Large deviations rates for stochastic gradient descent with strongly convex functions

Dragana Bajovic
dbajovic@uns.ac.rs
Faculty of Technical Sciences
University of Novi Sad, Serbia

Dusan Jakovetic
dusan.jakovetic@dmi.uns.ac.rs
Faculty of Sciences,
University of Novi Sad, Serbia

Soummya Kar
soummyak@andrew.cmu.edu
Carnegie Mellon University, USA

Abstract

Recent works have shown that high probability metrics with stochastic gradient descent (SGD) exhibit informativeness and in some cases advantage over the commonly adopted mean-square error-based ones. In this work we provide a formal framework for the study of general high probability bounds with SGD, based on the theory of large deviations. The framework allows for a generic (not-necessarily bounded) gradient noise satisfying mild technical assumptions, allowing for the dependence of the noise distribution on the current iterate. Under the preceding assumptions, we find an upper large deviations bound for SGD with strongly convex functions. The corresponding rate function captures analytical dependence on the noise distribution and other problem parameters. This is in contrast with conventional mean-square error analysis that captures only the noise dependence through the variance and does not capture the effect of higher order moments nor interplay between the noise geometry and the shape of the cost function. We also derive exact large deviation rates for the case when the objective function is quadratic and show that the obtained function matches the one from the general upper bound hence showing the tightness of the general upper bound. Numerical examples illustrate and corroborate theoretical findings.

1 INTRODUCTION

The large deviations theory represents a well-established principled approach for studying *rare events* that occur with stochastic processes, e.g., (Dembo et al. 1993). Typically, we are concerned with a sequence of rare events E_k related with the stochastic process of interest, indexed by, e.g., time k. In this setting, the probability of event E_k, $k = 1, 2, ...$ typically decays exponentially in k; the large deviations theory then enables to quantify this exponential rate. Such an approach has found many applications in statistics (Bucklew 1990), mechanics (Touchette 2009), communications (Shwartz et al. 1995), and information theory (Cover et al. 1991).

To be more concrete, consider an example of a sequence of random vectors X_k taking values in $\mathbb{R}^d$ that converge, e.g., almost surely, to a (deterministic) limit point $x^\star \in \mathbb{R}^d$. The rare event of interest E_k can then be, for example, $E_k = \{\|X_k - x^\star\| \geq \delta\}$, for some positive quantity δ, with $\|\cdot\|$ denoting the Euclidean norm. Equivalently, E_k can be represented as $\{X_k \in C_\delta\}$, where C_δ is the complement of the l_2 ball of radius δ centered at $x^\star$. Large deviations analysis then aims at discovering the corresponding rate of decay, i.e., the inaccuracy rate $\mathbf{I}(C_\delta)$:

$$\mathbb{P}\left(X_k \in C_\delta\right) = e^{-k\,\mathbf{I}(C_\delta)+o(k)}, \tag{1}$$

where $o(k)$ denotes terms growing slower than linearly with k. The inaccuracy rate $\mathbf{I}(C_\delta)$ can usually be expressed via the so called *rate function* $I : \mathbb{R}^d \mapsto \mathbb{R}$, according to the following formula (Bahadur 1960):

$$\mathbf{I}(C_\delta) = \inf_{x \in C_\delta} I(x). \tag{2}$$

Differently from the set function $\mathbf{I}$, the rate function I does not depend on the region C_δ; that is, when C_δ changes, only the region over which we minimize in (2) changes, while the function remains unchanged. Furthermore, this is true for arbitrary set C_δ. This means that, once the rate function is computed, the corresponding inaccuracy rate can be obtained via (2) for a new given region of interest.

In this paper, we are interested in applying the large deviations theory to analyzing the stochastic gradient descent (SGD) method. SGD is a simple but widely used optimization method that finds numerous practical applications, such as training machine learning and deep learning

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

models, e.g., (Niu et al. 2011; Gorbunov, Hanzely, et al. 2020; Lei et al. 2020). More precisely, we consider unconstrained optimization problems where the goal is to minimize a smooth, strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, via the SGD method of the form:

$$X_{k+1} = X_k - \alpha_k \left(\nabla f(X_k) - Z_k\right). \tag{3}$$

Here, $k = 1, 2, ...$ is the iteration counter, $\alpha_k = a/k$, $a > 0$ is the step-size, and Z_k is a zero-mean gradient noise that may depend on X_k. In this context, we are interested in solving for (1) and (2) for the SGD method (3), where now $x^\star$ is interpreted as the (deterministic) global minimizer of f. In other words, we are interested in finding (or approximating) the rate function $I(x)$ that quantifies the "tails" or "rare events" of how the SGD sequence iterates X_k deviate from the solution $x^\star$.

Clearly, evaluating (2) for SGD is of significant interest. It readily provides insights into the high-probability bounds for SGD that have been subject of much research effort recently, (Ghadimi et al. 2012; Ghadimi et al. 2013; Juditsky et al. 2019; Gorbunov, Danilova, et al. 2020; Davis et al. 2021). However, unlike the typical high probability bound studies, the large deviations approach here is fully flexible with respect to the choice of set C_δ; e.g., the l_2-ball complement may be replaced with an arbitrary open set, such as l_p norm complement of an arbitrary l_p-norm. While large deviations theory is a well-established field, there has been a limited body of work that applies large deviations to the analysis of SGD. Reference (Woodroofe 1972) is concerned with large deviations analysis for a scalar stochastic process equivalent to SGD in one dimension. The authors of (W. Hu et al. 2019) study large deviations of SGD when the step-size converges to zero; however, they are not concerned with large deviations when the iteration counter k increases – the case of our interest here.

Contributions. In this paper, we are interested in evaluating the large deviations rates in (1) and (2) for the SGD method, when the objective function f is smooth and strongly convex. Our main contributions are as follows. When f is a (strongly convex) quadratic function, we establish the so-called full large deviations principle for the sequence X_k. This means that we evaluate rate function $I(x)$ exactly, i.e., the corresponding rare event probability is computed exactly, with upper and lower bounds matched, up to exponentially decaying factors. We further explicitly quantify the rate function $I(x)$ as a function of the distribution of the gradient noise. This reveals a significant influence of higher order moments on the performance (in the sense of rare event probabilities) of SGD. This is in contrast with conventional SGD analyses, that typically capture only the dependence on the gradient noise variance. The large deviations principle for quadratic functions is established under a very general class of gradient noise distributions that are essentially only required to have a finite moment generating function. Next, for generic smooth and strongly convex costs f, we establish a large deviations upper bound (a lower bound on function $I(x)$) that certifies an exponential decay of the rare event probabilities in (1) with SGD. This is achieved when the distribution of the gradient noise is sub-Gaussian. We further show that the obtained large deviations upper bound is tight, as the corresponding rate function actually matches, up to higher order factors, the exact rate function that we formerly establish for the quadratic costs.

Our results are related with high probability bounds-type studies of SGD and related stochastic methods (Harvey et al. 2019; Ghadimi et al. 2012; Ghadimi et al. 2013; Juditsky et al. 2019; Gorbunov, Danilova, et al. 2020). Therein, for a given $\delta > 0$ and a confidence level $1 - \beta$, $\beta \in (0, 1)$, the goal is to find $K(\delta, \beta)$ such that $f(X_k) - f(x^\star) \leq \delta$ with probability at least $1 - \beta$, for all $k \geq K(\delta, \beta)$. The works (Ghadimi et al. 2012; Ghadimi et al. 2013; Juditsky et al. 2019; Gorbunov, Danilova, et al. 2020) provide estimates of $K(\delta, \beta)$ that depend *logarithmically* on β. In more detail, (Ghadimi et al. 2012; Ghadimi et al. 2013) establish high probability bounds for the stochastic gradient methods therein assuming sub-Gaussian gradient noises. The work (Juditsky et al. 2019) calculates the corresponding bounds for the basic SGD and the mirror descent that utilize a gradient truncation technique, while relaxing the noise sub-Gaussianity. The work (Gorbunov, Danilova, et al. 2020) establishes high probability bounds for an accelerated SGD that also utilizes a clipping nonlinearity. The large deviations rates in (1) and (2) - give estimates of $K(\delta, \beta)$ that also depend logarithmically on β, when β is small (goes to zero).[1]

