and "Adagrad" [7], which is a stochastic gradient method with a form a diagonal scaling[2] that allows to reduce the convergence rate (which is still in theory proportional to $O(1/\sqrt{n})$). We show results in the two right plots of Figure 2 and Figure 3.

Averaged SGD with decaying step-size proportional to $1/R^2\sqrt{n}$ has the same behavior than for least-squares (step-size harder to tune, always inferior performance except for *sido*).

SAG, constant-step-size SGD and the novel Newton technique tend to behave similarly (good with theoretical step-size, always among the best methods). They differ notably in some aspects:

[1] The original algorithm from [27] is considering only strongly convex problems, we have used the step-size of $1/16R^2$, which achieves fast convergence rates in all situations (see http://research.microsoft.com/en-us/um/cambridge/events/mls2013/downloads/stochastic_gradient.pdf).

[2] Since a bound on $\|\theta_*\|$ is not available, we have used step-sizes proportional to $1/\sup_n \|x_n\|_\infty$.

Table 1. Datasets used in our experiments. We report the proportion of non-zero entries, as well as estimates for the constant κ and ρ used in our theoretical results, together with the non-sharp constant which is typically used in analysis of logistic regression and which our analysis avoids (these are computed for non-sparse datasets only).

Name	d	n	sparsity	κ	ρ	$1/\inf_n \ell''(y_n, \langle\theta_*, x_n\rangle)$
quantum	79	50 000	100 %	5.8×10^2	16	8.5×10^2
covertype	55	581 012	100 %	9.6×10^2	160	3×10^{12}
alpha	501	500 000	100 %	6	18	8×10^4
sido	4 933	12 678	10 %	1.3×10^4	×	×
rcv1	47 237	20 242	0.2 %	2×10^4	×	×
news	1 355 192	19 996	0.03 %	2×10^4	×	×

(1) SAG converges quicker for the training errors (shown in the appendix) while it is a bit slower for the testing error, (2) in some instances, constant-step-size averaged SGD does underfit (*covertype*, *alpha*, *news*), which is consistent with the lack of convergence to the global optimum mentioned earlier, (3) the novel Newton approximation is consistently better.

On the non-sparse datasets, Adagrad performs similarly to the Newton-type method (often better in early iterations and worse later), except for the *alpha* dataset where the step-size is harder to tune (the best step-size tends to have early iterations that make the cost go up significantly). On sparse datasets like *rcv1*, the performance is essentially the same as Newton. On the *sido* data set, Adagrad (with fixed steps size, left column) achieves a good testing loss quickly then levels off, for reasons we cannot explain. On the *news* dataset, it is inferior without parameter-tuning and a bit better with. Adagrad uses a diagonal rescaling; it could be combined with our technique, early experiments show that it improves results but that it is more sensitive to the choice of step-size.

Overall, even with d and κ very large (where our bounds are vacuous), the performance of our algorithm still achieves the state of the art, while being more robust to the selection of the step-size: finer quantities likes degrees of freedom [13] should be able to quantify more accurately the quality of the new algorithms.

5 Conclusion

In this paper, we have considered the stochastic approximation problem where a convex function must be minimized using only unbiased estimates of its gradients, a common framework in machine learning based on empirical risk minimization. We focused on problems without strong convexity, where previously known algorithms typically achieve a convergence rate of $O(1/\sqrt{n})$. We presented two stochastic approximation algorithms that instead achieve an optimal rate of $O(1/n)$ for classical supervised learning tasks, such as least-squares and logistic regression. For least-squares regression, we showed that *averaged* stochastic gradient descent with a *constant step-size* attains this rate. For logistic regression, we introduced a novel stochastic gradient algorithm that constructs local quadratic approximations while maintaining the same computational complexity as standard SGD.

Our theoretical contributions include a non-asymptotic analysis of the generalization error, both in expectation and with high probability for least-squares regression. Extensive experiments on standard machine learning benchmarks demonstrate that our algorithms outperform existing methods, especially in high-dimensional settings.

Future directions for extending this work include: (a) an analysis of algorithms that update the quadratic approximation's support point at every iteration, (b) proximal extensions that may be more complex to analyze but easier to implement, (c) adaptive strategies for determining the constant step-size, (d) step-sizes that vary with the iterates to enhance robustness, as in normalized LMS, and (e) non-parametric analysis to improve theoretical results for large-dimensional problems.

References

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

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

[3] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Adv. NIPS*, 2008.

[4] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. ICML*, 2007.

[5] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[6] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*, 2011.

[7] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2010.

[8] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley & Sons, 1983.

[9] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Adv. NIPS*, 2009.

[10] Y. Nesterov. *Introductory lectures on convex optimization: a basic course.* Kluwer Academic Publishers, 2004.

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

[12] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications.* Springer-Verlag, second edition, 2003.

[13] C. Gu. *Smoothing spline ANOVA models.* Springer, 2002.

[14] R. Aguech, E. Moulines, and P. Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control and Optimization*, 39(3):872–899, 2000.

[15] A. B. Tsybakov. Optimal rates of aggregation. In *Proc. COLT*, 2003.

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

[17] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Cambridge University Press, London, 2009.

[18] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[19] N.J. Bershad. Analysis of the normalized lms algorithm with gaussian inputs. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):793–806, 1986.

[20] A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. *Stochastic Optimization: Algorithms and Applications*, pages 263–304, 2000.

[21] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. Technical Report 00804431, HAL, 2013.

[22] V. S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[23] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge Univ. press, 2000.

[24] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming.* SIAM studies in Applied Mathematics, 1994.

[25] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[26] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proc. COLT*, 2001.

[27] N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Adv. NIPS*, 2012.

[28] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):pp. 1679–1706, 1994.

[29] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2003.

In the appendix we provide proofs of all three theorems, as well as additional experimental results (all training objectives, and two additional datasets *quantum* and *rcv1*).

Notations. Throughout this appendix material, we are going to use the notation $\|X\|_p = \big[\mathbb{E}(\|X\|^p)\big]^{1/p}$ for any random vector X and real number $p \geqslant 1$. By Minkowski's inequality, we have the triangle inequality $\|X+Y\|_p \leqslant \|X\|_p + \|Y\|_p$ whenever the expression makes sense.

A Proof of Theorem 1

We first denote by $\eta_n = \theta_n - \theta_* \in \mathcal{H}$ the deviation to θ_*. Since we consider quadratic functions, it satisfies a simplified recursion:

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma(x_n \otimes x_n)\theta_n + \gamma\xi_n \\ &= \big(I - \gamma x_n \otimes x_n\big)\eta_{n-1} + \gamma\xi_n \end{aligned} \qquad (12)$$

We also consider $\bar\eta_n = \frac{1}{n+1}\sum_{k=0}^n \eta_k = \bar\theta_n - \theta_*$ the averaged iterate. We have $f(\theta_n) - f(\theta_*) = \frac{1}{2}\langle \eta_n, H\eta_n\rangle$ and $f(\bar\theta_n) - f(\theta_*) = \frac{1}{2}\langle \bar\eta_n, H\eta_n\rangle$.

The crux of the proof is to consider the same recursion as Eq. (12), but replacing $x_n \otimes x_n$ by its expectation H (which is related to fixed design analysis in linear regression). This is of course only an approximation, and thus one has to study the remainder term; it happens to satisfy a similar recursion, on which we can apply the same technique, and so on. This proof technique is taken from [14]. Here we push it to arbitrary orders with explicit constants for *averaged* constant-step-size stochastic gradient descent.

Consequences of assumptions. Note that Assumption **(A6)** implies that $\mathbb{E}\|x_n\|^2 \leqslant R^2$ (indeed, taking the trace of $\mathbb{E}\big(\|x_n\|^2 x_n \otimes x_n\big) \preccurlyeq R^2 H$, we get $\mathbb{E}\|x_n\|^4 \leqslant R^2\mathbb{E}\|x_n\|^2$, and we always have by Cauchy-Schwarz inequality, $\mathbb{E}\|x_n\|^2 \leqslant \sqrt{\mathbb{E}\|x_n\|^4} \leqslant R\sqrt{\mathbb{E}\|x_n\|^2}$). This then implies that $\operatorname{tr} H \leqslant R^2$ and thus $H \preccurlyeq (\operatorname{tr} H)I \preccurlyeq R^2 I$. Thus, whenever $\gamma \leqslant 1/R^2$, we have $\gamma H \preccurlyeq I$, for the order between positive definite matrices.

We denote by $\mathcal{F}_n$ the σ-algebra generated by $(x_1, z_1, \dots, x_n, z_n)$. Both θ_n and $\bar\theta_n$ are $\mathcal{F}_n$-measurable.

A.1 Two main lemmas

The proof relies on two lemmas, one that provides a weak result essentially equivalent (but more specific and simpler because the step-size is constant) to non-strongly-convex results from [6], and one that replaces $x_n \otimes x_n$ by its expectation H in Eq. (12), which may then be seen as a non-asymptotic counterpart to the similar set-tup in [2].

Lemma 1. *Assume $(x_n, \xi_n) \in \mathcal{H} \times \mathcal{H}$ are $\mathcal{F}_n$-measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$, $n \geqslant 1$. Assume that $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$, $\mathbb{E}\big[\|\xi_n\|^2|\mathcal{F}_{n-1}\big]$ is finite and $\mathbb{E}\big[\big(\|x_n\|^2 x_n \otimes x_n\big)|\mathcal{F}_{n-1}\big] \preccurlyeq R^2 H$, with $\mathbb{E}\big[x_n \otimes x_n\big|\mathcal{F}_{n-1}\big] = H$ for all $n \geqslant 1$, for some $R > 0$ and invertible operator H. Consider the recursion $\alpha_n = \big(I - \gamma x_n \otimes x_n\big)\alpha_{n-1} + \gamma\xi_n$, with $\gamma R^2 \leqslant 1$. Then:*

$$(1-\gamma R^2)\mathbb{E}\big[\langle\bar\alpha_{n-1}, H\bar\alpha_{n-1}\rangle\big] + \frac{1}{2n\gamma}\mathbb{E}\|\alpha_n\|^2 \leqslant \frac{1}{2n\gamma}\|\alpha_0\|^2 + \frac{\gamma}{n}\sum_{k=1}^n \mathbb{E}\|\xi_k\|^2$$

Proof. We follow the proof technique of [6] (which relies only on smoothness) and get:

$$\begin{aligned}\|\alpha_n\|^2 &= \|\alpha_{n-1}\|^2 + \gamma^2\|\xi_n - (x_n\otimes x_n)\alpha_{n-1}\|^2 + 2\gamma\langle\alpha_{n-1}, \xi_n - (x_n\otimes x_n)\alpha_{n-1}\rangle\\ &\leqslant \|\alpha_{n-1}\|^2 + \Big\{2\gamma^2\|\xi_n\|^2 + 2\gamma^2\|(x_n\otimes x_n)\alpha_{n-1}\|^2\Big\} + 2\gamma\langle\alpha_{n-1}, \xi_n - (x_n\otimes x_n)\alpha_{n-1}\rangle\end{aligned}$$

By taking expectations, we obtain:

$$\begin{aligned}\mathbb{E}\big[\|\alpha_n\|^2|\mathcal{F}_{n-1}\big] &\leqslant \|\alpha_{n-1}\|^2 + 2\gamma^2\|\xi_n\|^2 + 2\gamma^2\langle\alpha_{n-1}, \mathbb{E}\big[\|x_n\|^2 x_n\otimes x_n\big]\alpha_{n-1}\rangle - 2\gamma\langle\alpha_{n-1}, H\alpha_{n-1}\rangle\\ &\leqslant \|\alpha_{n-1}\|^2 + 2\gamma^2\|\xi_n\|^2 + 2\gamma^2R^2\langle\alpha_{n-1}, H\alpha_{n-1}\rangle - 2\gamma\langle\alpha_{n-1}, H\alpha_{n-1}\rangle\\ &= \|\alpha_{n-1}\|^2 + 2\gamma^2\|\xi_n\|^2 + 2\gamma^2R^2\langle\alpha_{n-1}, H\alpha_{n-1}\rangle - 2\gamma\langle\alpha_{n-1}, H\alpha_{n-1}\rangle\\ &\leqslant \|\alpha_{n-1}\|^2 + 2\gamma^2\|\xi_n\|^2 - 2\gamma(1-\gamma R^2)\langle\alpha_{n-1}, H\alpha_{n-1}\rangle\end{aligned}$$

By taking another expectation, we get

$$\mathbb{E}\|\alpha_n\|^2 \leqslant \mathbb{E}\|\alpha_{n-1}\|^2 + 2\gamma^2\mathbb{E}\|\xi_n\|^2 - 2\gamma(1-\gamma R^2)\mathbb{E}\langle\alpha_{n-1}, H\alpha_{n-1}\rangle$$

This leads to the desired result, because, by convexity, $\langle\overline{\alpha}_{n-1}, H\overline{\alpha}_{n-1}\rangle \leqslant \frac{1}{n}\sum_{k=0}^{n-1}\langle\alpha_k, H\alpha_k\rangle$. □

Lemma 2. *Assume $\xi_n \in \mathcal{H}$ is $\mathcal{F}_n$-measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$, $n \geqslant 1$. Assume $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$, $\mathbb{E}\big[\|\xi_n\|^2\big]$ is finite, and for all $n \geqslant 1$, $\mathbb{E}\big[\xi_n\otimes\xi_n\big] \preccurlyeq C$. Consider the recursion $\alpha_n = \big(I-\gamma H\big)\alpha_{n-1} + \gamma\xi_n$, with $\gamma H \preccurlyeq I$ for some invertible H. Then:*

$$\mathbb{E}[\alpha_n\otimes\alpha_n] = (I-\gamma H)^n\alpha_0\otimes\alpha_0(I-\gamma H)^n + \gamma^2\sum_{k=1}^{n}(I-\gamma H)^{n-k}C(I-\gamma H)^{n-k} \tag{13}$$

$$\mathbb{E}\big[\langle\overline{\alpha}_{n-1}, H\overline{\alpha}_{n-1}\rangle\big] \leqslant \frac{1}{n\gamma}\|\alpha_0\|^2 + \frac{\operatorname{tr} CH^{-1}}{n} \tag{14}$$

Proof. The proof relies on the fact that cost functions are quadratic and our recursions are thus linear, allowing to obtain α_n in closed form. The sequence (α_n) satisfies a linear recursion, from which we get, for all $n \geqslant 1$:

$$\alpha_n = (I-\gamma H)^n\alpha_0 + \gamma\sum_{k=1}^{n}(I-\gamma H)^{n-k}\xi_k$$

which leads to the first result using classical martingale second moment expansions (which amount to considering ξ_i, $i = 1,\dots,n$ independent, so that the variance of the sum is the sum of variances). Moreover, using the identity $\sum_{k=0}^{n-1}(I-\gamma H)^k = \big(I-(I-\gamma H)^n\big)\big(\gamma H\big)^{-1}$, we get:

$$\begin{aligned}\overline{\alpha}_{n-1} &= \frac{1}{n}\sum_{k=0}^{n-1}(I-\gamma H)^k\alpha_0 + \frac{\gamma}{n}\sum_{k=1}^{n-1}\sum_{j=1}^{k}(I-\gamma H)^{k-j}\xi_j\\ &= \frac{1}{n}\big(I-(I-\gamma H)^n\big)\big(\gamma H\big)^{-1}\alpha_0 + \frac{\gamma}{n}\sum_{k=1}^{n-1}\sum_{j=1}^{k}(I-\gamma H)^{k-j}\xi_j\\ &= \frac{1}{n}\big(I-(I-\gamma H)^n\big)\big(\gamma H\big)^{-1}\alpha_0 + \frac{\gamma}{n}\sum_{j=1}^{n-1}\bigg(\sum_{k=j}^{n-1}(I-\gamma H)^{k-j}\bigg)\xi_j\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\big(I-(I-\gamma H)^n\big)\big(\gamma H\big)^{-1}\alpha_0 + \frac{\gamma}{n}\sum_{j=1}^{n-1}\bigg(\sum_{k=0}^{n-1-j}(I-\gamma H)^k\bigg)\xi_j \\
&= \frac{1}{n}\big(I-(I-\gamma H)^n\big)\big(\gamma H\big)^{-1}\alpha_0 + \frac{\gamma}{n}\sum_{j=1}^{n-1}\big(I-(I-\gamma H)^{n-j}\big)\big(\gamma H\big)^{-1}\xi_j
\end{aligned}
$$

We then get, using standard martingale square moment inequalities (which here also amount to considering ξ_i, $i=1,\dots,n$ independent, so that the variance of the sum is the sum of variances):

$$
\begin{aligned}
\mathbb{E}\langle\overline{\alpha}_{n-1}, H\overline{\alpha}_{n-1}\rangle &= \frac{1}{n\gamma}\langle\alpha_0, \big[I-(I-\gamma H)^n\big]^2\big(n\gamma H\big)^{-1}\alpha_0\rangle \\
&\qquad\qquad + \frac{1}{n^2}\sum_{j=1}^{n-1}\operatorname{tr}\big(I-(I-\gamma H)^{n-j}\big)^2 H^{-1}C \\
&\leqslant \frac{1}{n\gamma}\|\alpha_0\|^2 + \frac{1}{n}\operatorname{tr}H^{-1}C
\end{aligned}
$$

because for all $u\in[0,1]$, $\frac{(1-(1-u)^n)^2}{nu}\leqslant 1$ (see Lemma 3 in Section A.6), and the second term is the sum of terms which are all less than $\operatorname{tr}H^{-1}C$.

Note that we may replace the term $\frac{1}{n\gamma}\|\alpha_0\|^2$ by $\dfrac{1}{n^2\gamma^2}\langle\alpha_0, H^{-1}\alpha_0\rangle$, which is only interesting when $\langle\alpha_0, H^{-1}\alpha_0\rangle$ is small. □

A.2 Proof principle

The proof relies on an expansion of η_n and $\overline{\eta}_{n-1}$ as polynomials in γ due to [14]. This expansion is done separately for the noise process (i.e., when assuming $\eta_0=0$) and for the noise-free process that depends only on the initial conditions (i.e., when assuming that $\sigma=0$). The bounds may then be added.

Indeed, we have $\eta_n = M_1^n\eta_0 + \gamma\sum_{k=1}^n M_{k+1}^n\xi_k$, with $M_i^j = (I-\gamma x_j\otimes x_j)\cdots(I-\gamma x_i\otimes x_i)$ and $M_i^{i-1}=I$, and thus $\overline{\eta}_n = \frac{1}{n+1}\sum_{i=0}^n\Big[M_1^i\eta_0 + \gamma\sum_{k=1}^i M_{k+1}^i\xi_k\Big] = \frac{1}{n+1}\sum_{i=0}^n M_1^i\eta_0 + \gamma\sum_{k=1}^n\bigg(\sum_{i=k}^n M_{k+1}^i\bigg)\xi_k$, leading to

$$
\|H^{1/2}\overline{\eta}_n\|_p \leqslant \bigg\|\frac{1}{n+1}\sum_{i=0}^n M_1^j\eta_0\bigg\|_p + \bigg\|\gamma\sum_{k=1}^n\bigg(\sum_{i=k}^n M_{k+1}^i\bigg)\xi_k\bigg\|_p
$$

for any $p\geqslant 2$ for which it is defined: the left term depends only on initial conditions and the right term depends only on the noise process (note the similarity with bias-variance decompositions).