Compared with existing high probability bound works, our results give the *exact* (tight) exponential decay rate in (2), and for an *arbitrary set* that does not contain $x^\star$, not only the Euclidean ball complements. To be concrete, the closest results to ours are obtained in (Harvey et al. 2019). While they are not directly concerned with obtaining large deviations rates, their results (with some additional work) lead to an exponential decay rates for Euclidean ball complements. In contrast, our results work for arbitrary open sets. Furthermore, focusing only on Euclidean ball complements, our results provide much tighter exponential rate bounds. Specifically, as we show in the paper, the exponential rate that we provide captures the interplay between the noise geometry and the cost function curvature, see Section 4.2 for details. From the technical perspective, this is achieved by working directly with the SGD iterates, as opposed to working with the distance of the iterates from the solution. To do so, we derive a novel set of techniques

[1] It is easy to see this by noting that, for μ-strongly convex costs, we have $f(x) - f(x^\star) \geq \frac{\mu}{2}\|x - x^\star\|^2$, for all $x \in \mathbb{R}^d$, requiring that the the right hand side of (1) be less than β, and reverse-engineering the smallest iterate k for which the latter holds.

that build upon the large deviations theory rather than on martingale concentration inequalities.

The current paper is also related with large deviations analyses of stochastic processes that arise with distributed inference, such as estimation and detection. Distributed detection has been studied in (Bajovic, Jakovetic, Xavier, et al. 2011), for Gaussian observations, and in (Bajovic, Jakovetic, Moura, et al. 2012), for generic observations. The work (Matta et al. 2016a) evaluates large deviations of the local states with a distributed detection method, when the step size parameter decreases. Reference (Matta et al. 2016b) further analyzes the non-exponential terms and consider directed networks for a similar problem. The paper (Marano et al. 2019) considers distributed detection with 1-bit messages. (P. Hu et al. 2022) consider social learning problems. Reference (Bajovic 2022) analyzes large deviations for distributed estimation and social learning. Unlike these works on distributed inference, we are not directly concerned with distributed systems; also, the cost functions that we consider are more general and, unlike the works above, do not result in linear (distributed averaging) dynamics; hence, novel tools for large deviations analysis are required here.

The rest of the paper is organized as follows. Section 2 explains the problem that we consider and gives the required preliminaries. Section 3 provides the main results of the paper – a large deviations upper bound for generic costs, and the full (exact) large deviations rates for quadratic costs. Specializing to the Gaussian noise, Section 4 provides analytical, closed-form expressions for the large deviations rate function. Finally, we conclude in Section 5. Appendix contains additional insights and examples, numerical results, and missing proofs.

2 SETUP AND PRELIMINARIES

We consider unconstrained optimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x). \tag{4}$$

We assume that f is L-smooth and μ-strongly convex, and that the stepsize in algorithm (3) is of the form $\alpha_k = a/(k+b)$, where $a,\, b > 0$.

Assumption 1. *We assume that f is twice differentiable, L-smooth and μ-strongly convex, where $0 < \mu \leq L$.*

Strong convexity implies uniqueness of the solution of (4), which is denoted by $x^\star$. We make the following assumption regarding the stepsize parameter a.

Assumption 2. *The stepsize parameter a satisfies $a\mu > 1$.*

Assumption 1 is standard in the analysis of optimization methods, i.e., it corresponds to a standard class of functions over which an optimization method analysis is carried out. Assumption 2 is required for some asymptotic arguments ahead, as $k \to \infty$. In practice, it may be restrictive that the constant a is too large in the step-size choice a/k, as at the initial iterations (small k's), we would have very large step-sizes. This is alleviated by having an appropriately chosen constant $b > 1$.

We denote by $\tilde{g}(X_k)$ the stochastic gradient of f returned by the gradient oracle at the current iterate X_k, and by $g(X_k)$ the (exact) gradient of f at the current iterate X_k. The difference between $\tilde{g}(X_k)$ and $g(X_k)$ (the gradient "noise") is denoted by $Z_k = g(X_k) - \tilde{g}(X_k)$. We make the following assumptions on Z_k.

Assumption 3. *1. For each k, Z_k depends on the past iterates only through X_k.*

2. For each k, the distribution of Z_k given X_k depends on X_k only through its realization and does not depend on the current iterate index, k.

3. For any given x, $\mathbb{E}[Z_k | X_k = x] = 0$, i.e., conditioned on the current iterate, the noise is zero-mean.

Assumption 3 allows for a general gradient noise that may actually depend on the current iterate X_k. This is a more general setting than the frequently studied case when Z_k is i.i.d. and independent of X_k. Item 3. of Assumption 3 says that, conditioned on the current iterate, the noise is zero-mean on average. This is also a standard bias-free noise assumption. Finally, note that items 1. and 2. in Assumption 3 typically hold in machine learning settings. Therein, the goal is typically to minimize a population loss $f(x) = \mathbb{E}[\phi(x, v)]$ where the expectation is taken over the distribution of the data v, and ϕ is an instantaneous loss function. Given that, at some iteration k, X_k takes a value x, the gradient noise equals $\nabla_x \phi(x, v_k) - \mathbb{E}[\nabla_x \phi(x, v)]$, where v_k is the data point sampled at iteration k. Then, items 1. and 2. are clearly satisfied, provided that the data sampling process is independent of the evolution of X_k.

For $x \in \mathbb{R}^d$, we denote by $H(x)$ the Hessian matrix of f computed at x. For compactness, we denote $H^\star = H(x^\star)$, i.e., $H^\star$ is the Hessian matrix of f computed at $x^\star$. For any $x \in \mathbb{R}^d$, define $h : \mathbb{R}^d \mapsto \mathbb{R}^d$ as the residual of the first order Taylor's approximation of the gradient g at $x^\star$,

$$h(x) = g(x) - H^\star(x - x^\star), \tag{5}$$

for $x \in \mathbb{R}^d$. For each $\delta > 0$, define also

$$\overline{h}(\delta) = \sup_{x\in\mathbb{B}_{x^\star}(\delta)} \|h(x)\|, \tag{6}$$

where $B_x(\delta)$ denotes the closed Euclidean ball in $\mathbb{R}^d$ of radius $\delta \geq 0$, centered at x. The following result holds by a well-known corollary of Taylor's remainder theorem.

Lemma 1. *There holds $\overline{h}(\delta) = o(\delta)$, i.e., $\lim_{\delta\to 0} \frac{\overline{h}(\delta)}{\delta} = 0$.*

Remark 1. *Clearly, when f is quadratic, $H(x)$ is constant for all $x \in \mathbb{R}^d$ and equal to $H^\star$, implying $h(x) \equiv 0$ and also $\overline{h}(\delta) \equiv 0$.*

Remark 2. *Lemma 1 holds by the twice continuous differentiability of f. The quantity $h(x)$ can be explicitly characterized if, in addition, it is assumed that the Hessian of function f is Lipschitz continuous, i.e., if $\|H(x) - H(y)\| \leq L_H \|x - y\|$, for all $x, y \in \mathbb{R}^d$, for some nonnegative constant L_H. It is easy to show that, in this case, we have $\|h(x)\| \leq L_H \|x - x^\star\|^2$, for any $x \in \mathbb{R}^d$. The latter implies a quadratic upper bound in δ on $\overline{h}(\delta)$, i.e., $\overline{h}(\delta) \leq L_H \delta^2$, for each $\delta \geq 0$.*

2.1 Distance to solution recursion

For analytical purposes, it is of interest to study the squared distance to solution of the current iterates $\|X_k - x^\star\|^2$. To characterize the evolution of this quantity, we use standard arguments that follow from strong convexity and Lipschitz smoothness:

$$\|X_{k+1} - x^\star\|^2 \leq \left(1 - 2\alpha_k \mu + 2\alpha_k^2 L^2\right) \|X_k - x^\star\|^2 + 2\alpha_k (X_k - x^\star)^\top Z_k + 2\alpha_k^2 \|Z_k\|^2; \tag{7}$$

details of the derivations can be found in Appendix A.

We introduce the function $\beta_k : \mathbb{R}^2 \mapsto \mathbb{R}$, defined by $\beta_k(u, v) = 1 - \alpha_k u + \alpha_k^2 v$. Similarly, for any two iteration indices $l \leq k$, we define $\beta_{k,l} : \mathbb{R}^2 \mapsto \mathbb{R}$ by $\beta_{k,l}(u, v) = \beta_k(u, v) \cdots \beta_l(u, v)$. The following technical lemma providing bounds on the product functions $\beta_{k,l}$ will be useful for the study of recursion (7) as well as other similar recursions that will emerge from the analysis.

Lemma 2. *Let l and k be two iteration indices such that $l < k$. For any nonnegative u, $v \in \mathbb{R}$, and $\alpha_k = a/(k+b)$, where $b \geq 1$, there holds:*

1. $\beta_{k,l}(u, v) \leq \left(\frac{l+b}{k+b+1}\right)^{au} e^{\frac{a^2 v}{l+b-1}}$;
2. *for each l such that $l + b \geq \frac{5au}{2}$, there holds* $\beta_{k,l}(u, v) \geq \left(\frac{l+b-1}{k+b}\right)^{au} e^{-\frac{a^2 u^2}{l+b-1}}$;

The proof of Lemma 2 is given in Appendix A.

Finally, for each iteration index k, we denote by μ_k the Borel measure on $\mathbb{R}^d$ induced by X_k. Similarly, we denote by ν_k the Borel measure induced by $\|X_k - x^\star\|$.

2.2 Large deviations preliminaries

We next give a definition of the rate function and the large deviations principle.

Rate function I and the large deviations principle.

Definition 1 (Rate function I (Dembo et al. 1993))**.** *Function $I : \mathbb{R}^d \mapsto [0, +\infty]$ is called a* rate function *if it is lower semicontinuous, or, equivalently, if its level sets are closed. If, in addition, the level sets of I are compact (i.e., closed and bounded), then I is called a good rate function.*

Definition 2 (The large deviations principle (Dembo et al. 1993))**.** *Suppose that $I : \mathbb{R}^d \mapsto [0, +\infty]$ is lower semicontinuous. A sequence of measures μ_k on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$, $k \geq 1$, is said to satisfy the large deviations principle (LDP) with rate function I if, for any measurable set $D \subseteq \mathbb{R}^d$, the following two conditions hold:*

1. $\limsup_{k \to +\infty} \frac{1}{k} \log \mu_k(D) \leq - \inf_{x \in \overline{D}} I(x)$;
2. $\liminf_{k \to +\infty} \frac{1}{k} \log \mu_k(D) \geq - \inf_{x \in D^\circ} I(x)$.

Log-moment generating functions of the noise Z_k and the iterates X_k. Following Assumption 3, we define the conditional LMGF of Z_k given the last iterate X_k.

Definition 3 (Conditional LMGF of Z_k)**.** *We denote by $\Lambda(\cdot; x)$ the log-moment generating function (LMGF) of Z_k given $X_k = x$,*

$$\Lambda(\lambda; x) := \log \mathbb{E}\left[e^{\lambda^\top Z_k} \middle| X_k = x\right], \text{ for } \lambda, x \in \mathbb{R}^d. \tag{8}$$

It will also be useful to define the conditional moment-generating function of $\|Z_k\|^2$, which we denote by $M(\cdot; x)$:

$$M(\nu; x) := \mathbb{E}\left[e^{\nu \|Z_k\|^2} \middle| X_k = x\right], \tag{9}$$

for $\nu \in \mathbb{R}$, $x \in \mathbb{R}^d$. By the inequality $e^x \leq x + e^{x^2}$, which holds for all $x \in \mathbb{R}$, we have $\mathbb{E}\left[e^{\lambda^\top Z_k} \middle| X_k\right] \leq \mathbb{E}[\lambda^\top Z_k | X_k] + \mathbb{E}\left[e^{(\lambda^\top Z_k)2} | X_k\right] \leq \mathbb{E}\left[e^{\|\lambda^2\| \|Z_k\|^2} \middle| X_k\right]$, where we used the Cauchy-Schwartz inequality, for the second term, and the fact that Z_k is zero-mean, for the first term. Thus,

$$\Lambda(\lambda; x) \leq \log M(\|\lambda^2\|; x) \tag{10}$$

for any realization x of X_k.

Lemma 3 lists properties of Λ that will be used in the paper.

Lemma 3 (Properties of Λ)**.** *For any given $x \in \mathbb{R}^d$ the following properties hold:*

1. *$\Lambda(\cdot; x)$ is convex and differentiable in the interior of its domain;*
2. *$\Lambda(0; x) = 0$ and $\nabla \Lambda(0; x) = \mathbb{E}[Z_k | X_k = x] = 0$;*
3. *$\Lambda(\lambda; x) \geq 0$, for each λ.*

Proof. Convexity and differentiability are general properties of log-moment generating functions (Dembo et al. 1993), as well as the zero value at the origin property and also that the gradient at the origin equals the mean vector; $\nabla\Lambda(0;x) = 0$ follows by the assumption that the noise is zero-mean, Assumption 3. The non-negativity from Part 3 follows by invoking convexity and exploiting the two properties from part 2, i.e., for any $x \in \mathbb{R}^d$: $\Lambda(\lambda;x) \geq \Lambda(0;x) + \nabla\Lambda(0;x)^\top\lambda = 0$. □

Example 1. *To illustrate the LMGF function Λ, we consider the case when, conditioned on an arbitrary realization $X_k = x$, the gradient noise Z_k is Gaussian, with mean vector equal to zero vector and covariance matrix $\Sigma(x)$. Using standard formula for the LMGF of a Gaussian multivariate, we have*

$$\Lambda(\lambda;x) = \frac{1}{2}\lambda^\top S(x)\lambda, \tag{11}$$

for $\lambda \in \mathbb{R}^d$. We note that when the gradient noise Z_k is independent of the current iterate X_k, the indices X_k in the preceding formula can be omitted, i.e., the expression for Λ simplifies to $\Lambda(\lambda;X_k) = \frac{1}{2}\lambda^\top S\lambda$, for all realizations X_k.

It will also be of interest to define the (unconditional) log-moment generating function of the iterates X_k.

Definition 4 (LMGF of $X_k - x^\star$)**.** *We let Γ_k denote the (unconditional) moment generating function of X_k,*

$$\Gamma_k(\lambda) := \mathbb{E}\left[e^{\lambda^\top(X_k - x^\star)}\right], \tag{12}$$

for $\lambda \in \mathbb{R}^d$. The (unconditional) log-moment generating function of X_k is then given by $\log\Gamma_k$.

We assume that the initial iterate X_1 is deterministic[2]. Hence, Γ_1 is finite for all $\lambda \in \mathbb{R}^d$.

We assume that the family of functions $\Lambda(\cdot;x)$ satisfy the following regularity conditions.

Assumption 4 (Lipschitz continuity in x)**.** *There exists a constant L_Λ such that for every λ, x, $y \in \mathbb{R}^d$, there holds:*

$$|\Lambda(\lambda;x) - \Lambda(\lambda;y)| \leq L_\Lambda\|\lambda\|^2\|x - y\|. \tag{13}$$

Remark 3. *We note that Assumption 4 is trivially satisfied when the noise distribution does not depend on the current iterate. For another illustration, consider Gaussian random noise distribution from Example 1, for which we have:*

$$\Lambda(\lambda;x) - \Lambda(\lambda;y) = \frac{1}{2}\lambda^\top(S(x) - S(y))\lambda \tag{14}$$

$$\leq \frac{1}{2}\|\lambda\|^2\|S(x) - S(y)\|. \tag{15}$$

[2]We note that this assumption can be relaxed to allow for random initial iterate; see Appendix D for details.

Comparing with the condition in (13)*, we see that* (13) *is satisfied when entries of the covariance matrix S, as functions of x, are Lipschitz continuous.*

The assumption below will be used for the proof of the main result of the paper, when the case of general convex functions is considered.