A.3 Initial conditions

In this section, we assume that ξ_n is uniformly equal to zero, and that $\gamma R^2\leqslant 1$.

We thus have $\eta_n = (I-\gamma x_n\otimes x_n)\eta_{n-1}$ and thus

$$
\|\eta_n\|^2 = \|\eta_{n-1}\|^2 - 2\gamma\langle\eta_{n-1}, (x_n\otimes x_n)\eta_{n-1}\rangle + \gamma^2\langle\eta_{n-1}, (x_n\otimes x_n)^2\eta_{n-1}\rangle
$$

By taking expectations (first given $\mathcal{F}_{n-1}$, then unconditionally), we get:

$$\begin{aligned}\mathbb{E}\|\eta_n\|^2 &\leqslant \mathbb{E}\|\eta_{n-1}\|^2 - 2\gamma\mathbb{E}\langle\eta_{n-1}, H\eta_{n-1}\rangle + \gamma^2 R^2\mathbb{E}\langle\eta_{n-1}, H\eta_{n-1}\rangle \text{ using } \mathbb{E}\|x_n\|^2 x_n\otimes x_n \preccurlyeq R^2 H,\\ &\leqslant \mathbb{E}\|\eta_{n-1}\|^2 - \gamma\mathbb{E}\langle\eta_{n-1}, H\eta_{n-1}\rangle \text{ using } \gamma R^2 \leqslant 1\end{aligned}$$

from which we obtain, by summing from 1 to n and using convexity (note that Lemma 1 could be used directly as well):

$$\mathbb{E}\langle\overline{\eta}_{n-1}, H\overline{\eta}_{n-1}\rangle \leqslant \frac{\|\eta_0\|^2}{n\gamma}$$

Here, it would be interesting to explore conditions under which the initial conditions may be forgotten at a rate $O(1/n^2)$, as obtained by [6] in the strongly convex case.

A.4 Noise process

In this section, we assume that $\eta_0 = \theta_0 - \theta_* = 0$ and $\gamma R^2 \leqslant 1$ (which implies $\gamma H \preccurlyeq I$). Following [14], we recursively define the sequences $(\eta_n^r)_{n\geqslant 0}$ for $r \geqslant 0$ (and their averaged counterparts $\overline{\eta}_n^r$):

- The sequence (η_n^0) is defined as $\eta_0^0 = \eta_0 = 0$ and for $n \geqslant 1$, $\eta_n^0 = (I-\gamma H)\eta_{n-1}^0 + \gamma\xi_n$.
- The sequence (η_n^r) is defined from (η_n^{r-1}) as $\eta_0^r = 0$ and, for all $n \geqslant 1$:

$$\eta_n^r = (I-\gamma H)\eta_{n-1}^r + \gamma(H - x_n\otimes x_n)\eta_{n-1}^{r-1} \qquad (15)$$

Recursion for expansion. We now show that the sequence $\eta_n - \sum_{i=0}^r \eta_n^i$ then satisfies the following recursion, for any $r \geqslant 0$ (which is of the same type than (η_n)):

$$\eta_n - \sum_{i=0}^r \eta_n^i = (I - \gamma x_n\otimes x_n)\Big(\eta_{n-1} - \sum_{i=0}^r \eta_{n-1}^r\Big) + \gamma(H - x_n\otimes x_n)\eta_{n-1}^r \qquad (16)$$

In order to prove Eq. (16) by recursion, we have, for $r = 0$,

$$\begin{aligned}\eta_n - \eta_n^0 &= (I-\gamma x_n\otimes x_n)\eta_{n-1} - (I-\gamma H)\eta_{n-1}^0\\ &= (I-\gamma x_n\otimes x_n)(\eta_{n-1} - \eta_{n-1}^0) + \gamma(H - x_n\otimes x_n)\eta_{n-1}^0,\end{aligned}$$

and, to go from r to $r+1$:

$$\begin{aligned}\eta_n - \sum_{i=0}^{r+1}\eta_n^i &= (I-\gamma x_n\otimes x_n)\Big(\eta_{n-1} - \sum_{i=0}^r \eta_{n-1}^i\Big) + \gamma(H - x_n\otimes x_n)\eta_{n-1}^r\\ &\qquad -(I-\gamma H)\eta_{n-1}^{r+1} - \gamma(H - x_n\otimes x_n)\eta_{n-1}^r\\ &= (I-\gamma x_n\otimes x_n)\Big(\eta_{n-1} - \sum_{i=0}^{r+1}\eta_{n-1}^i\Big) + \gamma(H - x_n\otimes x_n)\eta_{n-1}^{r+1}\end{aligned}$$

Bound on covariance operators. We now show that we also have a bound on the covariance operator of η_{n-1}^r, for any $r \geqslant 0$ and $n \geqslant 2$:

$$\mathbb{E}\big[\eta_{n-1}^r \otimes \eta_{n-1}^r\big] \preccurlyeq \gamma^{r+1}R^{2r}\sigma^2 I \qquad (17)$$

In order to prove Eq. (17) by recursion, we get for $r = 0$:

$$\begin{aligned}
\mathbb{E}\big[\eta_{n-1}^0 \otimes \eta_{n-1}^0\big] &\preccurlyeq \gamma^2\sigma^2 \sum_{k=1}^{n-1} (I-\gamma H)^{2n-2-2k} H \\
&\preccurlyeq \gamma^2\sigma^2 \big(I - (I-\gamma H)^{2n-2}\big)\big(I - (I-\gamma H)^2\big)^{-1} H \\
&= \gamma^2\sigma^2 \big(I - (I-\gamma H)^{2n-2}\big)\big(2\gamma H - \gamma^2 H^2\big)^{-1} H \\
&\preccurlyeq \gamma^2\sigma^2 \big(I - (I-\gamma H)^{2n-2}\big)\big(\gamma H\big)^{-1} H \preccurlyeq \gamma\sigma^2 I
\end{aligned}$$

In order to go from r to $r+1$, we have, using Lemma 2 and the fact that η_{k-1}^r and x_k are independent:

$$\begin{aligned}
&\mathbb{E}\big[\eta_{n-1}^{r+1} \otimes \eta_{n-1}^{r+1}\big] \\
\preccurlyeq\ & \gamma^2 \mathbb{E}\Big[\sum_{k=1}^{n-1} (I-\gamma H)^{n-1-k} (H - x_k\otimes x_k) \mathbb{E}\big[\eta_{k-1}^r \otimes \eta_{k-1}^r\big] (H - x_k \otimes x_k)(I-\gamma H)^{n-1-k}\Big] \\
\preccurlyeq\ & \gamma^{r+3} R^{2r} \sigma^2 \mathbb{E}\Big[\sum_{k=1}^{n-1} (I-\gamma H)^{n-1-k} (H - x_k\otimes x_k)^2 (I-\gamma H)^{n-1-k}\Big] \text{ using the result for } r, \\
\preccurlyeq\ & \gamma^{r+3} R^{2r+2}\sigma^2 \sum_{k=1}^{n-1} (I-\gamma H)^{2n-2-2k} H \text{ using } \mathbb{E}(x_k\otimes x_k - H)^2 \preccurlyeq \mathbb{E}\|x_k\|^2 x_k \otimes x_k \preccurlyeq R^2 H, \\
\preccurlyeq\ & \gamma^{r+2} R^{2r+2} \sigma^2 I
\end{aligned}$$

Putting things together. We may apply Lemma 1 to the sequence $\big(\eta_n - \sum_{i=0}^r \eta_n^i\big)$, to get

$$\begin{aligned}
\mathbb{E}\Big\langle \overline{\eta}_{n-1} - \sum_{i=0}^r \overline{\eta}_{n-1}^i, H\big(\overline{\eta}_{n-1} - \sum_{i=0}^r \overline{\eta}_{n-1}^i\big)\Big\rangle &\leqslant \frac{1}{1-\gamma R^2} \frac{\gamma}{n} \sum_{k=2}^n \mathbb{E}\|(H - x_k \otimes x_k)\eta_{k-1}^r\|^2 \\
&\leqslant \frac{1}{1-\gamma R^2} \gamma^{r+2}\sigma^2 R^{2r+4}
\end{aligned}$$

We may now apply Lemma 2 to Eq. (15), to get, with a noise process $\xi_n^r = (H - x_n \otimes x_n)\eta_{n-1}^{r-1}$ which is such that

$$\mathbb{E}\big[\xi_n^r \otimes \xi_n^r\big] \preccurlyeq \gamma^r R^{2r} \sigma^2 H$$

$$\mathbb{E}\langle \overline{\eta}_{n-1}^r, H\overline{\eta}_{n-1}^r\rangle \leqslant \frac{1}{n}\gamma^r R^{2r} d\sigma^2$$

We thus get, using Minkowski's inequality (i.e., triangle inequality for the norms $\|\cdot\|_p$):

$$\begin{aligned}
\big(\mathbb{E}\langle \overline{\eta}_{n-1}, H\overline{\eta}_{n-1}\rangle\big)^{1/2} &\leqslant \big(\frac{1}{1-\gamma R^2}\gamma^{r+2}\sigma^2 R^{2r+4}\big)^{1/2} + \frac{\sigma\sqrt{d}}{\sqrt{n}} \sum_{i=0}^r \gamma^{i/2} R^i \\
&\leqslant \big(\frac{1}{1-\gamma R^2}\gamma^{r+2}\sigma^2 R^{2r+4}\big)^{1/2} + \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1 - (\sqrt{\gamma R^2})^{r+1}}{1 - \sqrt{\gamma R^2}}
\end{aligned}$$

This implies that for any $\gamma R^2 < 1$, we obtain, by letting r tend to $+\infty$:

$$\big(\mathbb{E}\langle \overline{\eta}_{n-1}, H\overline{\eta}_{n-1}\rangle\big)^{1/2} \leqslant \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1}{1-\sqrt{\gamma R^2}}$$

A.5 Final bound

We can now take results from Appendices A.3 and A.4, to get

$$\big(\mathbb{E}\langle\overline{\eta}_{n-1},H\overline{\eta}_{n-1}\rangle\big)^{1/2}\leqslant\frac{\sigma\sqrt{d}}{\sqrt{n}}\frac{1}{1-\sqrt{\gamma R^2}}+\frac{\|\eta_0\|^2}{n\gamma}$$

which leads to the desired result.

A.6 Proof of Lemma 3

In this section, we state and prove a simple lemma.

Lemma 3. *For any $u\in[0,1]$ and $n>0$, $(1-(1-u)^n)^2\leqslant nu$.*

Proof. Since $u\in[0,1]$, we have, $1-(1-u)^n\leqslant 1$. Moreover, $n(1-u)^{n-1}\leqslant n$, and by integrating between 0 and u, we get $1-(1-u)^n\leqslant nu$. By multiplying the two previous inequalities, we get the desired result. □

B Proof of Theorem 2

Throughout the proof, we use the notation for $X\in\mathcal{H}$ a random vector, and p any *real* number greater than 1, $\|X\|_p=\big(\mathbb{E}\|X\|^p\big)^{1/p}$. We first recall the Burkholder-Rosenthal-Pinelis (BRP) inequality [28, Theorem 4.1]. Let $p\in\mathbb{R}$, $p\geqslant 2$ and $(\mathcal{F}_n)_{n\geqslant 0}$ be a sequence of increasing σ-fields, and $(x_n)_{n\geqslant 1}$ an adapted sequence of elements of $\mathcal{H}$, such that $\mathbb{E}\big[x_n|\mathcal{F}_{n-1}\big]=0$, and $\|x_n\|_p$ is finite. Then,

$$\begin{aligned}\bigg\|\sup_{k\in\{1,\dots,n\}}\bigg\|\sum_{j=1}^k x_j\bigg\|\bigg\|_p &\leqslant \sqrt{p}\bigg\|\sum_{k=1}^n\mathbb{E}\big[\|x_k\|^2|\mathcal{F}_{k-1}\big]\bigg\|_{p/2}^{1/2}+p\bigg\|\sup_{k\in\{1,\dots,n\}}\|x_k\|\bigg\|_p \qquad (18)\\ &\leqslant \sqrt{p}\bigg\|\sum_{k=1}^n\mathbb{E}\big[\|x_k\|^2|\mathcal{F}_{k-1}\big]\bigg\|_{p/2}^{1/2}+p\bigg\|\sup_{k\in\{1,\dots,n\}}\|x_k\|^2\bigg\|_{p/2}^{1/2}\end{aligned}$$

We use the same notations than the proof of Theorem 1, and the same proof principle: (a) splitting the contributions of the initial conditions and the noise, (b) providing a direct argument for the initial condition, and (c) performing an expansion for the noise contribution.

Consequences of assumptions. Note that by Cauchy-Schwarz inequality, assumption **(A7)** implies for all $z,t\in\mathcal{H}$, $\mathbb{E}\langle z,x_n\rangle^2\langle t,x_n\rangle^2\leqslant\kappa\langle z,Hz\rangle\langle t,Ht\rangle$. It in turn implies that for all positive semi-definite self-adjoint operators M,N, $\mathbb{E}\langle x_n,Mx_n\rangle\langle x_n,Nx_n\rangle\leqslant\kappa\,\mathrm{tr}(MH)\,\mathrm{tr}(NH)$.

B.1 Contribution of initial conditions

When the noise is assumed to be zero, we have $\eta_n=(I-\gamma x_n\otimes x_n)\eta_{n-1}$ almost surely, and thus, since $0\preccurlyeq\gamma x_n\otimes x_n\preccurlyeq I$, $\|\eta_n\|\leqslant\|\eta_0\|$ almost surely, and

$$\begin{aligned}\|\eta_n\|^2 &= \|\eta_{n-1}\|^2-2\gamma\langle\eta_{n-1},(x_n\otimes x_n)\eta_{n-1}\rangle+\gamma^2\langle\eta_{n-1},(x_n\otimes x_n)^2\eta_{n-1}\rangle\\ &\leqslant \|\eta_{n-1}\|^2-2\gamma\langle\eta_{n-1},(x_n\otimes x_n)\eta_{n-1}\rangle+\gamma\langle\eta_{n-1},(x_n\otimes x_n)\eta_{n-1}\rangle\end{aligned}$$

$$\text{using } \|x_n\|^2 \leqslant R^2 \text{ and } \gamma R^2 \leqslant 1,$$

$$= \quad \|\eta_{n-1}\|^2 - \gamma\langle \eta_{n-1}, (x_n \otimes x_n)\eta_{n-1}\rangle$$

which we may write as

$$\|\eta_n\|^2 - \|\eta_{n-1}\|^2 + \gamma\langle \eta_{n-1}, H\eta_{n-1}\rangle \leqslant \gamma\langle \eta_{n-1}, (H - x_n \otimes x_n)\eta_{n-1}\rangle \overset{\text{def}}{=} M_n$$

We thus have:

$$A_n \overset{\text{def}}{=} \|\eta_n\|^2 + \gamma \sum_{k=1}^{n} \langle \eta_{k-1}, H\eta_{k-1}\rangle \leqslant \|\eta_0\|^2 + \sum_{k=1}^{n} M_k$$

Note that we have

$$\mathbb{E}[M_n^2|\mathcal{F}_{n-1}] \quad \leqslant \quad \mathbb{E}\big[\gamma^2\langle \eta_{n-1}, (x_n \otimes x_n)\eta_{n-1}\rangle^2|\mathcal{F}_{n-1}\big] \leqslant \gamma^2 R^2 \|\eta_0\|^2 \langle \eta_{n-1}, H\eta_{n-1}\rangle$$

and $|M_n| \leqslant \gamma\|\eta_0\|^2 R^2$. We may now apply the Burkholder-Rosenthal-Pinelis inequality in Eq. (18), to get:

$$\begin{aligned}
\big\|A_n\big\|_p &\leqslant \|\eta_0\|^2 + \sqrt{p}\bigg\|\sum_{k=1}^{n} \mathbb{E}\big[M_k^2|\mathcal{F}_{k-1}\big]\bigg\|_{p/2}^{1/2} + p\bigg\|\sup_{k\in\{1,\dots,n\}} |M_k|\bigg\|_p \\
&\leqslant \|\eta_0\|^2 + \gamma\sqrt{p}\bigg\|\|\eta_0\|^2 R^2 \sum_{k=1}^{n} \langle \eta_{k-1}, H\eta_{k-1}\rangle\bigg\|_{p/2}^{1/2} + p\gamma R^2\|\eta_0\|^2 \\
&\leqslant \|\eta_0\|^2 + \gamma^{1/2} R\sqrt{p}\|\eta_0\|\big\|A_n\big\|_{p/2}^{1/2} + p\gamma R^2\|\eta_0\|^2 \\
&\leqslant \|\eta_0\|^2 + \gamma^{1/2} R\sqrt{p}\|\eta_0\|\big\|A_n\big\|_{p}^{1/2} + p\gamma R^2\|\eta_0\|^2
\end{aligned}$$

We have used above that (a) $\sum_{k=1}^{n} \langle \eta_{k-1}, H\eta_{k-1}\rangle \leqslant \frac{A_n}{\gamma}$ and that (b) $\big\|A_n\big\|_{p/2} \leqslant \big\|A_n\big\|_p$. This leads to

$$\Big(\big\|A_n\big\|_p^{1/2} - \frac{1}{2}\gamma^{1/2} R\sqrt{p}\|\eta_0\|\Big)^2 \leqslant \|\eta_0\|^2 + \frac{5p}{4}\gamma R^2\|\eta_0\|^2$$

which leads to

$$\big\|A_n\big\|_p^{1/2} - \frac{1}{2}\gamma^{1/2} R\sqrt{p}\|\eta_0\| = \|\eta_0\|\sqrt{1 + \frac{5p\gamma R^2}{4}}$$

$$\big\|A_n\big\|_p \quad \leqslant \quad \|\eta_0\|^2\bigg(2 + \frac{5p\gamma R^2}{2} + \frac{p\gamma R^2}{2}\bigg) \leqslant \|\eta_0\|^2(2 + 3p\gamma R^2)$$

Finally, we obtain, for any $p \geqslant 2$

$$\big\|\langle \overline{\eta}_{n-1}, H\overline{\eta}_{n-1}\rangle\big\|_p \leqslant \frac{\|\eta_0\|^2}{n\gamma}(2 + 3p\gamma R^2)$$

i.e., by a change of variable $p \to \frac{p}{2}$, for any $p \geqslant 4$, we get

$$\big\|H^{1/2}\overline{\eta}_{n-1}\big\|_p \leqslant \big\|\langle \overline{\eta}_{n-1}, H\overline{\eta}_{n-1}\rangle\big\|_{p/2}^{1/2} = \frac{\|\eta_0\|}{\sqrt{n\gamma}}\sqrt{2 + \frac{3p}{2}\gamma R^2}$$

By using monotonicity of norms, we get, for any $p \in [2, 4]$:

$$\big\|H^{1/2}\overline{\eta}_{n-1}\big\|_p \leqslant \big\|H^{1/2}\overline{\eta}_{n-1}\big\|_4 \leqslant \frac{\|\eta_0\|}{\sqrt{n\gamma}}\sqrt{2 + 6\gamma R^2} \leqslant \frac{\|\eta_0\|}{\sqrt{n\gamma}}\sqrt{2 + 3p\gamma R^2}$$

which is also valid for $p > 4$.

Note that the constants in the bound above could be improved by using a proof by recursion.

B.2 Contribution of the noise

We follow the same proof technique than for Theorem 1 and consider the expansion based on the sequences $(\eta_n^r)_n$, for $r \geqslant 0$. We need (a) bounds on η_n^0, (b) a recursion on the magnitude (in $\|\cdot\|_p$ norm) of η_n^r and (c) a control of the error made in the expansions.

Bound on $\overline{\eta}_n^0$. We start by a lemma similar to Lemma 2 but for all moments. This will show a bound for the sequence $\overline{\eta}_n^0$.

Lemma 4. *Assume $\xi_n \in \mathcal{H}$ is $\mathcal{F}_n$-measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$, $n \geqslant 1$. Assume $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$, $\mathbb{E}\big[\|\xi_n\|^2|\mathcal{F}_{n-1}\big]$ is finite. Assume moreover that for all $n \geqslant 1$, $\mathbb{E}\big[\xi_n \otimes \xi_n|\mathcal{F}_{n-1}\big] \preccurlyeq C$ and $\|\xi_n\|_p \leqslant \tau R$ almost surely for some $p \geqslant 2$.*

Consider the recursion $\alpha_n = \big(I - \gamma H\big)\alpha_{n-1} + \gamma\xi_n$, with $\alpha_0 = 0$ and $\gamma H \preccurlyeq I$. Let $p \in \mathbb{R}$, $p \geqslant 2$. Then:

$$\|H^{1/2}\overline{\alpha}_{n-1}\|_p \leqslant \frac{\sqrt{p}}{\sqrt{n}}\sqrt{\operatorname{tr} CH^{-1}} + \frac{\sqrt{\gamma}pR\tau}{\sqrt{n}} \tag{19}$$

Proof. We have, from the proof of Lemma 2:

$$\begin{aligned}
\overline{\alpha}_{n-1} &= \frac{\gamma}{n}\sum_{j=1}^{n-1}\big(I-(I-\gamma H)^{n-j}\big)\big(\gamma H\big)^{-1}\xi_j \\
\|H^{1/2}\overline{\alpha}_{n-1}\|_p &\leqslant \frac{\gamma}{n}\bigg\|\sum_{j=1}^{n-1}\big(I-(I-\gamma H)^{n-j}\big)\big(\gamma H\big)^{-1}H^{1/2}\xi_j\bigg\|_p \\
&\leqslant \frac{\gamma}{n}\bigg\|\sum_{j=1}^{n-1}\beta_j\bigg\|_p
\end{aligned}$$

with $\beta_j = \big(I-(I-\gamma H)^{n-j}\big)\big(\gamma H\big)^{-1}H^{1/2}\xi_j$. We have

$$\begin{aligned}
\sum_{j=1}^{n-1}\mathbb{E}[\|\beta_j\|^2|\mathcal{F}_{j-1}] &= \sum_{j=1}^{n-1}\operatorname{tr}\mathbb{E}\big[\xi_j\otimes\xi_j|\mathcal{F}_{j-1}\big]H\bigg(\frac{I-(I-\gamma H)^{n-j}}{\gamma H}\bigg)^2 \\
&\leqslant \gamma^{-2}\sum_{j=1}^{n-1}\mathbb{E}\big[\langle\xi_j, H^{-1}\xi_j\rangle|\mathcal{F}_{j-1}\big] \leqslant n\gamma^{-2}\operatorname{tr} CH^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\|\beta_j\|_p &\leqslant \lambda_{\max}\big[\big(I-(I-\gamma H)^{n-j}\big)\big(\gamma H\big)^{-1}H^{1/2}\big]\|\xi_j\|_p \\
&\leqslant \gamma^{-1/2}\|\xi_j\|_p\max_{u\in(0,1]}\frac{1-(1-u)^{n-j}}{u^{1/2}} \\
&\leqslant \frac{\sqrt{n-j}}{\sqrt{\gamma}}\|\xi_j\|_p \leqslant \frac{\tau R\sqrt{n}}{\sqrt{\gamma}}
\end{aligned}$$

using Lemma 3 in Section A.6, and assumption **(A7)**.

Using Burkholder-Rosenthal-Pinelis inequality in Eq. (18), we then obtain

$$\bigg\|\sup_{k\in\{1,\dots,n-1\}}\|H^{1/2}\overline{\alpha}_k\|\bigg\|_p \leqslant \frac{\sqrt{p}}{\sqrt{n}}\sqrt{\operatorname{tr} CH^{-1}} + p\frac{\sqrt{\gamma}}{\sqrt{n}}\sigma R$$

leading to the desired result. □

Bounds on η_n^0. Following the same proof technique as above, we have

$$\eta_n^0 = \gamma \sum_{j=1}^n (I-\gamma H)^{n-j} \xi_j$$

from which we get, for any positive semidefinite operator M such that $\operatorname{tr} M = 1$, using BRP's inequality:

$$\begin{aligned}
\Big\| \sup_{k \in \{1,\dots,n\}} \|M^{1/2} \eta_k^0\| \Big\|_p &\leqslant \sqrt{p}\gamma\sigma \Big\| \sum_{j=1}^n \operatorname{tr} H (I-\gamma H)^{n-j} M (I-\gamma H)^{n-j} \Big\|_{p/2}^{1/2} + p\gamma\tau R \\
&\leqslant \sqrt{p}\gamma\sigma \Big\| \frac{1}{\gamma} \operatorname{tr} M \Big\|_{p/2}^{1/2} + p\gamma\tau R \\
&\leqslant \frac{1}{R} \sqrt{p\gamma R^2} (\sigma + \tau \sqrt{p\gamma R^2})
\end{aligned}$$

leading to

$$\sup_{\operatorname{tr} M = 1} \Big\| \sup_{k \in \{1,\dots,n\}} \|M^{1/2} \eta_k^0\| \Big\|_p \leqslant \frac{1}{R} \sqrt{p\gamma R^2} (\sigma + \tau \sqrt{p\gamma R^2}) \tag{20}$$

Recursion on bounds on η_n^r. We introduce the following quantity to control the deviations of η_n^r:

$$A_r = \sup_{\operatorname{tr} M = 1} \Big\| \sup_{k \in \{1,\dots,n\}} \|M^{1/2} \eta_k^r\| \Big\|_p$$

We have from Eq. (20), $A_0 \leqslant \frac{1}{R} \sqrt{p\gamma R^2} (\sigma + \tau \sqrt{p\gamma R^2})$.

Since $\eta_n^r = (I-\gamma H)\eta_{n-1}^r + \gamma (H - x_n \otimes x_n) \eta_{n-1}^{r-1}$, for all $n \geqslant 1$, we have the closed form expression

$$\eta_n^r = \gamma \sum_{k=2}^n (I-\gamma H)^{n-k} (H - x_k \otimes x_k) \eta_{k-1}^{r-1}$$

and we may use BRP's inequality in Eq. (18) to get, for any M such that $\operatorname{tr} M = 1$:

$$A_r \leqslant B + C$$

with

$$\begin{aligned}
B &= \sqrt{p}\gamma \Big\| \sum_{k=2}^n \langle \eta_{k-1}^{r-1}, \mathbb{E}\big[(H - x_k \otimes x_k)(I-\gamma H)^{n-k} M (I-\gamma H)^{n-k} (H - x_k \otimes x_k)\big] \eta_{k-1}^{r-1} \rangle \Big\|_{p/2}^{1/2} \\
&\leqslant \sqrt{p}\gamma \Big\| \sum_{k=2}^n \langle \eta_{k-1}^{r-1}, \mathbb{E}\big[(x_k \otimes x_k)(I-\gamma H)^{n-k} M (I-\gamma H)^{n-k} (x_k \otimes x_k)\big] \eta_{k-1}^{r-1} \rangle \Big\|_{p/2}^{1/2} \\
&\quad \text{using } \mathbb{E} \operatorname{tr} N (H - x_k \otimes x_k) M (H - x_k \otimes x_k) \leqslant \mathbb{E} \operatorname{tr} N (x_k \otimes x_k) H (x_k \otimes x_k), \\
&\leqslant \sqrt{p}\gamma \Big\| \sum_{k=2}^n \kappa \langle \eta_{k-1}^{r-1}, H \eta_{k-1}^{r-1} \rangle \operatorname{tr} H (I-\gamma H)^{n-k} M (I-\gamma H)^{n-k} \Big\|_{p/2}^{1/2} \\
&\quad \text{using the kurtosis property,} \\
&\leqslant \sqrt{p}\gamma \sqrt{\kappa} A_{r-1} \Big(\sum_{k=2}^n \operatorname{tr} H (I-\gamma H)^{n-k} M (I-\gamma H)^{n-k} \Big)^{1/2}
\end{aligned}$$

$$\begin{aligned}
& \quad \text{using } \langle \eta_{k-1}^{r-1}, H\eta_{k-1}^{r-1}\rangle \leqslant \sup_{k\in\{1,\dots,n\}} \langle \eta_{k-1}^{r-1}, H\eta_{k-1}^{r-1}\rangle, \\
&\leqslant \sqrt{p}\gamma\sqrt{\kappa}RA_{r-1}\Big(\frac{1}{\gamma}\operatorname{tr} M\Big)^{1/2} = \sqrt{p\gamma R^2\kappa}A_{r-1}
\end{aligned}$$

and

$$\begin{aligned}
C &= p\gamma\Big\| \sup_{k\in\{2,\dots,n\}} \langle \eta_{k-1}^{r-1}, (H - x_k\otimes x_k)(I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}(H - x_k\otimes x_k)\eta_{k-1}^{r-1}\rangle \Big\|_{p/2}^{1/2} \\
&\leqslant p\gamma\Big\| \sup_{k\in\{2,\dots,n\}} \langle \eta_{k-1}^{r-1}, (x_k\otimes x_k)(I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}(x_k\otimes x_k)\eta_{k-1}^{r-1}\rangle \Big\|_{p/2}^{1/2} \\
&\quad + p\gamma\Big\| \sup_{k\in\{2,\dots,n\}} \langle \eta_{k-1}^{r-1}, H(I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}H\eta_{k-1}^{r-1}\rangle \Big\|_{p/2}^{1/2} \\
&\quad \text{using Minkowski's inequality,} \\
&\leqslant +p\gamma\Big(\sum_{k=2}^n \mathbb{E}\big[\langle \eta_{k-1}^{r-1}, (x_k\otimes x_k)(I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}(x_k\otimes x_k)\eta_{k-1}^{r-1}\rangle^{p/2}\big]\Big)^{1/p} \\
&\quad + p\gamma R^2 A_{r-1} \\
&\quad \text{bounding the supremum by a sum, and using } H(I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}H \preccurlyeq H^2, \\
&\leqslant p\gamma\Big(\sum_{k=2}^n \mathbb{E}\big[\langle \eta_{k-1}^{r-1}, x_k\rangle^p \langle x_k, (I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}x_k\rangle^{p/2}\big]\Big)^{1/p} + p\gamma R^2 A_{r-1} \\
&\leqslant p\gamma R A_{r-1}\Big(\sum_{k=2}^n \mathbb{E}\big[\langle x_k, (I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}x_k\rangle^{p/2}\big]\Big)^{1/p} + p\gamma R^2 A_{r-1} \\
&\quad \text{by conditioning with respect to } x_k, \\
&\leqslant p\gamma R A_{r-1}\Big(\sum_{k=2}^n \mathbb{E}\big[\langle x_k, (I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}x_k\rangle (R^2\operatorname{tr} M)^{p/2-1}\big]\Big)^{1/p} + p\gamma R^2 A_{r-1} \\
&\quad \text{bounding } \langle x_k, (I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}x_k\rangle \text{ by } R^2\operatorname{tr} M, \\
&\leqslant p\gamma R^2 A_{r-1} + p\gamma R A_{r-1}\Big((R^2)^{p/2-1}\sum_{k=2}^n \operatorname{tr} H(I-\gamma H)^{n-k}M(I-\gamma H)^{n-k}\Big)^{1/p} \\
&\leqslant p\gamma R^2 A_{r-1} + p\gamma R A_{r-1}\Big((R^2)^{p/2-1}R^2\gamma^{-1}\Big)^{1/p} = p\gamma R^2 A_{r-1} + p(\gamma R^2)^{1-1/p}RA_{r-1}
\end{aligned}$$

This implies that $A_r \leqslant A_0\big(\sqrt{p\gamma R^2\kappa} + p\gamma R^2 + p(\gamma R^2)^{1-1/p}\big)^r$, which in turn implies, from Eq. (20),

$$A_r \leqslant \sqrt{p\gamma R^2}(\sigma + \tau\sqrt{p\gamma R^2})\big(\sqrt{p\gamma R^2\kappa} + p\gamma R^2 + p(\gamma R^2)^{1-1/p}\big)^r \tag{21}$$

The condition on γ will come from the requirement that $\sqrt{p\gamma R^2\kappa} + p\gamma R^2 + p(\gamma R^2)^{1-1/p} < 1$.

Bound on $\|H^{1/2}\overline{\eta}_{n-1}^r\|$. We have the closed-form expression:

$$\overline{\eta}_{n-1}^r = \frac{\gamma}{n}\sum_{j=2}^{n-1} \frac{I-(I-\gamma H)^{n-j}}{\gamma H}(H - x_j\otimes x_j)\eta_{j-1}^{r-1}$$

leading to, using BRP's inequality in Eq. (18), similar arguments than in the previous bounds, $\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}H \preccurlyeq H^{-1}\gamma$ and $\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}H \preccurlyeq \frac{n}{\gamma}I$:

$$\begin{aligned}
& \|H^{1/2}\bar{\eta}^r_{n-1}\|_p \\
\leqslant\ & \frac{\gamma\sqrt{p}}{n}\Big\|\sum_{j=2}^{n-1}\langle\eta^{r-1}_{j-1},\mathbb{E}\Big[(H-x_j\otimes x_j)\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}H(H-x_j\otimes x_j)\Big]\eta^{r-1}_{j-1}\rangle\Big\|^{1/2}_{p/2} \\
& +\frac{\gamma p}{n}\Big\|\sup_{j\in\{2,\dots,n-1\}}\langle\eta^{r-1}_{j-1},(H-x_j\otimes x_j)\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}H(H-x_j\otimes x_j)\eta^{r-1}_{j-1}\rangle\Big\|^{1/2}_{p/2} \\
\leqslant\ & \frac{\gamma\sqrt{p}}{n}\Big\|\sum_{j=2}^{n-1}\langle\eta^{r-1}_{j-1},\mathbb{E}\Big[(H-x_j\otimes x_j)\gamma^{-2}H^{-1}(H-x_j\otimes x_j)\Big]\eta^{r-1}_{j-1}\rangle\Big\|^{1/2}_{p/2} \\
& +\frac{\gamma p}{n}\Big\|\sup_{j\in\{2,\dots,n-1\}}\langle\eta^{r-1}_{j-1},H\frac{n}{\gamma}H\eta^{r-1}_{j-1}\rangle\Big\|^{1/2}_{p/2} \\
& +\frac{\gamma p}{n}\Big\|\sup_{j\in\{2,\dots,n-1\}}\langle\eta^{r-1}_{j-1},(x_j\otimes x_j)\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}H(x_j\otimes x_j)\eta^{r-1}_{j-1}\rangle\Big\|^{1/2}_{p/2} \\
\leqslant\ & \frac{\sqrt{p}}{n}\Big\|\sum_{j=2}^{n-1}\kappa\langle\eta^{r-1}_{j-1},H\eta^{r-1}_{j-1}\rangle\operatorname{tr}H^{-1}H\Big\|^{1/2}_{p/2}+\frac{\sqrt{\gamma}pR^2}{\sqrt{n}}A_{r-1} \\
& +\frac{\gamma p}{n}\bigg(\sum_{j=2}^{n-1}\mathbb{E}\Big[\langle\eta^{r-1}_{j-1},x_j\rangle^p\langle x_j,\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}Hx_j\rangle^{p/2}\Big]\bigg)^{1/p} \\
\leqslant\ & \frac{\sqrt{p\kappa d}}{\sqrt{n}}RA_{r-1}+\frac{\sqrt{\gamma}pR^2}{\sqrt{n}}A_{r-1} \\
& +\frac{\gamma p}{n}RA_{r-1}\bigg(\sum_{j=2}^{n-1}\mathbb{E}\Big[\langle x_j,\big(I-(I-\gamma H)^{n-j}\big)^2\big(\gamma H\big)^{-2}Hx_j\rangle^{p/2}\Big]\bigg)^{1/p} \\
\leqslant\ & \frac{\sqrt{p}}{\sqrt{n}}RA_{r-1}(\sqrt{\gamma pR^2}+\sqrt{\kappa d})+\frac{\gamma p}{n}RA_{r-1}\bigg(\sum_{j=2}^{n-1}\frac{d}{\gamma^2}\big(\frac{n}{\gamma}R^2\big)^{p/2-1}\bigg)^{1/p} \\
\leqslant\ & \frac{\sqrt{p}}{\sqrt{n}}RA_{r-1}(\sqrt{\gamma pR^2}+\sqrt{\kappa d})+\frac{\gamma p}{n}RA_{r-1}n^{1/2}\gamma^{-1/2-1/p}R^{1-2/p}d^{1/p} \\
=\ & \frac{\sqrt{p}}{\sqrt{n}}RA_{r-1}(\sqrt{\gamma pR^2}+\sqrt{\kappa d})+\frac{p}{\sqrt{n}}RA_{r-1}(\gamma R^2)^{1/2-1/p}d^{1/p} \\
=\ & \frac{\sqrt{p}}{\sqrt{n}}RA_{r-1}\Big[\sqrt{\gamma pR^2}+\sqrt{\kappa d}+\sqrt{p}(\gamma R^2)^{1/2-1/p}d^{1/p}\Big] \\
\leqslant\ & \frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}(\sigma+\tau\sqrt{p\gamma R^2})(\sqrt{\gamma pR^2}+\sqrt{\kappa d}+\sqrt{p}d^{1/p}(\gamma R^2)^{1/2-1/p})\big(\sqrt{p\gamma R^2\kappa}+p\gamma R^2+p(\gamma R^2)^{1-1/p}\big)^{r-1}
\end{aligned}$$

using Eq. (21).