Assumption 5 (Sub-Gaussian noise)**.** *There exists a constant $C_1 > 0$ such that, for each λ, $x \in \mathbb{R}^d$*

$$\Lambda(\lambda;x) \leq C_1\frac{\|\lambda\|^2}{2}. \tag{16}$$

Remark 4. *Assumption 5 means that the gradient noise has "light tails," i.e., there exist positive constants c_1, c_2, such that the probability that the magnitude of the norm of the noise vector is above ϵ is upper bounded by $c_1\, e^{-c_2\epsilon^2}$, for any $\epsilon > 0$. Clearly, a Gaussian zero-mean multivariate distribution satisfies this property, and also any noise distribution with compact support.*

This assumption also ensures that, for each given λ, the value of the variance "proxy" C_1 cannot grow without bound as the domain of iterates x enlarges. For a Gaussian distribution, this means that the variance, as a function of the current iterate should be uniformly bounded over the domain of the iterates, which is a typical assumption in related works.

We also use the following implications of Assumption 5.

1. There exists $C_2 > 0$ such that

$$\mathbb{E}\left[\exp\left(\frac{\|Z_k^2\|}{C_2}\right)\middle|\, X_k\right] \leq e. \tag{17}$$

2. For any $\nu \in [0, 1/C_2]$ there holds

$$M(\nu;X_k) \leq \exp(\nu C_2). \tag{18}$$

Proof. The proof of part 1 can be derived by applying properties of sub-Gaussian random variables to $\|Z_k\|$; see, e.g., Proposition 2.5.2 in (Vershynin 2018) and also (Jin et al. 2019) for a treatment of sub-Gaussian random vectors.

To show part 2, fix $\nu \in [0, 1/C_2]$. By Hölder's inequality (applied for "p" $= 1/(\nu C_2) \geq 1$)

$$M(\nu;X_k) \leq \left(\mathbb{E}\left[\exp\left(1/C_2\|Z_k\|^2\right)\middle|\, X_k\right]\right)^{\nu C_2} \tag{19}$$

$$\leq \exp(\nu C_2) \tag{20}$$

where in the second inequality we used part 1. □

Remark 5. *When the distribution of Z_k is Gaussian, zero mean and with covariance matrix Σ, and independent of the current iterate, we have*

$$\Lambda(\lambda) = \frac{1}{2}\lambda^\top\Sigma\lambda \leq \frac{1}{2}\sigma_{\max}^2\|\lambda\|^2, \tag{21}$$

where $\sigma^2_{\max}$ is the maximal eigenvalue of Σ. Comparing with Assumption 5, we see that condition (16) *holds with $C_1 = \sigma^2_{\max}$. It can also be shown that part 1. of Proposition 2.2 holds for $C_2 \geq 2\sigma^2_{\max}$.*

2.3 Key technical lemma

Definition 5. *The Fenchel-Legendre transform, or the conjugate, of a given function $\Psi : \mathbb{R}^d \mapsto \mathbb{R}$ is defined by*

$$I(x) = \sup_{\lambda \in \mathbb{R}^d} x^\top \lambda - \Psi(\lambda), \text{ for } x \in \mathbb{R}^d. \tag{22}$$

Lemma 4. *Let Ψ_k be a sequence of log-moment generating functions associated to a given sequence of measures $\mu_k : \mathcal{B}(\mathbb{R}^d) \mapsto [0,1]$. Suppose that, for each $\lambda \in \mathbb{R}^d$, the following limit exists:*

$$\limsup_{k \to +\infty} \frac{1}{k} \Psi_k(k\lambda) \leq \Psi(\lambda). \tag{23}$$

If $\Psi(\lambda) < \infty$ for each $\lambda \in \mathbb{R}^d$, then the sequence μ_k satisfies the LDP upper bound with the rate function I equal to the Fenchel-Legendre transform of Ψ. If, in addition, (23) *holds as a limit and with equality, then the sequence of measures satisfies the LDP with rate function I.*

The second part of the lemma follows by the Gärtner-Ellis theorem. The first part can be proven by similar arguments as in the proof of the upper bound of the Gärtner-Ellis theorem; for details, see also the proof of Lemma 35 in (Bajovic 2022).

3 LARGE DEVIATIONS RATES FOR SGD ITERATES X_k

3.1 Large deviations rates for $\|X_k - x^\star\|$

To derive the main result – the large deviations rate function for the SGD sequence X_k, we first study large deviations properties of the sequence $\|X_k - x^\star\|$, $k = 1, 2, ...$ For the latter, we first exploit the idea from (Harvey et al. 2019) to obtain a high probability bound for the (scaled) quantity $\|X_k - x^\star\|^2$, via its moment generating function. We then use this bound to derive a rate function (bound) for $\|X_k - x^\star\|$. Since our assumptions are distinct than those in (Harvey et al. 2019) (e.g., the recursive form that we work with here contains factors that require special treatment than the one in (Harvey et al. 2019), also we do not assume bounded noisy gradient, as is the case with the proof available in (Harvey et al. 2019)), we provide full proof details, see Appendix B.

Lemma 5. *For any k, there holds*

$$\mathbb{P}\left(\|X_k - x^\star\| \geq \delta\right) \leq e e^{-(k+k_0)B\delta^2}, \tag{24}$$

where $B = \min\{\frac{1}{k_0 \|X_1 - x^\star\|}, \frac{2a\mu - 1}{4\max\{C_1, 2C_2\}a^2}\}$ and $k_0 = 4a^2L^2/(2a\mu - 1)$.

Remark 6. *The preceding theorem establishes a large deviations upper bound for the sequence of squared distance to solution iterates ξ_k, by exploiting noise sub-Gaussianity. By its nature, this result is a rough characterization of the large deviations rate function for the sequence X_k. In addition to being a result of independent interest, the utility consists in bounding the tails of distribution μ_k, as an enabling step towards deriving a fine, close to exact rate function for the SGD iterates X_k, as the main contribution of this paper. The latter is the subject of the next section.*

3.2 Main result: Large deviations rates for X_k

We now present our result for general convex functions satisfying assumptions from Section 2. The pillar of the analysis is the limit of the sequence of log-moment generating functions $\log \Gamma_k$ of the SGD iterates.

Lemma 6. *Suppose that Assumptions 1-5 hold and that the stepsize is given by $\alpha_k = a/(k + k_0)$. For any $\lambda \in \mathbb{R}^d$,*

$$\limsup_{k \to +\infty} \frac{1}{k} \log \Gamma_k(k\lambda) \leq \overline{\Psi}(\lambda) := \Psi^\star(\lambda) + r(\lambda), \tag{25}$$

where $\Psi^\star$ is defined by

$$\Psi^\star(\lambda) = \int_0^1 \Lambda(aQD(\theta)Q^\top \lambda; x^\star) d\theta, \tag{26}$$

where $H^\star = QDQ^\top$, $QQ^\top = I$, $D = \text{diag}\{\rho_1, ..., d_n\}$, $D(\theta) = \text{diag}\{\theta^{a\rho_1 - 1}, ..., \theta^{ad_n - 1}\}$, $r(\lambda) = \frac{4a^2\overline{\gamma}^2 L_\Lambda}{B^2}\|\lambda\|^4 + a\|\lambda\|\overline{h}\left(\frac{2\overline{\gamma}\|\lambda\|}{B}\right)$, and $\overline{\gamma} = \max\{1, \sqrt{(1-a\mu)^2 + a^2(L^2 - \mu^2)}\}$.

The proof of Lemma 6 is given in Appendix C. Having the limit in (25), LDP upper bound follows by Lemma 4.

Theorem 1. *Suppose that Assumptions 1-5 hold and that the stepsize is given by $\alpha_k = a/(k + k_0)$. Then, the sequence of iterates X_k satisfies the LDP upper bound with rate function $\overline{I}$ given as the Fenchel-Legendre transform of $\overline{\Psi}$ from Lemma 6, i.e., for any closed set F:*

$$\limsup_{k \to +\infty} \frac{1}{k} \log \mathbb{P}\left(X_k \in F\right) \leq - \inf_{x + x^\star \in F} \overline{I}(x). \tag{27}$$

Remark 7. *The rate function $\overline{I}$ depends on the Hessian matrix at the solution, $H(x^\star)$. However, coarser exponential rate bounds can be obtained by uniformly bounding the eigenvalues of $H(x^\star)$, as by our assumptions they are all confined in the interval $[\mu, L]$. See Appendix D for details.*

3.3 Discussions and interpretations

3.3.1 Positivity of $\overline{I}$ and exponential decay

From the fact that $\Psi^\star$, $r \geq 0$, and that both $\Psi^\star$ and r are finite on $\mathbb{R}^d$, it can be shown that $\overline{I} \geq 0$ and that $\overline{I}$ is a good

rate function. Specifically, $\overline{I}(0) = 0$ and $\overline{I}(x) > 0$ for any $x \neq 0$. Therefore, for any closed set F such that $x^\star \notin F$, we have

$$\inf_{x+x^\star \in F} \overline{I}(x) > 0, \tag{28}$$

that is, the exponent in (27) is strictly positive ensuring the exponential decay of the probabilities $\mathbb{P}(X_k \in F)$. To illustrate this in intuitive terms, we take as a special case the set $F = B^c_{x^\star}(\delta)$, for some $\delta > 0$. Then, the event of interest becomes $\{X_k \in F\} = \{\|X_k - x^\star\| \geq \delta\}$. Thus, for any $\delta > 0$, Theorem 1 implies that

$$\limsup_{k \to +\infty} \frac{1}{k} \log \mathbb{P}(\|X_k - x^\star\| \geq \delta) \leq -R(\delta), \tag{29}$$

where $R(\delta) = \inf_{\|x\| \geq \delta} I(x) > 0$.

3.3.2 Remainder term r

Recalling Lemma 1, it is easy to see that $r(\lambda) = o(\|\lambda\|^2)$, i.e., $\lim_{\|\lambda\| \to 0} \frac{r(\lambda)}{\|\lambda\|^2} = 0$. Also, for a function f that has Lipschitz Hessian, see Remark 2, the residual function r behaves roughly as $\sim \|\lambda\|^3$.

Further, for the special case when f is quadratic, $\overline{h}(\delta) = 0$, and hence r contains only the first term, and thus $r(\lambda) \sim \|\lambda\|^4$. Similarly, when the noise distribution does not depend on the current iterate, we have that $L_\Lambda = 0$, and hence $r(\lambda) = o(\|\lambda\|^2)$. Finally, for the case when both of the preceding conditions hold, the residual term is zero at all points: $r \equiv 0$, and hence the rate function $\overline{I} = I^\star$, where $I^\star$ is the Fenchel-Legendre transform of $\Psi^\star$.

3.3.3 Small deviations regime

When high precision estimates are sought, or equivalently, for small δ in (29), the candidate values of $\overline{I}$ in the minimization are very close to 0. By the fact that the remainder term $r(\lambda) = o(\|\lambda^2\|)$, it can be shown that, in the small deviations regime, $\overline{I}$ is determined by $\Psi^\star$ only, i.e., $\overline{I} \approx I^\star$, and, also, its behaviour is dominantly characterized by the noise variance.

3.4 LDP for quadratic functions

In this section we provide the full LDP for the case when f is a quadratic function. The proof of Theorem 2 is given in Appendix E.

Theorem 2. *Suppose that the objective function f is quadratic, that Assumptions 2-3 hold, with the step size given by $\alpha_k = a/k$. Suppose also that the noise distribution does not depend on the current iterate and that it has a finite log-moment generating function Λ. Then, the sequence X_k satisfies the large deviations principle with the rate function $I^\star$ given as the conjugate of $\Psi^\star$ defined in* (26)*, with $\Lambda(\cdot; x^\star)$ replaced by Λ.*

The rate function $I^\star$ depends on the distribution of Z_k and fully captures all moments of this distribution. In particular, for non-Gaussian distributions, it captures exactly the dependence not only on the variance, but also on higher order moments.

Remark 8. *We note that, in contrast with Theorem 1, for Theorem 2 the conditional distribution of Z_k can be arbitrary, as long as Λ is finite. In particular, it allows for distributions that are not light-tailed, such as Laplacian.*

Remark 9. *Recalling the discussion from subsection 3.3.2, we see that the upper bound rate function from Theorem 1 and the rate function from Theorem 2 match, hence showing that the bound in Theorem 1 is tight.*

4 GAUSSIAN NOISE: ANALYTICAL CHARACTERIZATION OF THE RATE FUNCTION

If the noise Z_k has a Gaussian distribution with mean value zero and covariance matrix Σ, then $\Psi^\star$ is computed by

$$\Psi^\star(\lambda) = \frac{a^2}{2} \int_0^1 \lambda^\top Q D(\theta) Q^\top \Sigma Q D(\theta) Q^\top \lambda d\theta. \tag{30}$$

To simplify the notation, let $S = Q^\top \Sigma Q$, and $M(\theta) = D(\theta) S D(\theta)$. It is easy to verify that $M_{ij}(\theta) = S_{ij} \theta^{a(\rho_i + \rho_j) - 2}$, for any $i, j = 1, ..., d$, and thus $\int_0^1 M_{ij}(\theta) d\theta = S_{ij}/(a(\rho_i + \rho_j) - 1)$. Hence, we obtain the following closed-form expression for $\Psi^\star$:

$$\Psi^\star(\lambda) = \frac{a^2}{2} \lambda^\top Q S^\star Q \lambda, \tag{31}$$

where $S^\star_{ij} = S_{ij}/(a(\rho_i + \rho_j) - 1)$, for $i, j = 1, ..., d$.

Recalling the Definition 5, it can be shown that the Fenchel-Legendre transform $I^\star$ of $\Psi^\star$ is given by

$$I^\star(z) = \frac{1}{2a^2} z^\top Q^\top S^{\star -1} Q z. \tag{32}$$

To obtain further intuition about the rate function $I^\star$, we consider the special case when the Hessian matrix $H^\star$ and the covariance matrix Σ share the same eigenspace (given by the columns of the matrix Q). Intuitively, the latter means that the orientation of the quadratic approximation of f at the origin is aligned with the gradient noise distribution in each of the axes. In this case, it follows that $S = Q^\top \Sigma Q$ is diagonal with $S_{ii} = \sigma^2_{ii}$, where σ^2_{ii} is the i-th eigenvalue of Σ (i.e., the eigenvalue of Σ corresponding to its eigenvector given by the i-th column of matrix Q). It follows that $S^\star$ is also diagonal with $S^\star_{ii} = \sigma^2_{ii}/(2a\rho_i - 1)$. Thus, the following neat expression for the rate function $I^\star$ emerges:

$$I(z) = \frac{1}{2a^2} z^\top Q^\top \mathrm{diag}\left(\frac{2a\rho_1 - 1}{\sigma^2_{11}}, ..., \frac{2a\rho_d - 1}{\sigma^2_{dd}} \right) Q z. \tag{33}$$

4.1 Decay rates with l_2 balls

We consider the case when in the large deviations event of interest $\{X_k \in F\}$ the set F is given as the complement of an l_2 ball around the solution $x^\star$: $F = B^{\mathrm{c}}_{x^\star}(\delta)$, i.e., $\{X_k \in F\} = \{\|X_k - x^\star\| \geq \delta\|$. Assuming that the residual is zero (see the result for quadratic functions in Section 3.4), by Theorem 1, we have

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(\|X_k - x^\star\| \geq \delta\right) \leq \inf_{\|z\|\geq\delta} I(z) =: \mathbf{I}(B^{\mathrm{c}}_{x^\star}(\delta)). \quad (34)$$

For the Gaussian noise assumed in this section, we have:

$$\begin{aligned}\mathbf{I}(B^{\mathrm{c}}_{x^\star}(\delta)) &= \inf_{\|z\|\geq\delta} \frac{1}{2a^2} z^\top Q^\top S^{\star -1} Q z \\ &= \frac{\delta^2}{2a^2} \inf_{\|w\|\geq 1} w^\top Q^\top S^{\star -1} Q w \\ &= \frac{\delta^2}{2a^2} \frac{1}{\lambda_{\max}(S^\star)}, \end{aligned} \quad (35)$$

where $\lambda_{\max}(S^\star)$ is the largest eigenvalue of the matrix $S^\star$. Hence, to find the value of the exponent $\mathbf{I}$ for any given ball-shaped set, it suffices to find (once) the maximal eigenvalue of $S^\star$, and the exponent $\mathbf{I}$ would be easily computed by the quadratic function (35).