We may then impose a restriction on γR^2, i.e., $\gamma R^2\leqslant\frac{1}{\alpha\kappa p}$ with $\alpha>1$. We then have

$$\begin{aligned}
\sqrt{p\gamma R^2\kappa}+p\gamma R^2+p(\gamma R^2)^{1-1/p} &\leqslant \frac{1}{\sqrt{\alpha}}+\frac{1}{\alpha}+(\alpha p)^{-1+1/p} \\
&\leqslant \frac{1}{\sqrt{\alpha}}+\frac{1}{\alpha}+(\alpha)^{-1}(\alpha p)^{1/p}
\end{aligned}$$

$$\begin{aligned} &\leqslant \frac{1}{\sqrt{\alpha}} + \frac{1}{\alpha} + (\alpha)^{-1}(\alpha 2)^{1/2} \text{ if } \alpha > 2, \\ &= \frac{1}{\sqrt{\alpha}} + \frac{1}{\alpha} + \sqrt{\frac{2}{\alpha}} \end{aligned}$$

With $\alpha = 12$, we obtain a bound of $0.781 \leqslant \frac{8}{10}$ above.

This leads to the bound

$$\begin{aligned} \|H^{1/2}\overline{\eta}_{n-1}^r\|_p &\leqslant \frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\big(\sigma + \frac{\tau}{\sqrt{12}\sqrt{\kappa}}\big)\big(\frac{1}{\sqrt{12\kappa}} + \sqrt{\kappa d} + d^{1/p}\sqrt{p}\big(\frac{1}{12p}\big)^{(1/2-1/p)/(1-1/p)}\big)\big(8/10\big)^{r-1} \\ &\leqslant \frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\tau(1 + \frac{1}{\sqrt{12}})\sqrt{\kappa d}(\frac{1}{\sqrt{12}} + 1 + \sup_{p\geqslant 2}\sqrt{p}\big(\frac{1}{12p}\big)^{(1/2-1/p)/(1-1/p)}\big)\big(8/10\big)^{r-1} \\ &\leqslant \frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\tau(1 + \frac{1}{\sqrt{12}})\sqrt{\kappa d}(\frac{1}{\sqrt{12}} + 1 + \sqrt{2}\big)\big(8/10\big)^{r-1} \\ &\leqslant \frac{7}{2}\frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\tau\sqrt{\kappa d}\big(8/10\big)^{r-1} \qquad (22) \end{aligned}$$

Bound on $\|H^{1/2}(\overline{\eta}_{n-1} - \sum_{i=0}^r \overline{\eta}_{n-1}^i)\|_p$. From Eq. (16) and the fact that $0 \preccurlyeq I - \gamma x_n \otimes x_n \preccurlyeq I$ almost surely, we get:

$$\begin{aligned} \Big\|\eta_n - \sum_{i=0}^r \eta_n^i\Big\|_p &\leqslant \Big\|\eta_{n-1} - \sum_{i=0}^r \eta_{n-1}^i\Big\|_p + \gamma\big\|(H - x_n \otimes x_n)\eta_{n-1}^r\big\|_p \\ &\leqslant \Big\|\eta_{n-1} - \sum_{i=0}^r \eta_{n-1}^i\Big\|_p + \gamma R\big\|\langle x_n, \eta_{n-1}^r\rangle\big\|_p \\ &\leqslant \Big\|\eta_{n-1} - \sum_{i=0}^r \eta_{n-1}^i\Big\|_p + \gamma R^2 A_r \end{aligned}$$

This implies that

$$\Big\|H^{1/2}(\overline{\eta}_{n-1} - \sum_{i=0}^r \overline{\eta}_{n-1}^i)\Big\|_p \leqslant R\Big\|\eta_n - \sum_{i=0}^r \eta_n^i\Big\|_p \leqslant n\gamma R^3 A_r \qquad (23)$$

Putting things together. We get by combining Lemma 4 with Eq. (22) and Eq. (23) and then letting r tends to infinity,

$$\begin{aligned} \|H^{1/2}\overline{\eta}_{n-1}^r\|_p &\leqslant \sum_{i=1}^r \big\|H^{1/2}\overline{\eta}_{n-1}^i\big\|_p + \big\|H^{1/2}\overline{\eta}_{n-1}^0\big\|_p + \Big\|H^{1/2}(\overline{\eta}_{n-1} - \sum_{i=0}^r \overline{\eta}_{n-1}^i)\Big\|_p \\ &\leqslant \Big\{\frac{7}{2}\frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\tau\sqrt{\kappa d}\frac{1-(8/10)^r}{1-8/10}\Big\} + \Big\{\frac{\sqrt{pd}\sigma}{\sqrt{n}} + \frac{\sqrt{\gamma}pR\tau}{\sqrt{n}}\Big\} + O((8/10)^r) \\ &\leqslant \frac{\sqrt{pd}\sigma}{\sqrt{n}} + \frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\tau\sqrt{\kappa d}(1 + \frac{7}{2}\frac{10}{2}) \leqslant \frac{\sqrt{pd}\sigma}{\sqrt{n}} + 18.5\frac{\sqrt{p}}{\sqrt{n}}\sqrt{p\gamma R^2}\tau\sqrt{\kappa d} \\ &\leqslant \frac{\sqrt{pd}\sigma}{\sqrt{n}} + \frac{18.5}{\sqrt{12}}\frac{\sqrt{pd}\tau}{\sqrt{n}} \leqslant \frac{\sqrt{pd}(\sigma + 6\tau)}{\sqrt{n}} \leqslant 7\frac{\sqrt{pd}\tau}{\sqrt{n}} \end{aligned}$$

B.3 Final bound

For $\gamma \leqslant \frac{1}{12\kappa p R^2}$, we obtain, from the last equations of Section B.1 and Section B.2,

$$\begin{aligned}\|H^{1/2}\overline{\eta}^r_{n-1}\|_p &\leqslant 7\frac{\sqrt{pd}\tau}{\sqrt{n}} + \frac{\|\eta_0\|}{\sqrt{n\gamma}}\sqrt{2+3p\gamma R^2} \\ &\leqslant \frac{\sqrt{p}}{\sqrt{n}}\bigg(7\sqrt{d}\tau + R\|\eta_0\|\sqrt{3+\frac{2}{\gamma p R^2}}\bigg)\end{aligned}$$

Moreover, when $\gamma = \frac{1}{12\kappa p R^2}$, we have:

$$\begin{aligned}\|H^{1/2}\overline{\eta}^r_{n-1}\|_p &\leqslant 7\frac{\sqrt{pd}\tau}{\sqrt{n}} + \frac{\|\eta_0\|}{\sqrt{n}}\sqrt{12\kappa p R^2}\sqrt{2+\frac{1}{4}} \\ &\leqslant 7\frac{\sqrt{pd}\tau}{\sqrt{n}} + \frac{6R\|\eta_0\|}{\sqrt{n}}\sqrt{\kappa p} = \frac{\sqrt{p}}{\sqrt{n}}\Big(7\sqrt{d}\tau + 6\sqrt{\kappa}R\|\theta_0-\theta_*\|\Big)\end{aligned}$$

B.4 Proof of Corollary 1

We have from the previous proposition, for $\gamma \leqslant \frac{1}{12\kappa p R^2}$:

$$\begin{aligned}\big(\mathbb{E}\big|f(\overline{\theta}_{n-1}) - f(\theta_*)\big|^p\big)^{1/p} &\leqslant \frac{1}{2n}\bigg(\sqrt{p}\big[7\tau\sqrt{d} + R\|\theta_0-\theta_*\|\sqrt{3}\big] + R\|\theta_0-\theta_*\|\sqrt{\frac{2}{\gamma R^2}}\bigg)^2 \\ &\leqslant \frac{1}{2n}\bigg(\sqrt{p}\,\Box + \triangle\sqrt{\frac{1}{\eta}}\bigg)^2\end{aligned}$$

with $\eta = 12\gamma\kappa R^2 \leqslant 1/p$, and $\Box = 7\tau\sqrt{d} + R\|\theta_0-\theta_*\|\sqrt{3}$ and $\triangle = R\|\theta_0-\theta_*\|\sqrt{24\kappa}$.

This leads to, using Markov's inequality:

$$\mathbb{P}\bigg(f(\overline{\theta}_{n-1}) - f(\theta_*) \geqslant \frac{t}{2n}\bigg) \leqslant \bigg(\frac{\sqrt{p}\,\Box + \triangle\sqrt{1/\eta}}{\sqrt{t}}\bigg)^{2p}$$

This leads to, with $p = \frac{1}{\eta}$,

$$\mathbb{P}\bigg(f(\overline{\theta}_{n-1}) - f(\theta_*) \geqslant \frac{t}{2n}\bigg) \leqslant \bigg(\frac{(\Box + \triangle)^2}{\eta t}\bigg)^{1/\eta}$$

This leads to

$$\mathbb{P}\bigg(f(\overline{\theta}_{n-1}) - f(\theta_*) \geqslant \frac{t}{2n}\big[7\tau\sqrt{d} + R\|\theta_0-\theta_*\|(\sqrt{3}+\sqrt{24\kappa})\big]^2\bigg) \leqslant \bigg(\frac{1}{12\gamma\kappa R^2 t}\bigg)^{1/(12\gamma\kappa R^2)} \qquad (24)$$

Thus the large deviations decay as power of t, with a power that decays as $1/(12\gamma\kappa R^2)$. If γ is small, the deviations are lighter.

In order to get the desired result, we simply take $t = \frac{1}{12\gamma\kappa R^2}\delta^{-12\kappa\gamma R^2}$.

C Proof of Theorem 3

The proof relies mostly on properties of approximate Newton steps: $\theta_1 \overset{\text{def}}{=} \widetilde{\theta}_n$ is an approximate minimizer of f, and $\theta_3 \overset{\text{def}}{=} \zeta_n$ is an approximate minimizer of the associated quadratic problem.

In terms of convergence rates, θ_1 will be $(1/\sqrt{n})$-optimal, while θ_3 will be $(1/n)$-optimal for the quadratic problem because of previous results on averaged LMS. A classical property is that a single Newton step squares the error. Therefore, the full Newton step should have an error which is the square of the one of θ_1, i.e., $O(1/n)$. Overall, since θ_3 approaches the full Newton step with rate $O(1/n)$, this makes a bound of $O(1/n)$.

In Section C.1, we provide a general *deterministic* result on the Newton step, while in Section C.2, we combine with two stochastic approximation results, making the informal reasoning above more precise.

C.1 Approximate Newton step

In this section, we study the effect of an approximate Newton step. We consider $\theta_1 \in \mathcal{H}$, the Newton iterate $\theta_2 = \theta_1 - f''(\theta_1)^{-1} f'(\theta_1)$, and an approximation θ_3 of θ_2. In the next proposition, we provide a bound on $f(\theta_3) - f(\theta_*)$, under different conditions, whether θ_1 is close to optimal for f, and/or θ_3 is close to optimal for the quadratic approximation around θ_1 (i.e., close to θ_2). Eq. (25) corresponds to the least-favorable situations where both errors are small, while Eq. (26) and Eq. (27) consider cases where θ_1 is sufficiently good. See proof in Section E.3. These three cases are necessary for the probabilistic control.

Proposition 1 (Approximate Newton step)**.** *Assume* ***(B3-4)****, and $\theta_1, \theta_2, \theta_3 \in \mathcal{H}$ such that $f(\theta_1) - f(\theta_*) = \varepsilon_1$, $\theta_2 = \theta_1 - f''(\theta_1)^{-1} f'(\theta_1)$ and $\frac{1}{2}\langle \theta_3 - \theta_2, f''(\theta_1)(\theta_3 - \theta_2)\rangle = \varepsilon_2$. Then, if $t^2 = \varepsilon_1 \kappa \rho$,*

$$f(\theta_3) - f(\theta_*) \leqslant \varepsilon_1 + \sqrt{2\rho\varepsilon_2} e^{\sqrt{3+t^2}t/2} + \sqrt{\rho} e^{\sqrt{3+t^2}t} (\sqrt{3} + 2t)\sqrt{\varepsilon_1} \tag{25}$$

If $t = \sqrt{\varepsilon_1 \kappa \rho} \leqslant 1/16$, then

$$f(\theta_3) - f(\theta_*) \leqslant 57\kappa\rho\varepsilon_1^2 + 2\sqrt{\rho\varepsilon_2}. \tag{26}$$

Moreover, if $t = \sqrt{\varepsilon_1 \kappa \rho} \leqslant 1/16$ and $\varepsilon_2 \kappa \rho \leqslant 1/16$, then

$$f(\theta_3) - f(\theta_*) \leqslant 57\kappa\rho\varepsilon_1^2 + 12\varepsilon_2 \tag{27}$$

Note that in the favorable situation in Eq. (26), we get error of the form $O(\varepsilon_1^2 + \varepsilon_2)$. It essentially suffices now to show that in our set-up, in a probabilistic sense to be determined, $\varepsilon_1 = O(1/\sqrt{n})$ and $\varepsilon_2 = O(1/n)$, while controlling the unfavorable situations.

C.2 Stochastic analysis

We consider the following two-step algorithm:

- Starting from any initialization θ_0, run n iterations of averaged stochastic gradient descent to get θ_1,
- Run from θ_1 n steps of LMS on the quadratic approximation around θ_1, to get θ_3, which is an approximation of the Newton step θ_2.

We consider the events

$$A_1 = \left\{ f(\theta_1) - f(\theta_*) \leqslant \frac{1}{16^2}(\kappa\rho)^{-1} \right\} = \left\{ \varepsilon_1 \leqslant \frac{1}{16^2}(\kappa\rho)^{-1} \right\}$$

and

$$A_2 = \left\{ \frac{1}{2}\langle \theta_3 - \theta_2, f''(\theta_1)(\theta_3 - \theta_2)\rangle \leqslant \frac{1}{16}(\kappa\rho)^{-1} \right\} = \left\{ \varepsilon_2 \leqslant \frac{1}{16}(\kappa\rho)^{-1} \right\}$$

We denote by $\mathcal{G}_1$ the σ-field generated by the first n observations (the ones used to define θ_1). We have, by separating all events, i.e., using $1 = 1_{A_1}1_{A_2} + 1_{A_1}1_{A_2^c} + 1_{A_1^c}$:

$$\begin{aligned}
& \mathbb{E}\big[f(\theta_3) - f(\theta_*)\big|\mathcal{G}_1\big] \\
= \ & \mathbb{E}\big[1_{A_1}1_{A_2}\big(f(\theta_3) - f(\theta_*)\big)\big|\mathcal{G}_1\big] + \mathbb{E}\big[1_{A_1}1_{A_2^c}\big(f(\theta_3) - f(\theta_*)\big)\big|\mathcal{G}_1\big] + \mathbb{E}\big[1_{A_1^c}\big(f(\theta_3) - f(\theta_*)\big)\big|\mathcal{G}_1\big] \\
\leqslant \ & \mathbb{E}\big[1_{A_1}1_{A_2}\big(57\kappa\rho\varepsilon_1^2 + 12\varepsilon_2\big)\big|\mathcal{G}_1\big] + \mathbb{E}\big[1_{A_1}1_{A_2^c}\big(57\kappa\rho\varepsilon_1^2 + 2\sqrt{\rho\varepsilon_2}\big)\big|\mathcal{G}_1\big] \\
& + \mathbb{E}\big[1_{A_1^c}\big(\varepsilon_1 + \sqrt{2\rho\varepsilon_2}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3} + 2t)\sqrt{\varepsilon_1}\big)\big|\mathcal{G}_1\big] \text{ using Prop. 1,} \\
\leqslant \ & 57\kappa\rho\varepsilon_1^2 + 12\mathbb{E}\big[1_{A_1}\varepsilon_2\big|\mathcal{G}_1\big] + \mathbb{E}\big[1_{A_1}1_{A_2^c}\big(2\sqrt{\rho\varepsilon_2}\big)\big|\mathcal{G}_1\big] \\
& + 1_{A_1^c}\bigg(\varepsilon_1 + \mathbb{E}\big[\sqrt{\varepsilon_2}\big|\mathcal{G}_1\big]\sqrt{2\rho}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3} + 2t)\sqrt{\varepsilon_1}\bigg) \\
\leqslant \ & 57\kappa\rho\varepsilon_1^2 + 12 \times 1_{A_1}\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big] + 2\sqrt{\rho}1_{A_1}\sqrt{\mathbb{P}(A_2^c|\mathcal{G}_1)}\sqrt{\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big]} \text{ using Cauchy-Schwarz inequality,} \\
& + 1_{A_1^c}\bigg(\varepsilon_1 + \mathbb{E}\big[\sqrt{\varepsilon_2}\big|\mathcal{G}_1\big]\sqrt{2\rho}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3} + 2t)\sqrt{\varepsilon_1}\bigg) \\
= \ & 57\kappa\rho\varepsilon_1^2 + 12 \times 1_{A_1}\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big] + 2\sqrt{\rho}1_{A_1}\sqrt{\mathbb{P}(\{\varepsilon_2 \geqslant \frac{1}{16\kappa\rho}|\mathcal{G}_1)}\sqrt{\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big]} \\
& + 1_{A_1^c}\bigg(\varepsilon_1 + \mathbb{E}\big[\sqrt{\varepsilon_2}\big|\mathcal{G}_1\big]\sqrt{2\rho}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3} + 2t)\sqrt{\varepsilon_1}\bigg) \\
\leqslant \ & 57\kappa\rho\varepsilon_1^2 + 12 \times 1_{A_1}\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big] + 2\sqrt{\rho}1_{A_1}\sqrt{16\kappa\rho\mathbb{E}(\varepsilon_2|\mathcal{G}_1)}\sqrt{\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big]} \text{ using Markov's inequality,} \\
& + 1_{A_1^c}\bigg(\varepsilon_1 + \mathbb{E}\big[\sqrt{\varepsilon_2}\big|\mathcal{G}_1\big]\sqrt{2\rho}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3} + 2t)\sqrt{\varepsilon_1}\bigg) \\
= \ & 57\kappa\rho\varepsilon_1^2 + 1_{A_1}\mathbb{E}\big[\varepsilon_2\big|\mathcal{G}_1\big](12 + 2\sqrt{\rho}\sqrt{16\kappa\rho}) \\
& + 1_{A_1^c}\bigg(\varepsilon_1 + \mathbb{E}\big[\sqrt{\varepsilon_2}\big|\mathcal{G}_1\big]\sqrt{2\rho}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3} + 2t)\sqrt{\varepsilon_1}\bigg) \qquad (28)
\end{aligned}$$

We now need to control ε_2, i.e., the error made by the LMS algorithm started from θ_1.