We close the analysis with a particularly elegant solution for the special case when $H^\star$ and Σ are axes-aligned. As detailed at the beginning of the section, in the latter case, $S^\star$ is diagonal, with $S^\star_{ii} = \sigma^2_{ii}/(2a\rho_i - 1)$, and the rate function is given by (33). Thus, to find the maximal eigenvalue of $S^\star$ reduces to finding the index i for which $\frac{\sigma^2_{ii}}{2a\rho_i - 1}$ is highest, or, equivalently, $\frac{2a\rho_i - 1}{\sigma^2_{ii}}$ the lowest, which then yields:

$$\mathbf{I}(B^{\mathrm{c}}_{x^\star}(\delta)) = \frac{\delta^2}{2a^2} \min\{\frac{2a\rho_i - 1}{\sigma^2_{ii}} : i = 1, ...d\}, \quad (36)$$

where, we recall, ρ_i is the i-th eigenvalue of $H^\star$. What the expression above is saying is that, in order to find the exponential decay rate for an l_2 ball, we should search for the direction i in which the value $\frac{\sigma^2_{ii}}{2\rho_i - 1}$ is highest. In a sense, the latter quantity can be thought of as the effective noise variance, capturing the interplay between the noise distribution and the shape of the function at the solution. Specifically, if along the direction where the noise variance is highest, say $i^\star$, the function has a high curvature (i.e., large $\rho_{i^\star}$), this will effectively alleviate the effects of noise and increase the rate function, in comparison to the case when the curvature along i is lower, and therefore result in faster convergence.

Finally, when the noise is isotropic, i.e., such that $\sigma^2_{ii} = \sigma^2$, for all i, exploiting the fact that the spectrum of $H^\star$ lies inside the interval $[\mu, L]$, the rate function is found by:

$$\mathbf{I}(B^{\mathrm{c}}_{x^\star}(\delta)) = \frac{\delta^2}{2a^2} \frac{2a\mu - 1}{\sigma^2}. \quad (37)$$

4.2 Comparison with the rate from Lemma 5

We now compare the rate function bounds obtained from Lemma 5 and Theorem 1. To gain deeper insights, we will assume that the residual term r equals zero (compare with Section 3.4). We also assume that the noise is Gaussian and axes-aligned with the matrix $H^\star$ (see the preceding subsection). The exponent B from 5 can be upper bounded by[3]

$$B \leq \frac{2a\mu - 1}{4{\sigma_{\max}}^2 a^2},$$

where we exploited the fact that, for Gaussian noise, $C_1 = \sigma^2_{\max}$, see Remark 5. Hence, for an l_2 ball of radius δ, the exponent that Lemma 5 provides is bounded by

$$B\delta^2 \leq \frac{\delta^2}{4a^2} \frac{2a\mu - 1}{\sigma^2_{\max}}. \quad (38)$$

The counterpart obtained from Theorem 1 is given by expression (36). To show direct comparison with (38), we further upper bound this value by decoupling the minimization over i :

$$\begin{aligned}\mathbf{I}(B^{\mathrm{c}}_{x^\star}(\delta)) &= \frac{\delta^2}{2a^2} \min\{\frac{2a\rho_i - 1}{\sigma^2_{ii}} : i = 1, ...d\} \\ &\geq \frac{\delta^2}{2a^2} \frac{\min\{2a\rho_i - 1 : i = 1, .., d\}}{\max\{\sigma^2_{ii} : i = 1, .., d\}} \\ &= \frac{\delta^2}{2a^2} \frac{2a\mu - 1}{\sigma^2_{\max}}. \end{aligned} \quad (39)$$

Comparing with (38) (and ignoring the scaling constant 2), the following important point can be noted: on an intuitive level, the derivation of the rate B is equivalent to that of decoupling the effects of the noise distribution and the shape of the function f at the origin. Hence, in contrast with $I^\star$, the rate B is oblivious to the interplay between these two quantities – from a purely technical perspective, this distinction is a consequence of relying on recursions on the iterates' distance to the solution, $\|X_k - x^\star\|$, as opposed to working directly with the iterates X_k, as is the case in the proof of Theorem 1.

5 CONCLUSIONS

We developed large deviations analysis for the stochastic gradient descent (SGD) method, when the objective function is smooth and strongly convex. For (strongly convex)

[3]The dependence in B on X_1 in Lemma 5 seems to be an artifact of the conducted proof method, rather than an essential property of the exponential rate that Lemma 5 pursues. Hence, for unbiased comparison, we omit this factor in the analysis of the rate B.

quadratic costs, we establish the full large deviations principle. That is, we derive the exact exponential rate of decay of the probability that the iterate sequence generated by SGD stays within an arbitrary set that is away from the problem solution. This is achieved for a very general class of gradient noises, that may be iteration-dependent and are required to have a finite log-moment generating function. For generic costs, we derive a tight large deviations upper bound that, up to higher order terms, matches the exact rate derived for the quadratics.

Acknowledgements

The work of D. Bajovic and D. Jakovetic was supported in part by the European Union's Horizon 2020 Research and Innovation program under grant agreement No 957337 and also by Serbian Ministry of Education, Science and Technological Development. The paper reflects only the view of the authors and the Commission is not responsible for any use that may be made of the information it contains. The work of S. Kar was supported in part by the Office of Naval Research under Grant No N00014-21-1-2547.

References

Bahadur, R. R. (1960). "On the Asymptotic Efficiency of Tests and Estimates". In: *Sankhya: The Indian Journal of Statistics, 1933-1960* 22.3/4, pp. 229–252. ISSN: 00364452. URL: http://www.jstor.org/stable/25048458.

Bajovic, D. (2022). *Inaccuracy rates for distributed inference over random networks with applications to social learning*. DOI: 10.48550/ARXIV.2208.05236. URL: https://arxiv.org/abs/2208.05236.

Bajovic, D., D. Jakovetic, J. M. F. Moura, J. Xavier, and B. Sinopoli (Nov. 2012). "Large Deviations Performance of Consensus+Innovations Distributed Detection with Non-Gaussian Observations". In: *IEEE Transactions on Signal Processing* 60.11, pp. 5987–6002.

Bajovic, D., D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura (Sept. 2011). "Distributed Detection via Gaussian Running Consensus: Large Deviations Asymptotic Analysis". In: *IEEE Transactions on Signal Processing* 59.9, pp. 4381–4396.

Bucklew, J. A. (1990). *Large Deviations Techniques in Decision, Simulation and Estimation*. New York: Wiley.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. New York: John Wiley and Sons.

Davis, D., D. Drusvyatskiy, L. Xiao, and J. Zhang (2021). "From Low Probability to High Confidence in Stochastic Convex Optimization." In: *J. Mach. Learn. Res.* 22, pp. 49–1.

Dembo, A. and O. Zeitouni (1993). *Large Deviations Techniques and Applications*. Boston, MA: Jones and Barlett.

Ghadimi, S. and G. Lan (2012). "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework". In: *SIAM J. Optim.* 22.4, pp. 1469–1492.

– (2013). "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms". In: *SIAM J. Optim.* 23.4, pp. 2061–2089.

Gorbunov, E., M. Danilova, and A. Gasnikov (2020). "Stochastic optimization with heavy-tailed noise via accelerated gradient clipping". In: *arXiv preprint arXiv:2005.10785*.

Gorbunov, E., F. Hanzely, and P. Richtarik (2020). "A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 680–690.

Harvey, N. J. A., C. Liaw, Y. Plan, and S. Randhawa (2019). "Tight analyses for non-smooth stochastic gradient descent". In: *32nd Annual Conference on Learning Theory*. Vol. 99. PMLR, pp. 1–35.

Hu, P., V. Bordignon, S. Vlaski, and A. H. Sayed (2022). *Optimal Aggregation Strategies for Social Learning over Graphs*. DOI: 10.48550/ARXIV.2203.07065. URL: https://arxiv.org/abs/2203.07065.