LMS on the second-order Taylor approximation. We consider the quadratic approximation around $\theta_1 \in \mathcal{H}$, and write is as an expectation, i.e.,

$$\begin{aligned}
g(\theta) &= f(\theta_1) + \langle f'(\theta_1), \theta - \theta_1\rangle + \frac{1}{2}\langle \theta - \theta_1, f''(\theta_1)(\theta - \theta_1)\rangle \\
&= f(\theta_1) + \mathbb{E}\big[\big\langle \ell'(y, \langle x, \theta_1\rangle)x, \theta - \theta_1\big\rangle\big] + \frac{1}{2}\mathbb{E}\big[\big\langle \theta - \theta_1, \ell''(y, \langle x, \theta_1\rangle)x \otimes x(\theta - \theta_1)\big\rangle\big] \\
&= f(\theta_1) + \big\langle \mathbb{E}\big[\ell'(y, \langle x, \theta_1\rangle)x\big], \theta - \theta_1\big\rangle + \frac{1}{2}\big\langle \theta - \theta_1, \mathbb{E}\big[\ell''(y, \langle x, \theta_1\rangle)x \otimes x\big](\theta - \theta_1)\big\rangle
\end{aligned}$$

We consider $\widetilde{x}_n = \sqrt{\ell''(y_n, \langle x_n, \theta_1\rangle)} x_n$ and $\widetilde{z}_n = -\ell'(y_n, \langle x_n, \theta_1\rangle) x_n$, so that

$$g(\theta) \quad = \quad f(\theta_1) + \mathbb{E}\bigg[\frac{1}{2}\langle \theta - \theta_1, \widetilde{x}_n\rangle^2 - \langle \widetilde{z}_n, \theta - \theta_1\rangle\bigg]$$

We denote by $\theta_2 = \theta_1 - f''(\theta_1)^{-1} f'(\theta_1)$ the output of the Newton step, i.e., the global minimizer of g, and $\widetilde{\xi}_n = \widetilde{z}_n - \langle \theta_2 - \theta_1, \widetilde{x}_n\rangle \widetilde{x}_n$ the residual.

We have $\mathbb{E}\widetilde{\xi}_n = 0$, $\mathbb{E}\big[\widetilde{x}_n \otimes \widetilde{x}_n\big] = f''(\theta_1)$, and, for any $z \in \mathcal{H}$:

$$\begin{aligned}
\big(\mathbb{E}\big[\langle z, \widetilde{\xi}_n\rangle\big]^2\big)^{1/2} \quad &= \quad \bigg(\mathbb{E}\big[\big(\ell''(y_n, \langle x_n, \theta_1\rangle)\langle \theta_2 - \theta_1, x_n\rangle + \ell'(y_n, \langle x_n, \theta_1\rangle)\big)\langle z, x_n\rangle\big]^2\bigg)^{1/2} \\
&\leqslant \quad \bigg(\mathbb{E}\big[\big(\ell''(y_n, \langle x_n, \theta_1\rangle)\big)\langle \theta_2 - \theta_1, x_n\rangle \langle z, x_n\rangle\big]^2\bigg)^{1/2} + \bigg(\mathbb{E}\big[\ell'(y_n, \langle x_n, \theta_1\rangle \langle z, x_n\rangle\big]^2\bigg)^{1/2} \\
&\qquad \text{using the triangle inequality,} \\
&\leqslant \quad \sqrt{\kappa}\sqrt{\langle z, f''(\theta_1) z\rangle \langle \theta_2 - \theta_1, f''(\theta_1)(\theta_2 - \theta_1)\rangle} + \bigg(\mathbb{E}\big[\langle z, x_n\rangle\big]^2\bigg)^{1/2} \\
&\leqslant \quad \sqrt{\kappa}\sqrt{\langle z, f''(\theta_1) z\rangle \langle \theta_2 - \theta_1, f''(\theta_1)(\theta_2 - \theta_1)\rangle} + \sqrt{\rho}\sqrt{\langle z, f''(\theta_*) z\rangle} \\
&\leqslant \quad \sqrt{\langle z, f''(\theta_1) z\rangle}\bigg[\sqrt{\kappa}\|f''(\theta_1)^{-1/2} f'(\theta_1)\| + \sqrt{\rho} e^{\sqrt{\kappa\rho} d_1/2}\bigg]
\end{aligned}$$

where we denote $d_1^2 = \langle \theta_1 - \theta_*, H(\theta_1 - \theta_*)\rangle$, and we have used assumption **(B4)** , $|\ell'| \leqslant 1$ and Prop. 5 relating H and $f''(\theta_1)$. This leads to

$$\mathbb{E}\big[\widetilde{\xi}_n \otimes \widetilde{\xi}_n\big] \preccurlyeq \bigg[\sqrt{\kappa}\|f''(\theta_1)^{-1/2} f'(\theta_1)\| + \sqrt{\rho} e^{\sqrt{\kappa\rho} d_1/2}\bigg]^2 f''(\theta_1)$$

Thus, we have:

- $\mathbb{E}\big[\widetilde{\xi}_n \otimes \widetilde{\xi}_n\big] \preccurlyeq \sigma^2 f''(\theta_1)$ with $\sigma = \sqrt{\kappa}\|f''(\theta_1)^{-1/2} f'(\theta_1)\| + \sqrt{\rho} e^{\sqrt{\kappa\rho} d_1/2}$.
- $\|x_n\|^2 \leqslant R^2/4$ almost surely.

We may thus apply the previous results, i.e., Theorem 1, to obtain with the LMS algorithm a $\theta_3 \in \mathcal{H}$ such that, with $\gamma = \frac{1}{R^2}$:

$$\begin{aligned}
\mathbb{E}\big[\varepsilon_2 | \mathcal{G}_1\big] \quad &\leqslant \quad \frac{2}{n}\bigg[\sqrt{d}\sqrt{\kappa}\|f''(\theta_1)^{-1/2} f'(\theta_1)\| + \sqrt{d}\sqrt{\rho} e^{\sqrt{\kappa\rho} d_1/2} + \frac{R}{2}\|f''(\theta_1)^{-1} f'(\theta_1)\|\bigg]^2 \\
&\leqslant \quad \frac{2}{n}\bigg[\sqrt{d}\sqrt{\kappa}\|H^{-1/2} f'(\theta_1)\| e^{\sqrt{\kappa\rho} d_1/2} + \sqrt{d}\sqrt{\rho} e^{\sqrt{\kappa\rho} d_1/2} + \frac{R}{2} e^{\sqrt{\kappa\rho} d_1}\|H^{-1} f'(\theta_1)\|\bigg]^2 \\
&\qquad \text{using Prop. 5 ,} \\
&\leqslant \quad \frac{2}{n}\bigg[\sqrt{d}\sqrt{\kappa}(\sqrt{3} + 2\sqrt{\kappa\rho\varepsilon_1})\sqrt{\varepsilon_1} e^{\sqrt{\kappa\rho} d_1/2} + \sqrt{d}\sqrt{\rho} e^{\sqrt{\kappa\rho} d_1/2} + \frac{R}{2} e^{\sqrt{\kappa\rho} d_1}\frac{e^{\sqrt{\kappa\rho} d_1} - 1}{\sqrt{\kappa\rho} d_1}\|\theta_1 - \theta_*\|\bigg]^2 \\
&\qquad \text{using Prop. 11 and Eq. (49) from Section E.3}
\end{aligned}$$

Thus, $\mathbb{E}\big[\varepsilon_2 | \mathcal{G}_1\big] \leqslant \frac{2}{n}\bigg[R\|\theta_1 - \theta_*\|\triangle_2(t) + \triangle_3(t)\sqrt{d\rho}\bigg]^2$, with increasing functions

$$\triangle_2(t) = \frac{1}{2} e^{\sqrt{3+t^2} t}\frac{e^{\sqrt{3+t^2} t} - 1}{\sqrt{3+t^2} t}$$

$$\triangle_3(t) = \big[(\sqrt{3}+2t)t+1\big]e^{\sqrt{3+t^2}t/2}$$

which are such that $\triangle_2(t) \leqslant 0.6$ and $\triangle_3(t) \leqslant 1.2$ if $t \leqslant 1/16$.

We then get from Eq. (28):

$$\begin{aligned}
& \mathbb{E}\big[f(\theta_3) - f(\theta_*)\big|\mathcal{G}_1\big] \\
\leqslant\ & 57\kappa\rho\varepsilon_1^2 + \frac{2}{n}(12 + 2\sqrt{\rho}\sqrt{16\kappa\rho})\big[0.6R\|\theta_1-\theta_*\| + \sqrt{d\rho}1.2\big]^2 \\
& +1_{A_1^c}\bigg(\varepsilon_1 + \sqrt{\frac{2}{n}}\big(R\|\theta_1-\theta_*\|\triangle_2(t) + \triangle_3(t)\sqrt{d\rho}\big)\sqrt{2\rho}e^{\sqrt{3+t^2}t/2} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3}+2t)\sqrt{\varepsilon_1}\bigg) \\
\leqslant\ & 57\kappa\rho\varepsilon_1^2 + \frac{1}{n}(12 + 8\sqrt{\rho}\sqrt{\kappa\rho})\big(6d\rho + \frac{3}{2}R^2\|\theta_1-\theta_*\|^2\big) \\
& +1_{A_1^c}\bigg(\sqrt{\varepsilon_1}\big[\frac{\sqrt{t}}{\sqrt{\kappa\rho}} + \sqrt{\rho}e^{\sqrt{3+t^2}t}(\sqrt{3}+2t)\big] + \sqrt{\frac{2}{n}}\big(R\|\theta_1-\theta_*\|\triangle_2(t) + \triangle_3(t)\sqrt{d\rho}\big)\sqrt{2\rho}e^{\sqrt{3+t^2}t/2}\bigg) \\
=\ & 57\kappa\rho\varepsilon_1^2 + \frac{12}{n}(3 + 2\sqrt{\rho}\sqrt{\kappa\rho})\big(2d\rho + \frac{1}{2}R^2\|\theta_1-\theta_*\|^2\big) \\
& +1_{A_1^c}\bigg(\sqrt{\rho\varepsilon_1}\big[\frac{\sqrt{t}}{\rho\sqrt{\kappa}} + e^{\sqrt{3+t^2}t}(\sqrt{3}+2t)\big] + \sqrt{\frac{4\rho}{n}}\triangle_2(t)e^{\sqrt{3+t^2}t/2}R\|\theta_1-\theta_*\| + \sqrt{\frac{4\rho}{n}}\triangle_3(t)e^{\sqrt{3+t^2}t/2}\sqrt{\rho d}\bigg) \\
\leqslant\ & 57\kappa\rho\varepsilon_1^2 + \frac{12}{n}(3 + 2\sqrt{\rho}\sqrt{\kappa\rho})\big(2d\rho + \frac{1}{2}R^2\|\theta_1-\theta_*\|^2\big) \\
& +1_{A_1^c}\bigg(\sqrt{\rho\varepsilon_1}\triangle_4(t) + \sqrt{\frac{\rho}{n}}R\|\theta_1-\theta_*\|\triangle_5(t) + \sqrt{\frac{\rho}{n}}\sqrt{\rho d}\triangle_6(t)\bigg) \qquad (29)
\end{aligned}$$

with (using $\rho \geqslant 4$):

$$\begin{aligned}
\triangle_4(t) &= \frac{\sqrt{t}}{4} + e^{\sqrt{3+t^2}t}(\sqrt{3}+2t) \leqslant 5\exp(2t^2) \\
\triangle_5(t) &= 2e^{\sqrt{3+t^2}t/2}\Delta_2(t) \leqslant 4\exp(3t^2) \\
\triangle_6(t) &= 2e^{\sqrt{3+t^2}t/2}\Delta_3(t) \leqslant 6\exp(3t^2)
\end{aligned}$$

The last inequalities may be checked graphically.

By taking expectations and using $\mathbb{E}|XYZ| \leqslant (\mathbb{E}|X|^2)^{1/2}(\mathbb{E}|X|^4)^{1/4}(\mathbb{E}|X|^4)^{1/4}$, this leads to, from Eq. (29):

$$\begin{aligned}
& \mathbb{E}\big[f(\theta_3) - f(\theta_*)\big] \\
\leqslant\ & 57\kappa\rho\mathbb{E}\big[\varepsilon_1^2\big] + \frac{12}{n}(3 + 2\sqrt{\rho}\sqrt{\kappa\rho})\big(2d\rho + \frac{1}{2}\mathbb{E}\big[R^2\|\theta_1-\theta_*\|^2\big]\big) \\
& +\sqrt{\rho}\sqrt{\mathbb{P}(A_1^c)}\bigg(\big(\mathbb{E}\big[\varepsilon_1^2\big]\big)^{1/4}\big(\mathbb{E}\big[\triangle_4(t)\big]^4\big)^{1/4} \\
& +\sqrt{\frac{1}{n}}\big(\mathbb{E}\big[R^4\|\theta_1-\theta_*\|^4\big]\big)^{1/4}\big(\mathbb{E}\big[\triangle_5(t)\big]^4\big)^{1/4} + \sqrt{\frac{1}{n}}\sqrt{\rho d}\big(\mathbb{E}\big[\triangle_6(t)\big]^4\big)^{1/4}\bigg) \qquad (30)
\end{aligned}$$

We now need to use bounds on the behavior of the first n steps of regular averaged stochastic gradient descent.

Fine results on averaged stochastic gradient descent. In order to get error bounds on θ_1, we run n steps of averaged stochastic gradient descent with constant-step size $\gamma = 1/(2R^2\sqrt{n})$. We need the following bounds from [21, Appendix E and Prop. 1]:

$$\begin{aligned}
\mathbb{E}[(f(\theta_1) - f(\theta_*))^2] &\leqslant \frac{1}{n}(R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{4})^2 \\
\mathbb{E}[(f(\theta_1) - f(\theta_*))^3] &\leqslant \frac{1}{n^{3/2}}(R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{2})^3 \\
\mathbb{E}\big[R^4\|\theta_1 - \theta_*\|^4\big] &\leqslant (R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{4})^2 \\
\mathbb{E}\big[R^2\|\theta_1 - \theta_*\|^2\big] &\leqslant R^2\|\theta_0 - \theta_*\|^2 + \frac{1}{4} \\
\mathbb{P}\bigg[f(\theta_1) - f(\theta_*) \geqslant \frac{1}{16^2}(\kappa\rho)^{-1}\bigg] &\leqslant 16^6(\kappa\rho)^3\mathbb{E}\big[f(\theta_1) - f(\theta_*)\big]^3 \text{ using Markov's inequality,} \\
&\leqslant 16^6(\kappa\rho)^3\frac{1}{n^{3/2}}(R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{2})^3
\end{aligned}$$

However, we need a finer control of the deviations in order to bound quantities of the form $e^{\alpha\varepsilon_1}$. In Section D, extending results from [21], we show in Prop. 3 that if $\frac{\alpha(10+2R^2\|\theta_0-\theta_*\|^2)}{\sqrt{n}} \leqslant \frac{1}{2e}$, then $Ee^{\alpha(f(\theta_1)-f(\theta_*))} \leqslant 1$.

Putting things together. From Eq. (30), we then get, if $\frac{6(10+2R^2\|\theta_0-\theta_*\|^2)}{\sqrt{n}} \leqslant \frac{1}{2e}$:

$$\begin{aligned}
&\mathbb{E}\big[f(\theta_3) - f(\theta_*)\big] \\
\leqslant\ & 57\kappa\rho\frac{1}{n}(R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{4})^2 + \frac{12}{n}(3 + 2\sqrt{\rho}\sqrt{\kappa\rho})\big(2d\rho + \frac{1}{4} + \frac{1}{2}R^2\|\theta_0 - \theta_*\|^2\big) \\
&+\sqrt{\rho}\sqrt{16^6(\kappa\rho)^3\frac{1}{n^{3/2}}(R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{2})^3} \\
&\times\bigg(5\big(\frac{1}{n}(R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{4})^2\big)^{1/4} + 4\sqrt{\frac{1}{n}}\big((R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{4})^2\big)^{1/4} + 6\sqrt{\frac{1}{n}}\sqrt{\rho d}\bigg)
\end{aligned}$$

Using the notation $D = (R^2\|\theta_0 - \theta_*\|^2 + \frac{3}{2})$, we obtain:

$$\begin{aligned}
&\mathbb{E}\big[f(\theta_3) - f(\theta_*)\big] \\
\leqslant\ & 57\kappa\rho\frac{1}{n}D^2 + \frac{12}{n}(3 + 2\sqrt{\rho}\sqrt{\kappa\rho})\big(2d\rho + \frac{D}{2}\big) \\
&+\sqrt{\rho}\sqrt{16^6(\kappa\rho)^3\frac{1}{n^{3/2}}D^3} \times \bigg(5\big(\frac{1}{n}D^2\big)^{1/4} + 4\sqrt{\frac{1}{n}}\big(D^2\big)^{1/4} + 6\sqrt{\frac{1}{n}}\sqrt{\rho d}\bigg) \\
=\ & 57\kappa\rho\frac{1}{n}D^2 + \frac{12}{n}(3 + 2\sqrt{\rho}\sqrt{\kappa\rho})\big(2d\rho + \frac{D}{2}\big) \\
&+\sqrt{\rho}16^3(\kappa\rho)^{3/2}\frac{1}{n^{3/4}}D^{3/2} \times \bigg(5\frac{1}{n^{1/4}}D^{1/2} + 4\frac{1}{n^{1/2}}D^{1/2} + 6\frac{1}{\sqrt{n}}\sqrt{\rho d}\bigg) \\
\leqslant\ & \frac{\kappa^{3/2}\rho^2}{n}\bigg[\frac{57}{4}D^2 + 12(\frac{3}{16} + \frac{2}{4})\big(2d\rho + \frac{D}{2}\big) + 16^3D^{3/2} \times \bigg(5D^{1/2} + 4\frac{1}{n^{1/4}}D^{1/2} + 6\frac{1}{n^{1/4}}\sqrt{\rho d}\bigg)\bigg] \\
&\qquad\qquad\text{using } \rho \geqslant 4 \text{ and } \kappa \geqslant 1, \\
\leqslant\ & \frac{\kappa^{3/2}\rho^2D^2}{n}\bigg[\frac{57}{4}1 + 12(\frac{3}{16} + \frac{2}{4})\big(2d\rho\frac{4}{9} + \frac{1}{2}\frac{2}{3}\big) + 16^3 \times \bigg(5 + 4 + 6\frac{1}{n^{1/4}}\frac{\sqrt{2}}{\sqrt{3}}\sqrt{\rho d}\bigg)\bigg] \text{ using } D \geqslant \frac{3}{2},
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{\kappa^{3/2}\rho^2 D^2}{n}\left[36881 + 20067\frac{\sqrt{\rho d}}{n^{1/4}} + 17d\rho\right] \leqslant \frac{\kappa^{3/2}\rho^3 d}{n}56965(R^2\|\theta_0-\theta_*\|^2+\frac{3}{2})^2 \\
&\leqslant \frac{\kappa^{3/2}\rho^3 d}{n}(16R\|\theta_0-\theta_*\|+19)^4
\end{aligned}
$$

The condition $\frac{6(10+2R^2\|\theta_0-\theta_*\|^2)}{\sqrt{n}} \leqslant \frac{1}{2e}$ is implied by $n \geqslant (19+9R\|\theta_0-\theta_*\|)^4$.

D Higher-order bounds for stochastic gradient descent

In this section, we provide high-order bounds for averaged stochastic gradient for logistic regression. The first proposition gives a finer result than [21], with a simpler proof, while the second proposition is new.

Proposition 2. *Assume* ***(B1-4)****. Consider the stochastic gradient recursion $\theta_n = \theta_{n-1} - \gamma\ell'(y_n, \langle\theta_{n-1}, x_n\rangle)x_n$ and its averaged version $\overline{\theta}_{n-1}$. We have, for all real $p \geqslant 1$,*

$$
\big\|f(\overline{\theta}_{n-1}) - f(\theta_*)\big\|_p \leqslant \frac{17\gamma R^2}{2}(\sqrt{p}+\frac{p}{\sqrt{n}})^2 + \frac{1}{\gamma n}\|\theta_0-\theta_*\|^2 \qquad (31)
$$

$$
\big\|\|\theta_n-\theta_*\|^2\big\|_p \leqslant 17\gamma^2 R^2 n(\sqrt{p}+\frac{p}{\sqrt{n}})^2 + 2\|\theta_0-\theta_*\|^2 \qquad (32)
$$

Proof. Following [21], we have the recursion:

$$
2\gamma\big[f(\theta_{n-1}) - f(\theta_*)\big] + \|\theta_n-\theta_*\|^2 \quad \leqslant \quad \|\theta_{n-1}-\theta_*\|^2 + \gamma^2R^2 + M_n
$$

with

$$
M_n = -2\gamma\langle\theta_{n-1}-\theta_*, f_n'(\theta_{n-1}) - f'(\theta_{n-1})\rangle
$$

This leads to

$$
2\gamma n f\bigg(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\bigg) - 2\gamma n f(\theta^*) + \|\theta_n-\theta_*\|^2 \leqslant A_n
$$

with $A_n = \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + \sum_{k=1}^n M_k$. Note that $\mathbb{E}(M_k|\mathcal{F}_{k-1}) = 0$ and $|M_k| \leqslant 4\gamma R\|\theta_{k-1}-\theta_*\| \leqslant 4\gamma R A_{k-1}^{1/2}$ almost surely. We may now use BRP's inequality in Eq. (18) to get:

$$
\begin{aligned}
\bigg\|\sup_{k\in\{0,\dots,n\}} A_k\bigg\|_p &\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + \sqrt{p}\bigg\|16\gamma^2R^2\sum_{k=1}^n\|\theta_{k-1}-\theta_*\|^2\bigg\|_{p/2}^{1/2} \\
&\quad +p\bigg\|\sup_{k\in\{1,\dots,n\}} 4\gamma R\|\theta_{k-1}-\theta_*\|\bigg\|_p \\
&\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + \sqrt{p}4\gamma R\sqrt{n}\bigg\|\sup_{k\in\{0,\dots,n-1\}} A_k\bigg\|_{p/2}^{1/2} \\
&\quad +p4\gamma R\bigg\|\sup_{k\in\{0,\dots,n-1\}} A_k^{1/2}\bigg\|_p \\
&\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + 4\gamma R\bigg\|\sup_{k\in\{0,\dots,n-1\}} A_k\bigg\|_{p/2}^{1/2}\big(\sqrt{pn}+p\big)
\end{aligned}
$$

Thus if $B = \Big\| \sup_{k\in\{0,\dots,n\}} A_k \Big\|_p$, we have

$$B \leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma R B^{1/2}\big(\sqrt{pn} + p\big)$$

By solving this quadratic inequality, we get:

$$\big(B^{1/2} - 2\gamma R(\sqrt{pn}+p)\big)^2 \leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma^2 R^2\big(\sqrt{pn}+p\big)^2$$

$$B^{1/2} \leqslant 2\gamma R(\sqrt{pn}+p) + \sqrt{\|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma^2 R^2\big(\sqrt{pn}+p\big)^2}$$

$$\begin{aligned} B &\leqslant 8\gamma^2 R^2(\sqrt{pn}+p)^2 + 2\|\theta_0 - \theta_*\|^2 + 2n\gamma^2 R^2 + 8\gamma^2 R^2\big(\sqrt{pn}+p\big)^2 \\ &\leqslant 16\gamma^2 R^2(\sqrt{pn}+p)^2 + 2\|\theta_0 - \theta_*\|^2 + 2n\gamma^2 R^2 \\ &\leqslant 17\gamma^2 R^2(\sqrt{pn}+p)^2 + 2\|\theta_0 - \theta_*\|^2 \end{aligned}$$

The previous statement leads to the desired result if $p \geqslant 2$. For $p \in [1,2]$, we may bound it by the value at $p = 2$, and a direct calculation shows that the bound is still correct. □

Proposition 3. *Assume **(B1-4)**. Consider the stochastic gradient recursion $\theta_n = \theta_{n-1} - \gamma \ell'(y_n, \langle \theta_{n-1}, x_n\rangle)x_n$ and its averaged version $\overline{\theta}_{n-1}$. If $\dfrac{\alpha e(10 + 2R^2\|\theta_0 - \theta_*\|^2)}{\sqrt{n}} \leqslant \dfrac{1}{2}$, then $\mathbb{E}\exp\big(\alpha\big[f(\overline{\theta}_{n-1}) - f(\theta_*)\big]\big) \leqslant 1$.*

Proof. Using that almost surely, $\|\overline{\theta}_{n-1} - \theta_*\| \leqslant \|\theta_0 - \theta_*\| + n\gamma R$ we obtain that almost surely $f(\overline{\theta}_{n-1}) - f(\theta_*) \leqslant R\|\theta_0 - \theta_*\| + n\gamma R^2$.

Moreover, from the previous proposition, we have for $p \leqslant \frac{n}{4}$,

$$\big\|f(\overline{\theta}_{n-1}) - f(\theta_*)\big\|_p \leqslant \frac{17\gamma R^2}{2}\frac{9}{4}p + \frac{1}{\gamma n}\|\theta_0 - \theta_*\|^2$$

For $\gamma = \frac{1}{2R^2\sqrt{n}}$, we get:

$$\big\|f(\overline{\theta}_{n-1}) - f(\theta_*)\big\|_p \leqslant \frac{10p}{\sqrt{n}} + \frac{2R^2}{\sqrt{n}}\|\theta_0 - \theta_*\|^2$$

and

$$f(\overline{\theta}_{n-1}) - f(\theta_*) \leqslant R\|\theta_0 - \theta_*\| + \frac{\sqrt{n}}{2} \text{ almost surely}$$

This leads to the bound valid for all p:

$$\big\|f(\overline{\theta}_{n-1}) - f(\theta_*)\big\|_p \leqslant \frac{p}{\sqrt{n}}(10 + 2R^2\|\theta_0 - \theta_*\|^2)$$

We then get

$$\begin{aligned} \mathbb{E}\exp\big(\alpha\big[f(\overline{\theta}_{n-1}) - f(\theta_*)\big]\big) &= \sum_{k=0}^{\infty} \frac{\alpha^p}{p!}\mathbb{E}[|f(\overline{\theta}_{n-1}) - f(\theta_*)|^p] \\ &\leqslant \sum_{k=0}^{\infty} \frac{\alpha^p}{p!}\frac{p^p}{n^{p/2}}\Big(10 + 2R^2\|\theta_0 - \theta_*\|^2\Big)^p \end{aligned}$$

$$\begin{aligned}
&\leqslant \sum_{k=0}^{\infty} \frac{\alpha^p}{2(p/e)^p} \frac{p^p}{n^{p/2}} \Big(10 + 2R^2 \|\theta_0 - \theta_*\|^2 \Big)^p \text{ using Stirling's formula,} \\
&\leqslant \frac{1}{2} \sum_{k=0}^{\infty} \frac{(e\alpha)^p}{n^{p/2}} \Big(10 + 2R^2 \|\theta_0 - \theta_*\|^2 \Big)^p \\
&\leqslant \frac{1}{2} \frac{1}{1-1/2} = 1 \text{ if } \frac{\alpha e (10 + 2R^2 \|\theta_0 - \theta_*\|^2)}{\sqrt{n}} \leqslant \frac{1}{2}
\end{aligned}$$

□

E Properties of self-concordance functions

In this section, we review various properties of self-concordant functions, that will prove useful in proving Theorem 3. All these properties rely on bounding the third-order derivatives by second-order derivatives. More precisely, from assumptions **(B3-4)**, we have for any $\theta, \delta, \eta \in \mathcal{H}$, where $f^{(r)}[\delta_1, \dots, \delta_k]$ denotes the k-th order differential of f:

$$\begin{aligned}
f'''(\theta)[\delta, \delta, \eta] &= \mathbb{E}\big[\ell'''(y_n, \langle \theta, x_n \rangle) \langle \delta, x_n \rangle^2 \langle \eta, x_n \rangle\big] \\
|f'''(\theta)[\delta, \delta, \eta]| &\leqslant \mathbb{E}\big[\ell''(y_n, \langle \theta, x_n \rangle) \langle \delta, x_n \rangle^2 |\langle \eta, x_n \rangle|\big] \\
&\leqslant \sqrt{\mathbb{E}\big[\ell''(y_n, \langle \theta, x_n \rangle)^2 \langle \delta, x_n \rangle^4\big]} \sqrt{\mathbb{E}\big[\langle \eta, x_n \rangle^2\big]} \text{ using Cauchy-Schwarz,} \\
&\leqslant \sqrt{\kappa\rho} f''(\theta)[\delta, \delta] \sqrt{\langle \eta, H\eta \rangle} \text{ using the two assumptions}
\end{aligned}$$

E.1 Global Taylor expansions

In this section, we derive global non-asymptotic Taylor expansions for self-concordant functions, which show that they behave similarly to like quadratic functions.

The following proposition shows that having a small excess risk $f(\theta) - f(\theta_*)$ implies that the weighted distance to optimum $\langle \theta - \theta_*, H(\theta - \theta_*) \rangle$ is small. Note that for quadratic functions, these two quantities are equal and that throughout this section, we always consider norms weighted by the matrix H (Hessian at optimum).

Proposition 4 (Bounding weighted distance to optimum from function values)**.** *Assume **(B3-4)**. Then, for any $\theta \in \mathcal{H}$:*

$$\langle \theta - \theta_*, H(\theta - \theta_*) \rangle \leqslant 3\big[f(\theta) - f(\theta_*)\big] + \kappa\rho\big[f(\theta) - f(\theta_*)\big]^2 \tag{33}$$

Proof. Let $\varphi : t \mapsto f\big[\theta_* + t(\theta - \theta_*)\big]$. Denoting $d = \sqrt{\langle \theta - \theta_*, f''(\theta_*)(\theta - \theta_*) \rangle}$, we have:

$$\begin{aligned}
|\varphi'''(t)| &\leqslant \mathbb{E}\big[\ell'''(y, \langle \theta_* + t(\theta - \theta_*), x \rangle) |\langle \theta - \theta_*, x \rangle|^3\big] \\
&\leqslant \mathbb{E}\big[\ell''(y, \langle \theta_* + t(\theta - \theta_*), x \rangle) \langle \theta - \theta_*, x \rangle^2\big] \sqrt{\kappa\rho} d = \sqrt{\kappa\rho} d \varphi''(t)
\end{aligned}$$

from which we obtain $\varphi''(t) \geqslant \varphi''(0) e^{-\sqrt{\kappa\rho} dt}$. Following [25], by integrating twice (and noting that $\varphi'(0) = 0$ and $\varphi''(0) = d^2$), we get

$$\begin{aligned}
f(\theta) = \varphi(1) &\geqslant \varphi(0) + \varphi''(0) \frac{1}{S^2 d^2} \big(e^{-\sqrt{\kappa\rho} d} + \sqrt{\kappa\rho} d - 1 \big) \\
&\geqslant f(\theta_*) + \frac{1}{\kappa\rho} \big(e^{-\sqrt{\kappa\rho} d} + \sqrt{\kappa\rho} d - 1 \big)
\end{aligned}$$

Thus

$$e^{-\sqrt{\kappa\rho}d} + \sqrt{\kappa\rho}d - 1 \leqslant \kappa\rho\big[f(\theta) - f(\theta_*)\big]$$

The function $\kappa : u \mapsto e^{-u} + u - 1$ is an increasing bijection from $\mathbb{R}_+$ to itself. Thus this implies $d \leqslant \frac{1}{\sqrt{\kappa\rho}}\kappa^{-1}\Big(\kappa\rho\big[f(\theta) - f(\theta_*)\big]\Big)$. We show below that $\kappa^{-1}(v) \leqslant \sqrt{3v + v^2}$, leading to the desired result.

The identity $\kappa^{-1}(v) \leqslant \sqrt{3v + v^2}$ is equivalent to $e^{-u} + u - 1 \geqslant \sqrt{u^2 + \alpha^2} - \alpha$, for $\alpha = \frac{3}{2}$. It then suffices to show that $1 - e^{-u} \geqslant \frac{u}{\sqrt{u^2+\alpha^2}}$. This can be shown by proving the monotonicity of $u \mapsto e^{-u} + u - 1 - \sqrt{u^2 + \alpha^2} + \alpha$, and we leave this exercise to the reader. □

The next proposition shows that Hessians between two points which are close in weighted distance are close to each other, for the order between positive semi-definite matrices.

Proposition 5 (Expansion of Hessians)**.** *Assume **(B3-4)**. Then, for any $\theta_1, \theta_2 \in \mathcal{H}$:*

$$f''(\theta_1)e^{\sqrt{\kappa\rho}\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}} \succcurlyeq f''(\theta_2) \succcurlyeq f''(\theta_1)e^{-\sqrt{\kappa\rho}\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}} \tag{34}$$

$$\big\|f''(\theta_1)^{-1/2}f''(\theta_2)f''(\theta_1)^{-1/2} - I\big\|_{\mathrm{op}} \leqslant e^{\sqrt{\kappa\rho}\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}} - 1 \tag{35}$$

Proof. Let $z \in \mathcal{H}$ and $\psi(t) = z^\top f''(\theta_1 + t(\theta_2 - \theta_1))z$. We have:

$$\begin{aligned} |\psi'(t)| &= |f'''(\theta_1 + t(\theta_2 - \theta_1))[z, z, \theta_2 - \theta_1]| \\ &\leqslant f''(\theta_1 + t(\theta_2 - \theta_1))[z, z]\sqrt{\kappa\rho}d = \psi(t)\sqrt{\kappa\rho}d \end{aligned}$$

with $d_{12} = \sqrt{\langle\theta_2 - \theta_1, H(\theta_2 - \theta_1)\rangle}$. Thus $\psi(0)e^{\sqrt{\kappa\rho}d_{12}t} \geqslant \psi(t) \geqslant \psi(0)e^{-\sqrt{\kappa\rho}d_{12}t}$. This implies, for $t = 1$, that

$$f''(\theta_1)e^{\sqrt{\kappa\rho}d_{12}} \succcurlyeq f''(\theta_2) \succcurlyeq f''(\theta_1)e^{-\sqrt{\kappa\rho}d_{12}}$$

which implies the desired results. $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm (largest singular value). □

The following proposition gives an approximation result bounding the first order expansion of gradients by the first order expansion of function values.

Proposition 6 (Expansion of gradients)**.** *Assume **(B3-4)**. Then, for any $\theta_1, \theta_2 \in \mathcal{H}$ and $\Delta \in \mathcal{H}$:*

$$\langle\Delta, f'(\theta_2) - f'(\theta_1) - f''(\theta_1)(\theta_2 - \theta_1)\rangle \leqslant \sqrt{\kappa\rho}\langle\Delta, H\Delta\rangle^{1/2}\big[f(\theta_2) - f(\theta_1) - \langle f'(\theta_1), \theta_2 - \theta_1\rangle\big] \tag{36}$$

Proof. Let $\varphi(t) = \langle\Delta, f'(\theta_1 + t(\theta_2 - \theta_1)) - f'(\theta_1) - tf''(\theta_1)(\theta_2 - \theta_1)\rangle$. We have $\varphi'(t) = \langle\Delta, f''(\theta_1 + t(\theta_2 - \theta_1))(\theta_1 - \theta_2\rangle - \langle\Delta, f''(\theta_1)(\theta_1 - \theta_2)\rangle$ and $|\varphi''(t)| = |f'''(\theta_1 + t(\theta_2 - \theta_1))[\theta_2 - \theta_1, \theta_2 - \theta_1, \Delta]| \leqslant \sqrt{\kappa\rho}\langle\Delta, H\Delta\rangle^{1/2}\langle\theta_1 - \theta_2, f''(\theta_1 + t(\theta_2 - \theta_1))(\theta_1 - \theta_2)\rangle$. This leads to

$$\langle\Delta, f'(\theta_2) - f'(\theta_1) - f''(\theta_1)(\theta_2 - \theta_1)\rangle \leqslant \sqrt{\kappa\rho}\langle\Delta, H\Delta\rangle^{1/2}\big[f(\theta_2) - f(\theta_1) - \langle f'(\theta_1), \theta_2 - \theta_1\rangle\big]$$

Note that one may also use the bound

$$|\varphi'(t)| \leqslant \|\theta_1 - \theta_2\|\langle\Delta, f''(\theta_1)^2\Delta\rangle^{1/2}\big[e^{t\sqrt{\kappa\rho}\|H^{1/2}(\theta_1-\theta_2)\|} - 1\big]$$

leading to

$$\begin{aligned} &\langle\Delta, f'(\theta_2) - f'(\theta_1) - f''(\theta_1)(\theta_2 - \theta_1)\rangle \\ \leqslant\ &\|\theta_1 - \theta_2\|\langle\Delta, f''(\theta_1)^2\Delta\rangle^{1/2}\frac{e^{\sqrt{\kappa\rho}\|H^{1/2}(\theta_1-\theta_2)\|} - 1 - \sqrt{\kappa\rho}\|H^{1/2}(\theta_1 - \theta_2)\|}{\sqrt{\kappa\rho}\|H^{1/2}(\theta_1 - \theta_2)\|} \end{aligned} \tag{37}$$

□

The following proposition considers a global Taylor expansion of function values. Note that when $\kappa\rho\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle$ tends to zero, we obtain *exactly* the second-order Taylor expansion. For more details, see [25]. This is followed by a corrolary that upper bounds excess risk by distance to optimum (this is thus the other direction than Prop. 4).

Proposition 7 (Expansion of function values)**.** *Assume **(B3-4)**. Then, for any $\theta_1, \theta_2 \in \mathcal{H}$ and $\Delta \in \mathcal{H}$:*

$$\begin{aligned} & f(\theta_2) - f(\theta_1) - \langle f'(\theta_1), \theta_2 - \theta_1\rangle \\ \leqslant\; & \langle \theta_2 - \theta_1, f''(\theta_1)(\theta_2-\theta_1)\rangle \frac{e^{\sqrt{\kappa\rho}\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}} - 1 - \sqrt{\kappa\rho}\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}}{\kappa\rho\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle} \end{aligned} \tag{38}$$

Proof. Let $\varphi(t) = f\big[\theta_1 + t(\theta_2-\theta_1)\big] - f(\theta_1) - t\langle f'(\theta_1), \theta_2-\theta_1\rangle$. We have $\varphi'(t) = \langle f'\big[\theta_1 + t(\theta_2 - \theta_1)\big], \theta_2 - \theta_1\rangle - \langle f'(\theta_1), \theta_2 - \theta_1\rangle$ and $\varphi''(t) = \langle \theta_2 - \theta_1, f''\big[\theta_1 + t(\theta_2-\theta_1)\big](\theta_2-\theta_1)\rangle$. Moreover, $\varphi'''(t) \leqslant \sqrt{\kappa\rho}\varphi''(t)\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}$, leading to $\varphi''(t) \leqslant e^{\sqrt{\kappa\rho}t\sqrt{\langle\theta_2-\theta_1, H(\theta_2-\theta_1)\rangle}}\varphi''(0)$. Integrating twice between 0 and 1 leads to the desired result. □

Corollary 2 (Excess risk)**.** *Assume **(B3-4)**, and $\theta_1 \in \mathcal{H}$ and $\theta_2 = \theta_1 - f''(\theta_1)^{-1}f'(\theta_1)$. Then*

$$f(\theta) - f(\theta^*) \leqslant \frac{e^{\sqrt{\kappa\rho}d} - \sqrt{\kappa\rho}d - 1}{\kappa\rho} \tag{39}$$

where $d = \sqrt{\langle \theta - \theta_*, H(\theta-\theta_*)\rangle}$.