Hu, W., C. J. Li, L. Li, and J. G. Liu (2019). "On the diffusion approximation of nonconvex stochastic gradient descent". In: *Annals of Mathematical Sciences and Applications* 4.1.

Jin, C., P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan (2019). *A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm*. DOI: 10.48550/ARXIV.1902.03736. URL: https://arxiv.org/abs/1902.03736.

Juditsky, A., A. Nazin, A. Nemirovsky, and A. Tsybakov (2019). "Algorithms of robust stochastic optimization based on mirror descent method". In: *arXiv:1907.02707*.

Lei, L. and M. I. Jordan (2020). "On the adaptivity of stochastic gradient-based optimization". In: *SIAM Journal on Optimization* 30.2, pp. 1473–1500.

Marano, S. and A. H. Sayed (Oct. 2019). "Detection under one-bit messaging over adaptive networks". In: *IEEE Trans. Information Theory* 65.10, pp. 6519–6538.

Matta, V., P. Braca, S. Marano, and A. H. Sayed (Aug. 2016a). "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime". In: *IEEE Trans. Information Theory* 62.8, pp. 4710–4732.

– (Dec. 2016b). "Distributed detection over adaptive networks: Refined asymptotics and the role of connectivity". In: *IEEE Trans. Signal and Information Processing over Networks* 2.4, pp. 442–460.

Niu, F., B. Recht, C. Ré, and S. J. Wright (2011). "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent". In: *arXiv preprint arXiv:1106.5730*.

Shwartz, A. and A. Weiss (1995). *Large Deviations for Performance Analysis: Queues, Communications, and Computing*. New York: Chapman and Hall.

Touchette, H. (2009). "The large deviation approach to statistical mechanics". In: *Physics Reports* 478.1, pp. 1–69. ISSN: 0370-1573.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/9781108231596.

Woodroofe, M. (1972). "Normal Approximation and Large Deviations for the Robbins-Monro Process". In: *Z. Wahrscheinlichkeitstheorie verw. Geb.* 21, pp. 329–338.

A PROOFS OF AUXILIARY RESULTS

A.1 Proof of recursion (7)

For any $k \geq 1$, we have:

$$\begin{aligned}\|X_{k+1} - x^\star\| &= \|X_k - \alpha_k g(X_k) + \alpha_k Z_k - x^\star\| \\ &= \|X_k - x^\star\|^2 - 2\alpha_k (X_k - x^\star)^\top (g(X_k) - Z_k) \\ &+ \alpha_k^2 \|g(X_k) - Z_k\|^2 \\ &\leq (1 - 2\alpha_k \mu)\xi_k + 2\alpha_k (X_k - x^\star)^\top Z_k \\ &+ 2\alpha_k^2 \|g(X_k)\| + 2\alpha_k^2 \|Z_k\|^2 \\ &\leq \left(1 - 2\alpha_k \mu + 2\alpha_k^2 L^2\right) \xi_k + 2\alpha_k (X_k - x^\star)^\top Z_k \\ &+ 2\alpha_k^2 \|Z_k\|^2, \end{aligned} \tag{40}$$

where the first inequality follows from the strong convexity of f, Assumption 1, and the fact that, for $a, b \in \mathbb{R}^d$, $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and the second inequality follows from the Lipschitz smoothness of f, Assumption 1.

A.2 Proof of Lemma 2

Fix l and k where $1 \leq l \leq k$. Fix $u,\ v \geq 0$. From the upper and the lower Darboux sum for the logarithmic function applied to the interval $[l, k]$, we obtain:

$$\log \frac{k+1}{l} \leq \frac{1}{l} + \ldots + \frac{1}{k} \leq \log \frac{k}{l-1}. \tag{41}$$

For the 2-sum we use the following simple bound $1/l^2 \leq 1/(l(l-1)) = 1/(l-1) - 1/l$ to obtain:

$$\frac{1}{l^2} + \ldots + \frac{1}{k^2} \leq \frac{1}{l-1} - \frac{1}{l} + \ldots + \frac{1}{k-1} - \frac{1}{k} \leq \frac{1}{l-1}. \tag{42}$$

To prove part 1, we use that $1 + x \leq e^x$ applied to each of the terms in the product $\beta_{k,l}$, together with the left hand-side inequality of (41) and the right hand-side inequality of (42):

$$\begin{aligned}\beta_{k,l}(u,v) &\leq e^{-au \sum_{j=l}^k \frac{1}{j+b} + a^2 v \sum_{j=l}^k \frac{1}{(j+b)^2}} \\ &\leq e^{-au \log\left(\frac{k+b+1}{l+b}\right) + \frac{a^2 v}{l+b-1}} \\ &= \left(\frac{l+b}{k+b+1}\right)^{au} e^{\frac{a^2 v}{l+b-1}}. \end{aligned} \tag{43}$$

To prove part 2, we first note that, since $v \geq 0$, there holds $\beta_{k,l}(u,v) \geq \beta_{k,l}(u,0)$, i.e., $\beta_{k,l}(u,v) \geq (1-\alpha_k u) \cdots (1-\alpha_l u)$. We now use that, for $x \leq \frac{2}{5}$, $1 - x \geq e^{-x-x^2}$:

$$\begin{aligned}\beta_{k,l}(u,v) &\geq e^{-au \sum_{j=l}^k \frac{1}{j+b} - a^2 u^2 \sum_{j=l}^k \frac{1}{(j+b)^2}} \\ &\geq \left(\frac{l+b-1}{k+b}\right)^{au} e^{-\frac{a^2 u^2}{l+b-1}}. \end{aligned} \tag{44}$$

This completes the proof of the lemma.

A.3 Proof of recursion (69)

Here we prove an alternative recursion on $\|X_k - x^\star\|$, used within the proof of Lemma 6. Specifically, we show that, for any k,

$$\|X_{k+1} - x^\star\| \leq \gamma_k \|X_k - x^\star\| + \alpha_k \|Z_k\|, \tag{45}$$

where, we recall, $\gamma_k = \left(1 - 2\alpha_k \mu + \alpha_k^2 L^2\right)^{1/2}$.

From the triangle inequality applied to the Euclidean norm,

$$\begin{aligned}\|X_{k+1} - x^\star\| &= \|X_k - \alpha_k g(X_k) + \alpha_k Z_k - x^\star\| \\ &\leq \|X_k - \alpha_k g(X_k) - x^\star\| + \alpha_k \|Z_k\| \end{aligned} \tag{46}$$

Exploiting L-smoothness and μ- convexity of f for the second term:

$$\begin{aligned}&\|X_k - \alpha_k g(X_k) - x^\star\|^2 \leq \\ &\|X_k - x^\star\|^2 - 2\alpha_k (X_k - x^\star)^\top g(X_k) + \alpha_k^2 \|g(X_k)\| \\ &\leq \|X_k - x^\star\|^2 - 2\alpha_k \mu \|X_k - x^\star\|^2 + \alpha_k^2 L^2 \|X_k - x^\star\|^2 \\ &= \gamma_k^2 \|X_k - x^\star\|^2. \end{aligned} \tag{47}$$

Taking the square root and replacing in (46) yields (45).

B PROOF OF LEMMA 5

First, we transform the recursion in (7) by defining $Y_{k+1} = (k + k_0)\|X_{k+1} - x^\star\|^2$, to obtain:

$$Y_{k+1} \leq a_k Y_k - b_k \sqrt{k + k_0 - 1} (X_k - x^\star)^\top Z_k + c_k \|Z_k\|^2, \tag{48}$$

where

$$a_k = \frac{k + k_0}{k + k_0 - 1} (1 - 2\alpha_k \mu + 2\alpha_k^2 L^2) \tag{49}$$

$$b_k = \frac{a}{\sqrt{k + k_0 - 1}} \tag{50}$$

$$c_k = \frac{a^2}{k + k_0}. \tag{51}$$

The key technical result behind Lemma 5 is the following upper bound on the tail probability of the Y_k iterates:

$$\mathbb{P}\left(Y_k \geq \epsilon\right) \leq e e^{-B\epsilon}, \tag{52}$$

which holds for each $k \geq 1$, and $\epsilon \geq 0$. The result of Lemma 5 directly follows from (52) by taking $\epsilon_k = k\delta^2$, for each k.