Proof. Applying Prop. 7 to $\theta_2 = \theta$ and $\theta_1 = \theta_*$, we get the desired result. □

The following proposition looks at a similar type of bounds than Prop. 7; it is weaker when θ_2 and θ_1 are close (it does not converge to the second-order Taylor expansion), but stronger for large values (it does not grow exponentially fast).

Proposition 8 (Bounding function values with fewer assumptions)**.** *Assume **(B3-4)**, and $\theta 1, \theta_2 \in \mathcal{H}$. Then*

$$f(\theta_2) - f(\theta_1) \leqslant \sqrt{\rho}\|H^{1/2}(\theta_1 - \theta_2)\| \tag{40}$$

Proof. Let $\varphi(t) = f(\theta_1 + t(\theta_2 - \theta_1)) - f(\theta_1)$. We have $|\varphi'(t)| = |\mathbb{E}\ell'(y_n, \langle x_n, \theta 1 + t(\theta_2 - \theta_1)\rangle)\langle \theta_2 - \theta_1 t, x_n\rangle| \leqslant \sqrt{\rho}\|H^{1/2}(\theta_1 - \theta_2)\|$. Integrating between 0 and 1 leads to the desired result. □

E.2 Analysis of Newton step

Self-concordance has been traditionally used in the analysis of Newton's method (see [29, 24]). In this section, we adapt classical results to our specific notion of self-concordance (see also [25]). A key quantity is the so-called "Newton decrement" at a certain point θ_1, equal to $\langle f'(\theta_1), f''(\theta_1)^{-1}f'(\theta_1)\rangle$, which governs the convergence behavior of Newton methods (this is the quantity which is originally shown to be quadratically convergent). In this paper, we consider a slightly different version where the Hessian is chosen to be the one at θ_*, i.e., $\langle f'(\theta_1), H^{-1}f'(\theta_1)\rangle$.

The following proposition shows how a full Newton step improves the Newton decrement (by taking a square).

Proposition 9 (Effect of Newton step on Newton decrement)**.** *Assume* ***(B3-4)****, and $\theta_1 \in \mathcal{H}$ and $\theta_2 = \theta_1 - f''(\theta_1)^{-1} f'(\theta_1)$. Then*

$$\langle f'(\theta_2), H^{-1} f'(\theta_2)\rangle \leqslant \kappa\rho e^{2\sqrt{\kappa\rho}d_1} \langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle^2 \bigg(\frac{e^{\sqrt{\kappa\rho}d_{12}} - \sqrt{\kappa\rho}d_{12} - 1}{\kappa\rho d_{12}^2} \bigg)^2 \tag{41}$$

where $d_{12}^2 = \langle \theta_2 - \theta_1, H(\theta_2 - \theta_1)\rangle \leqslant e^{\sqrt{\kappa\rho}d_1} \langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle$ and $d_1 = \langle \theta_1 - \theta_*, H(\theta_1 - \theta_*)\rangle^{1/2}$.

Proof. When applying the two previous propositions to the Newton step $\theta_2 = \theta_1 - f''(\theta_1)^{-1} f'(\theta_1)$, we get:

$$\begin{aligned}\langle \Delta, f'(\theta_2)\rangle &\leqslant \sqrt{\kappa\rho} \langle \Delta, H\Delta\rangle^{1/2} \langle f'(\theta_1), f''(\theta_1)^{-1} f'(\theta_1)\rangle \frac{e^{\sqrt{\kappa\rho}d_{12}} - \sqrt{\kappa\rho}d_{12} - 1}{\kappa\rho d_{12}^2} \\ &\leqslant \sqrt{\kappa\rho} S e^{\sqrt{\kappa\rho}d_1} \langle \Delta, H\Delta\rangle^{1/2} \langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle \frac{e^{\sqrt{\kappa\rho}d_{12}} - \sqrt{\kappa\rho}d_{12} - 1}{\kappa\rho d_{12}^2}\end{aligned}$$

We then optimize with respect to Δ to obtain the desired result. □

The following proposition shows how the Newton decrement is upper bounded by a function of the excess risk.

Proposition 10 (Newton decrement)**.** *Assume* ***(B3-4)****, and $\theta_1 \in \mathcal{H}$, then,*

$$\langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle \leqslant \bigg(\frac{1}{2}\sqrt{\kappa\rho}\Delta_1 + \sqrt{d_1^2 + \sqrt{\kappa\rho}d_1\Delta_1 + \frac{1}{4}\kappa\rho\Delta_1^2} \bigg)^2 \tag{42}$$

with $d_1 = \sqrt{\langle \theta_1 - \theta_*, H(\theta_1 - \theta_*)\rangle}$ and $\Delta_1 = f(\theta_1) - f(\theta_*)$.

Proof. We may bound the Newton decrement as follows:

$$\begin{aligned}\langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle &\leqslant \langle f'(\theta_1) - H(\theta_1 - \theta_*), H^{-1} f'(\theta_1)\rangle + \langle H(\theta_1 - \theta_*), H^{-1} f'(\theta_1)\rangle \\ &\leqslant \sqrt{\kappa\rho}\big[f(\theta_1) - f(\theta_*)\big] \langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle^{1/2} + \langle f'(\theta_1), \theta_1 - \theta_*\rangle \qquad (43)\end{aligned}$$

This leads to

$$\langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle \leqslant \bigg(\frac{1}{2}\sqrt{\kappa\rho}\big[f(\theta_1) - f(\theta_*)\big] + \sqrt{\langle f'(\theta_1), \theta_1 - \theta_*\rangle + \frac{1}{4}\kappa\rho\big[f(\theta_1) - f(\theta_*)\big]^2} \bigg)^2$$

Moreover,

$$\begin{aligned}\langle f'(\theta_1), \theta_1 - \theta_*\rangle &= \langle H(\theta_1 - \theta_*), \theta_1 - \theta_*\rangle + \langle f'(\theta_1) - H(\theta_1 - \theta_*), \theta_1 - \theta_*\rangle \\ &\leqslant \langle H(\theta_1 - \theta_*), \theta_1 - \theta_*\rangle + \sqrt{\kappa\rho}\langle \theta_1 - \theta_*, H(\theta_1 - \theta_*)\rangle^{1/2} \big[f(\theta_1) - f(\theta_*)\big] \\ &\leqslant d_1^2 + \sqrt{\kappa\rho}d_1\big[f(\theta_1) - f(\theta_*)\big]\end{aligned}$$

Overall, we get

$$\begin{aligned}&\langle f'(\theta_1), H^{-1} f'(\theta_1)\rangle \\ &\leqslant \bigg(\frac{1}{2}\sqrt{\kappa\rho}\big[f(\theta_1) - f(\theta_*)\big] + \sqrt{d_1^2 + \sqrt{\kappa\rho}d_1\big[f(\theta_1) - f(\theta_*)\big] + \frac{1}{4}\kappa\rho\big[f(\theta_1) - f(\theta_*)\big]^2} \bigg)^2\end{aligned}$$

□

The following proposition provides a bound on a quantity which is not the Newton decrement. Indeed, this is (up to the difference in the Hessians), the norm of the Newton step. This will be key in the following proofs.

Proposition 11 (Bounding gradients from unweighted distance to optimum)**.** *Assume* ***(B3-4)****, and $\theta_1 \in \mathcal{H}$, then,*

$$\|H^{-1} f'(\theta_1)\| \leqslant \frac{e^{\sqrt{\kappa\rho} d_1} - 1}{\sqrt{\kappa\rho} d_1} \|\theta_1 - \theta_*\| \tag{44}$$

with $d_1 = \sqrt{\langle \theta_1 - \theta_*, H(\theta_1 - \theta_*)\rangle}$.

Proof. We have:

$$\begin{aligned}\|H^{-1} f'(\theta_1)\| &\leqslant \|H^{-1}\big[f'(\theta_1) - H(\theta_1 - \theta_*)\big]\| + \|H^{-1}\big[H(\theta_1 - \theta_*)\big]\| \\ &\leqslant \|\theta_1 - \theta_*\| \bigg(1 + \frac{e^{\sqrt{\kappa\rho} d_1} - 1 - \sqrt{\kappa\rho} d_1}{\sqrt{\kappa\rho} d_1}\bigg)\end{aligned}$$

□

The next proposition shows that having a small Newton decrement implies that the weighted distance to optimum is small.

Proposition 12 (Weighted distance to optimum)**.** *Assume* ***(B3-4)****. If we have $\sqrt{\kappa\rho} e^{\sqrt{\kappa\rho} d} \langle f'(\theta), H^{-1} f'(\theta)\rangle^{1/2} \leqslant \frac{1}{2}$, with $d = \sqrt{\langle \theta - \theta_*, H(\theta - \theta_*)\rangle}$, then*

$$d \leqslant 4 e^{\sqrt{\kappa\rho} d} \langle f'(\theta), H^{-1} f'(\theta)\rangle^{1/2}$$

Proof. For any $\Delta \in \mathcal{H}$ such that $\langle \Delta, H\Delta\rangle = 1$, and $t \geqslant 0$, we have, following the same reasoning than for Prop. 7:

$$\begin{aligned}f(\theta + t\Delta) &\geqslant f(\theta) + t\langle \Delta, f'(\theta)\rangle + \langle \Delta, f''(\theta)\Delta\rangle \frac{e^{-vt} + vt - 1}{v^2} \\ &\geqslant f(\theta) + \frac{\langle \Delta, f''(\theta)\Delta\rangle}{v^2}\Big[e^{-vt} - 1 + tv\big(1 - s\big)\Big]\end{aligned}$$

with $v = \sqrt{\kappa\rho}\sqrt{\langle \Delta, H\Delta\rangle} = \sqrt{\kappa\rho}$ and

$$s = \frac{v|\langle \Delta, f'(\theta)\rangle|}{\langle \Delta, f''(\theta)\Delta\rangle} \leqslant \frac{\sqrt{\kappa\rho}\langle f'(\theta), f''(\theta)^{-1} f'(\theta)\rangle^{1/2}}{\langle \Delta, f''(\theta)\Delta\rangle^{1/2}} \leqslant \sqrt{\kappa\rho} e^{\sqrt{\kappa\rho} d} \langle f'(\theta), H^{-1} f'(\theta)\rangle^{1/2}$$

It is shown in [25] that if $s \in [0, 1)$, then

$$e^{-2s/(1-s)} + (1 - s) 2s(1 - s)^{-1} - 1 \geqslant 0$$

This implies that if $s \leqslant 1/2$, for $t = \frac{2\sqrt{\kappa\rho}^{-1} s}{1-s}$, $f(\theta_2 + t\Delta) \geqslant f(\theta_2)$. Thus,

$$d = \sqrt{\langle \theta - \theta_*, H(\theta - \theta_*)\rangle} \leqslant t \leqslant 4\sqrt{\kappa\rho}^{-1} s \leqslant 4 e^{\sqrt{\kappa\rho} d} \langle f'(\theta), H^{-1} f'(\theta)\rangle^{1/2} \tag{45}$$

Note that the quantity d appears twice in the result above. □

E.3 Proof of Prop. 1

In this section, we prove Prop. 1 using tools from self-concordance analysis. These tools are described in the previous Sections E.1 and E.2. In order to understand the proof, it is preferable to read these sections first.

We use the notation $t^2 = \kappa\rho\varepsilon_1$. We then get $d_1^2 \overset{\text{def}}{=} \langle \theta_1 - \theta_*, H(\theta_1 - \theta_*) \rangle \leqslant (3 + t^2)\varepsilon_1$ from Prop. 4.

Proof of Eq. (25). We have, from Prop. 8,

$$\begin{aligned} f(\theta_3) - f(\theta_*) &\leqslant f(\theta_2) - f(\theta_*) + \sqrt{\rho}\|H^{1/2}(\theta_3 - \theta_2)\| \\ &\leqslant f(\theta_2) - f(\theta_*) + \sqrt{2\rho\varepsilon_2}e^{\sqrt{\kappa\rho}d_1/2} \end{aligned} \qquad (46)$$

Moreover, we have, also from Prop. 8, $f(\theta_2) - f(\theta_*) \leqslant \sqrt{\rho}\|H^{1/2}f''(\theta_1)^{-1}f'(\theta_1)\|$, and using Prop. 5, we get

$$f(\theta_2) - f(\theta_*) \leqslant e^{\sqrt{\kappa\rho}d_1}\sqrt{\rho}\|H^{-1/2}f'(\theta_1)\| \qquad (47)$$

We may now use Prop. 10 and use the bound:

$$\begin{aligned} \langle f'(\theta_1), H^{-1}f'(\theta_1) \rangle &\leqslant \left(\frac{1}{2}\sqrt{\kappa\rho}\varepsilon_1 + \sqrt{(3+t^2)\varepsilon_1 + \sqrt{\kappa\rho}\sqrt{(3+t^2)\varepsilon_1}\varepsilon_1 + \frac{1}{4}\kappa\rho\varepsilon_1^2} \right)^2 \\ &\leqslant \left(\frac{1}{2}t\sqrt{\varepsilon_1} + \sqrt{(3+t^2)\varepsilon_1 + t\sqrt{(3+t^2)}\varepsilon_1 + \frac{1}{4}t^2\varepsilon_1} \right)^2 \\ &= \left(\frac{1}{2}t + \sqrt{(3+t^2) + t\sqrt{(3+t^2)} + \frac{1}{4}t^2} \right)^2 \varepsilon_1 \overset{\text{def}}{=} \Box_1(t)^2\varepsilon_1 \end{aligned} \qquad (48)$$

A simple plot shows that for all $t > 0$,

$$\Box_1(t) = \frac{1}{2}t + \sqrt{(3+t^2) + t\sqrt{(3+t^2)} + \frac{1}{4}t^2} \leqslant \sqrt{3} + 2t \qquad (49)$$

Combining with Eq. (46) and Eq. (47), we get

$$f(\theta_3) - f(\theta_*) \leqslant e^{\sqrt{3+t^2}t}\sqrt{\rho\varepsilon_1}(\sqrt{3} + 2t) + \sqrt{2\rho\varepsilon_2}e^{\sqrt{3+t^2}t/2}$$

which is exactly Eq. (25).

Proof of Eq. (26) and Eq. (27). For these two inequalities, the starting point is the same. Using Eq. (48) (i.e., the Newton decrement at θ_1), we first show that the distances d_{12} and d_2 are bounded. Using $f''(\theta_1) \succcurlyeq e^{-\sqrt{\kappa\rho}d_1}H$ (Prop. 5):

$$d_{12}^2 \leqslant e^{\sqrt{\kappa\rho}d_1}\langle f'(\theta_1), H^{-1}f'(\theta_1) \rangle \leqslant e^{t\sqrt{3+t^2}}\Box_1(t)^2\varepsilon_1 \overset{\text{def}}{=} \Box_2(t)^2\varepsilon_1$$

and thus

$$d_2 \leqslant d_1 + d_{12} \leqslant \left[\sqrt{3+t^2} + \Box_2(t) \right] \sqrt{\varepsilon_1}$$

Now, we can bound the Newton decrement at θ_2, using Prop. 9:

$$\langle f'(\theta_2), H^{-1}f'(\theta_2) \rangle \leqslant \kappa\rho e^{2\sqrt{\kappa\rho}d_1}\langle f'(\theta_1), H^{-1}f'(\theta_1) \rangle^2 \left(\frac{e^{\sqrt{\kappa\rho}d_{12}} - \sqrt{\kappa\rho}d_{12} - 1}{(\sqrt{\kappa\rho}d_{12})^2} \right)^2$$

$$\leqslant \kappa\rho\varepsilon_1^2\square_2(t)^4\bigg(\frac{e^{t\square_2(t)} - t\square_2(t) - 1}{(t\square_2(t))^2}\bigg)^2 \overset{\text{def}}{=} \square_3(t)\kappa\rho\varepsilon_1^2$$

Thus, using Prop. 12, if $\kappa\rho e^{2\sqrt{\kappa\rho}d_2}\langle f'(\theta_2), H^{-1}f'(\theta_2)\rangle \leqslant t^4 e^{2t[\sqrt{3+t^2}+\square_2(t)]}\square_3(t) \leqslant \frac{1}{4}$, then

$$d_2 \leqslant 4e^{\sqrt{\kappa\rho}d_2}\sqrt{\square_3(t)\kappa\rho\varepsilon_1^2} \leqslant 4e^{t[\sqrt{3+t^2}+\square_2(t)]}\sqrt{\square_3(t)\kappa\rho\varepsilon_1^2} \overset{\text{def}}{=} \square_4(t)\sqrt{\kappa\rho\varepsilon_1^2}$$

We then have

$$\begin{aligned}
d_3 = \sqrt{\langle \theta_3 - \theta_*, H(\theta_3 - \theta_*)\rangle} &\leqslant \sqrt{\langle \theta_3 - \theta_2, H(\theta_3 - \theta_2)\rangle} + \sqrt{\langle \theta_2 - \theta_*, H(\theta_2 - \theta_*)\rangle} = d_{23} + d_2 \\
&\leqslant \square_4(t)\sqrt{\kappa\rho\varepsilon_1^2} + \sqrt{2\varepsilon_2}e^{t\sqrt{3+t^2}/2} \\
d_3\sqrt{\kappa\rho} &\leqslant \square_4(t)t^2 + \sqrt{2\varepsilon_2\kappa\rho}e^{t\sqrt{3+t^2}/2} \leqslant \square_4(t)t^2 + \sqrt{2u^2}e^{t\sqrt{3+t^2}/2}
\end{aligned}$$

where $\varepsilon_2\kappa\rho \leqslant u^2$.

We now have two separate paths to obtain Eq. (26) and Eq. (27).

If we assume that ε_2 is bounded, i.e., with $t = 1/16$ and $u = 1/4$, then, one can check computationally that we obtain $d_3\sqrt{\kappa\rho} \leqslant 0.41$ and thus $b = 0.576$ below:

$$\begin{aligned}
f(\theta_3) - f(\theta^*) &\leqslant \frac{e^{\sqrt{\kappa\rho}d_3} - \sqrt{\kappa\rho}d_3 - 1}{\kappa\rho} \text{ using Prop. 2,} \\
&\leqslant d_3^2 \max_{\alpha\in[0,0.41} \frac{e^\alpha - \alpha - 1}{\alpha^2} \leqslant 0.576\big(\square_4(t)\sqrt{\kappa\rho\varepsilon_1^2} + \sqrt{2\varepsilon_2}e^{t\sqrt{3+t^2}/2}\big)^2 \\
&\leqslant 0.576(1+1/c)\square_4(t)^2\kappa\rho\varepsilon_1^2 + 2\times 0.576(1+c)e^{t\sqrt{3+t^2}}\varepsilon_2 \\
&\leqslant 57\kappa\rho\varepsilon_1^2 + 12\varepsilon_2, \text{ with } c = 8.1
\end{aligned}$$

which is exactly Eq. (27).

If we only assume ε_1 bounded, then we have (from the beginning of the proof):

$$\begin{aligned}
f(\theta_3) - f(\theta_*) &\leqslant f(\theta_2) - f(\theta_*) + \sqrt{\rho}\|H^{1/2}(\theta_3 - \theta_2)\| \\
&\leqslant 57\kappa\rho\varepsilon_1^2 + \sqrt{2\rho\varepsilon_2}e^{t\sqrt{3+t^2}/2} \leqslant 57\kappa\rho\varepsilon_1^2 + 2\sqrt{\rho\varepsilon_2}
\end{aligned}$$

because we may use the earlier reasoning with $\varepsilon_3 = 0$. This is Eq. (26).

F Additional experiments

In Table 2, we describe the datasets we have used in experiments and where they were downloaded from.

In Figure 4, we provide similar results than in Section 4, for two additional datasets, *quantum* and *rcv1*, while in Figure 5, Figure 6 and Figure 7, we provide training objectives for all methods. We can make the following observations:

- For non-sparse datasets, SAG manages to get the smallest training error, confirming the results of [27].
- For the high-dimensional sparse datasets, constant step-size SGD is performing best (note that as shown in Section 3, it is not converging to the optimal value in general, this happens notably for the *alpha* dataset).

Table 2: Datasets used in our experiments . We report the proportion of non-zero entries.

Name	d	n	sparsity	
quantum	79	50 000	100 %	osmot.cs.cornell.edu/kddcup/
covertype	55	581 012	100 %	www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
alpha	501	500 000	100 %	ftp://largescale.ml.tu-berlin.de/largescale/
sido	4 933	12 678	10 %	www.causality.inf.ethz.ch/
rcv1	47 237	20 242	0.2 %	www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
news	1 355 192	19 996	0.03 %	www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

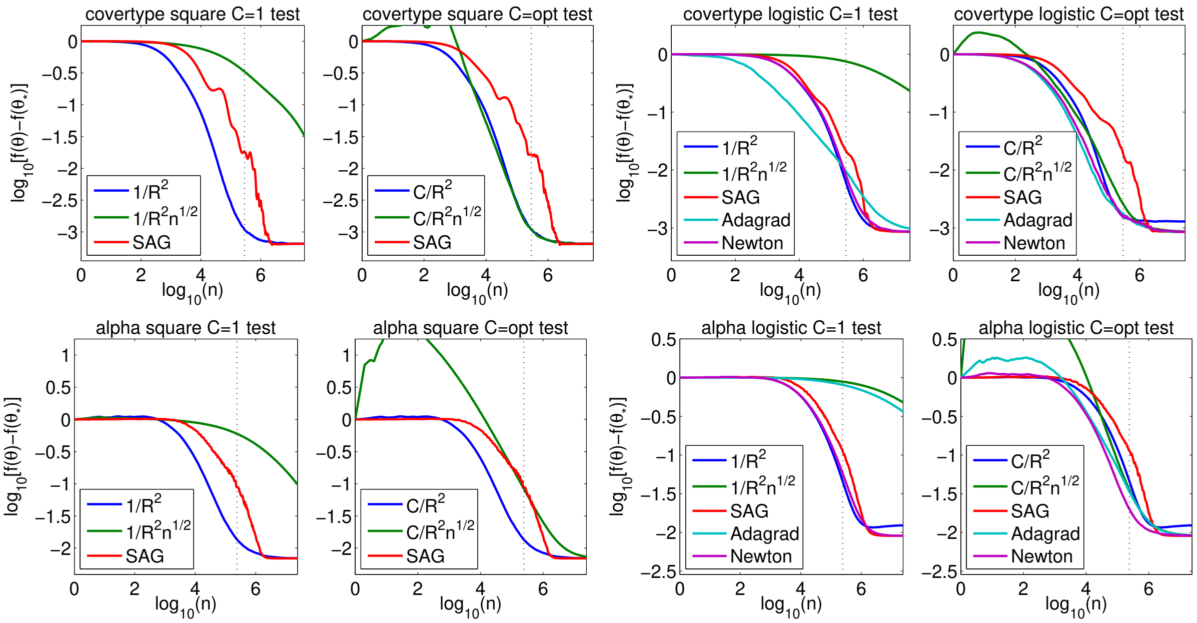

Figure 2. Test performance for least-square regression (two left plots) and logistic regression (two right plots). From top to bottom: *covertype*, *alpha*. Left: theoretical steps, right: steps optimized for performance after one effective pass through the data. Best seen in color.

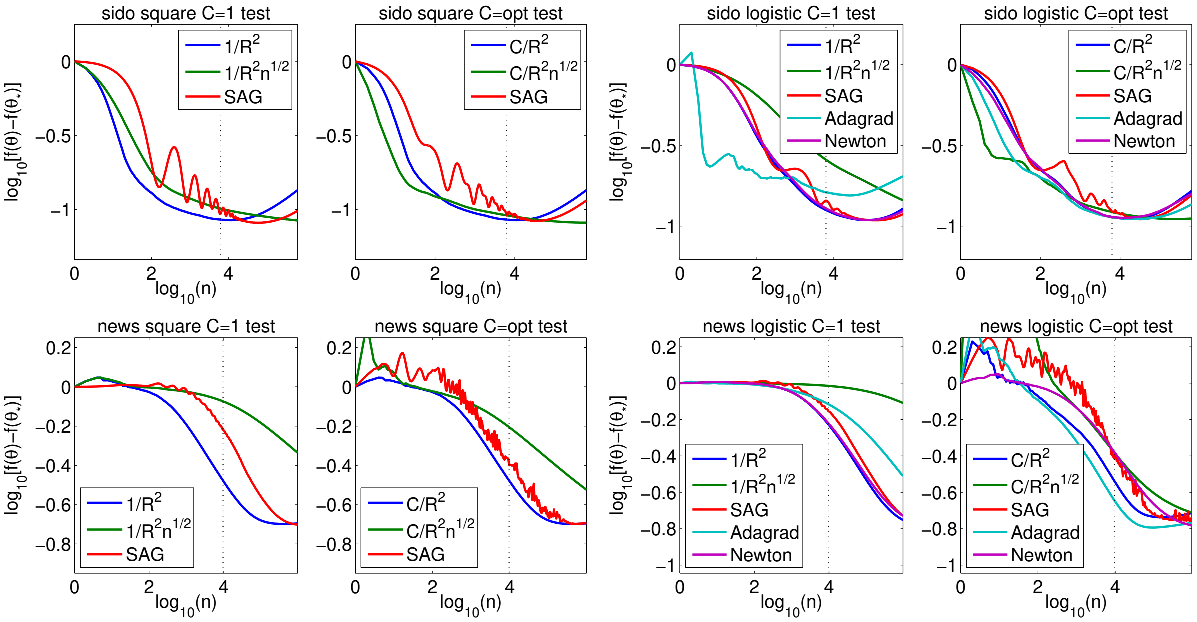

Figure 3. Test performance for least-square regression (two left plots) and logistic regression (two right plots). From top to bottom: *sido*, *news*. Left: theoretical steps, right: steps optimized for performance after one effective pass through the data. Best seen in color.

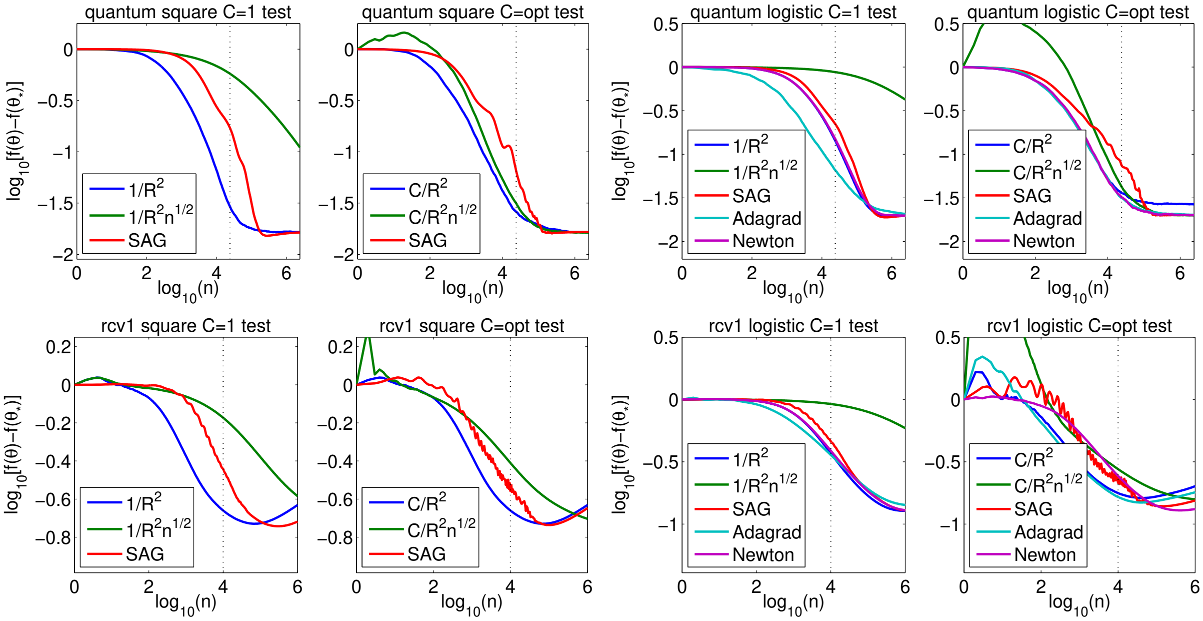

Figure 4. Test performance for least-square regression (two left plots) and logistic regression (two right plots). From top to bottom: *quantum*, *rcv1*. Left: theoretical steps, right: steps optimized for performance after one effective pass through the data.

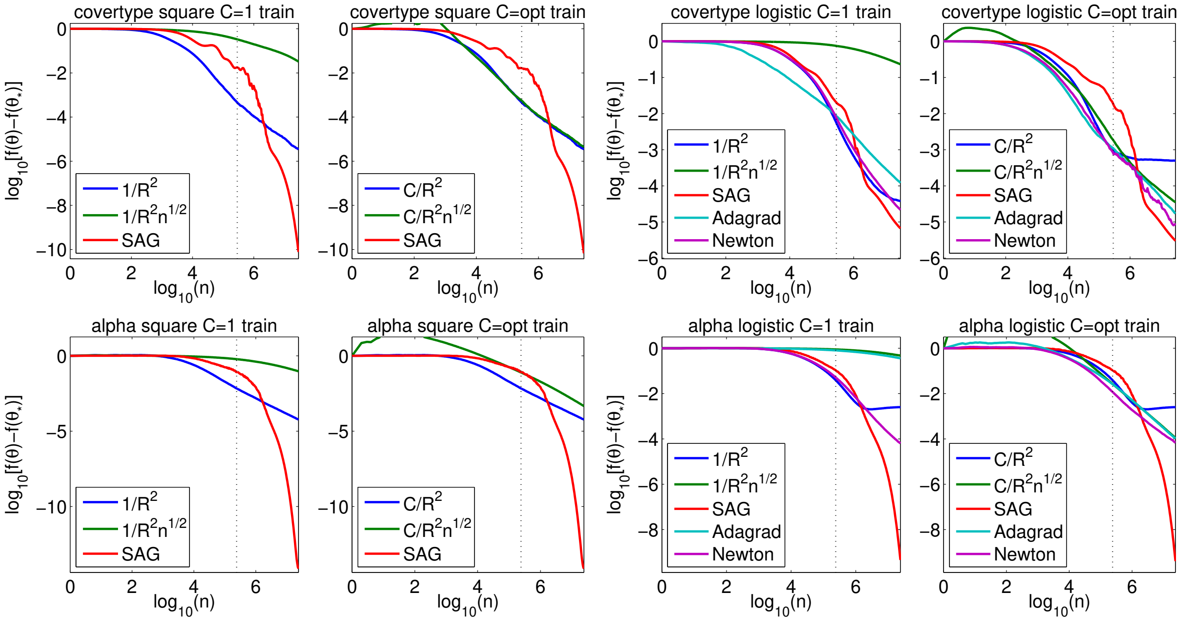

Figure 5. Training objective for least square regression (two left plots) and logistic regression (two right plots). From top to bottom: *covertype*, *alpha*. Left: theoretical steps, right: steps optimized for performance after one effective pass through the data.

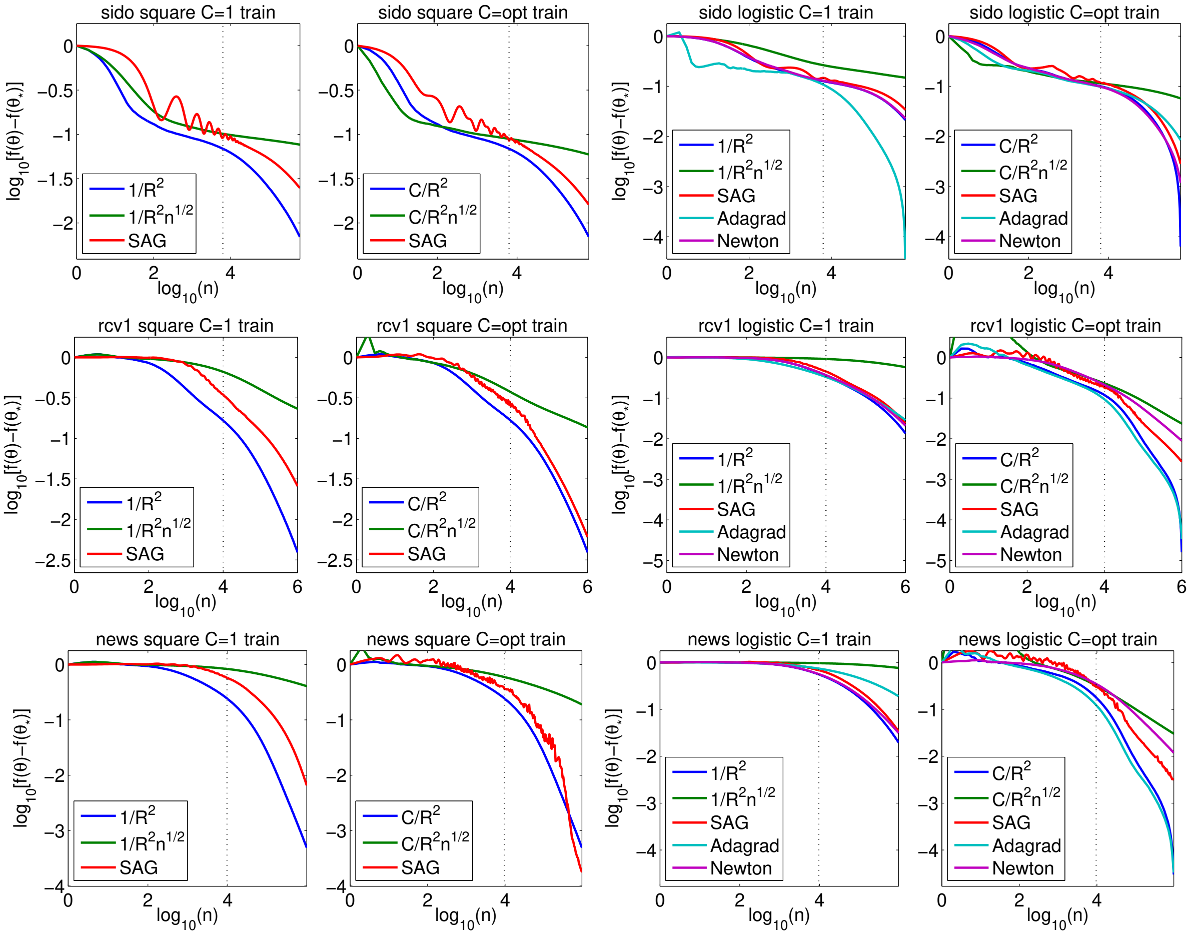

Figure 6. Training objective for least square regression (two left plots) and logistic regression (two right plots). From top to bottom: *sido*, *news*. Left: theoretical steps, right: steps optimized for performance after one effective pass through the data.

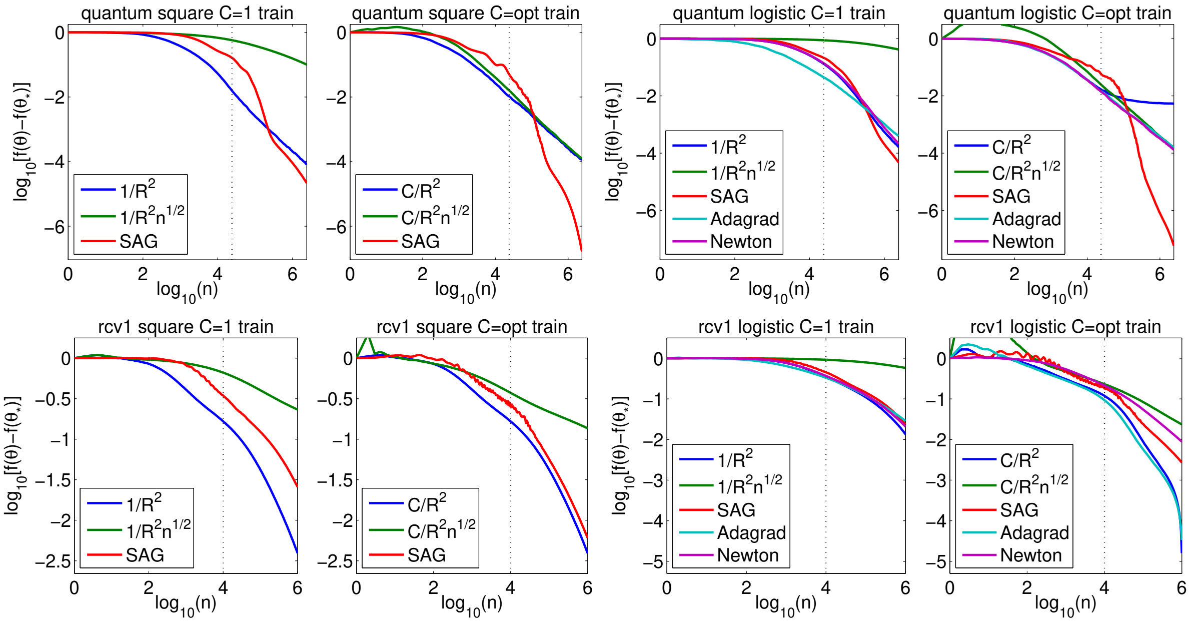

Figure 7. Training objective for least square regression (two left plots) and logistic regression (two right plots). From top to bottom: *quantum*, *rcv1*. Left: theoretical steps, right: steps optimized for performance after one effective pass through the data.