Thus, in the remainder of the proof we focus on proving (24). It can be easily verified that, for each k,

$$a_k = 1 - \frac{2a\mu - 1}{k + k_0 - 1} \left(1 - \frac{2a^2 L^2}{(2a\mu - 1)(k + k_0 - 1)}\right). \tag{53}$$

Recalling Assumption 2 and the value of k_0, we see that the above quantity is smaller than 1 for each k.

Denote by Φ_k the moment generating function of Y_k, and by $\Phi_{k+1|k}(\cdot; X_k)$ the moment generating function of Y_k conditioned on X_k :

$$\Phi_k(\nu) := \mathbb{E}\left[\exp(\nu Y_k)\right] \tag{54}$$

$$\Phi_{k+1|k}(\nu; X_k) := \mathbb{E}\left[\exp(\nu Y_k) \middle| X_k\right], \tag{55}$$

for $\nu \in \mathbb{R}$; note that $\Phi_{k+1}(\nu) = \mathbb{E}\left[\Phi_{k+1|k}(\nu; X_k)\right]$, for each $\nu \in \mathbb{R}$. From the recursion (7), we have:

$$\begin{aligned}&\Phi_{k+1|k}(\nu; X_k) = \\ &\exp(a_k \nu Y_k) \mathbb{E}\left[\exp(-b_k \sqrt{k + k_0 - 1} (X_k - x^\star)^\top Z_k + c_k \|Z_k\|^2) \middle| X_k\right] \\ &\leq \exp(a_k \nu Y_k) \left(\mathbb{E}\left[\exp(-2b_k \nu \sqrt{k + k_0 - 1} (X_k - x^\star)^\top Z_k) \middle| X_k\right]\right)^{1/2} \times \\ &\quad \left(\mathbb{E}\left[\exp(2c_k \nu \|Z_k\|^2) \middle| X_k\right]\right)^{1/2} \\ &\leq \exp(a_k \nu Y_k) \exp(2b_k^2 \nu^2 Y_k) \left(\mathbb{E}\left[\exp(2c_k \nu \|Z_k\|^2) \middle| X_k\right]\right)^{1/2} \end{aligned} \tag{56}$$

Recalling (2), the last term is finite for $\nu \leq 1/(2a^2C_2) =: B_0$, and for such ν, the corresponding value is equal to $\exp(C_2c_k\nu)$. Thus, for each $\nu \leq B_0$,

$$\Phi_{k+1|k}(\nu; X_k) \leq \exp(\nu(a_k + 2b_k^2\nu)Y_k + C_2c_k\nu). \tag{57}$$

It is easy to see that $B \leq B_0$. Consider $\nu \leq B$. Taking the expectation on both sides of (57), the following recursive inequality on Φ_k is obtained for any $\nu \leq B$ and any $k \geq 1$:

$$\Phi_{k+1}(\nu) \leq \Phi((a_k + 2b_k^2B)\nu)\exp(C_2c_k\nu). \tag{58}$$

From this point, the proof proceeds similarly as in (Harvey et al. 2019), i.e., by induction, and using $k = 1$ as the base, it can be shown that, for each $\nu \leq B$,

$$\Phi_k(\nu) \leq e^{\frac{\nu}{B}}. \tag{59}$$

By exponential Markov, from (59), for each $\nu \leq B$,

$$\mathbb{P}\left(Y_k \geq \epsilon\right) \leq \mathbb{E}\left[\exp \nu Y_k e^{-\nu\epsilon}\right]. \tag{60}$$

Taking $\nu = B$ yields the desired result.

C PROOF OF LEMMA 6

Fix $\lambda \in \mathbb{R}^d$. Fix $k \geq 1$. Define $\eta_l = B_{k,l}\eta_k$, $B_{k,l} = (I - \alpha_l H^\star)\cdots(I - \alpha_k H^\star)$, $\eta_k = k\lambda$. By Lemma 2,

$$\|\eta_l\| \leq k\left(\frac{l + k_0}{k + k_0 + 1}\right)^{a\mu}\|\lambda\| \leq (l + k_0)\|\lambda\|. \tag{61}$$

For an arbitrary $l \leq k$, there holds

$$\begin{aligned}
\Gamma_{l+1}(\eta_l) &= \mathbb{E}\left[\exp\left(\eta_l^\top(X_{l+1} - x^\star)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\eta_l^\top(X_l - \alpha_l g(X_l) + \alpha_l Z_l - x^\star)\right)|X_l\right]\right] \\
&= \mathbb{E}\left[\exp\left(\Lambda(\alpha_l\eta_l; X_l) + \eta_l^\top(X_l - \alpha_l g(X_l) - x^\star)\right)\right] \\
&= \int_{x\in\mathbb{R}^d}\Gamma_{l+1|l}(\eta_l; x)\mu_l(dx),
\end{aligned} \tag{62}$$

where $\Gamma_{l+1|l}(\cdot; x)$ denotes the conditional moment generating function of X_{l+1}, given $X_l = x$. We now fix $\delta > 0$ (the exact value to be chosen later) and split the analysis in two cases: 1) $\mathcal{A}_{l,\delta} = \{X_l \in B_{x^\star}(\delta)\}$; and 2) $\mathcal{A}_{l,\delta}^{\mathrm{c}} = \{X_l \in B_{x^\star}^{\mathrm{c}}(\delta)\}$.

Introduce

$$\begin{aligned}
\Gamma_{l+1|\mathcal{A}_{l,\delta}}(\eta_l) &:= \mathbb{E}\left[1_{\|X_l - x^\star\|\leq\delta}\Gamma_{l+1|l}(\eta_l; X_l)\right] \\
&= \int_{\|x - x^\star\|\leq\delta}\Gamma_{l+1|l}(\eta_l; x)\mu_l(dx) \qquad (63) \\
\Gamma_{l+1|\mathcal{A}_{l,\delta}^{\mathrm{c}}}(\eta_l) &:= \mathbb{E}\left[1_{\|X_l - x^\star\|>\delta}\Gamma_{l+1|l}(\eta_l; X_l)\right] \\
&= \int_{\|x - x^\star\|>\delta}\Gamma_{l+1|l}(\eta_l; x)\mu_l(dx); \qquad (64)
\end{aligned}$$

note that

$$\Gamma_{l+1}(\eta_l) = \Gamma_{l+1|A_{l,\delta}}(\eta_l) + \Gamma_{l+1|A_{l,\delta}^{\mathrm{c}}}(\eta_l). \tag{65}$$

C.1 Case 1: $x \in \mathcal{A}_{l,\delta}$.

Fix $x \in \mathbb{R}^d$ such that $\|x\| \leq \delta$. We have:

$$\begin{aligned}
\Gamma_{l+1|l}(\eta_l; x) &= \exp\left(\Lambda(\alpha_l\eta_l; x) + \eta_l^\top(x - \alpha_l g(x) - x^\star)\right) \\
&= \exp\left(\Lambda(\alpha_l\eta_l; x) + \eta_l^\top((I - \alpha_l H^\star)(x - x^\star) - \alpha_l h(x)\right) \\
&\leq \exp\left(\Lambda(\alpha_l\eta_l; x^\star) + L_\Lambda\|\eta_l^2\|\|x - x^\star\| + \alpha_l\|\eta_l\|\|h(x)\|\right) \\
&\quad\times\exp\left(\eta_l^\top((I - \alpha_l H^\star)(x - x^\star)\right) \qquad (66) \\
&\leq \exp\left(\Lambda(\alpha_l\eta_l; x^\star) + L_\Lambda\alpha_l^2\|\eta_l^2\|\delta^2 + \alpha_l\|\eta_l\|\overline{h}(\delta) + \eta_{l-1}^\top(x - x^\star)\right), \qquad (67)
\end{aligned}$$

where in (66) we used Lipschitz continuity of Λ in x, Assumption 4, and in (67) we used the fact that $\|x - x^\star\| \leq \delta$. It follows that

$$\Gamma_{l+1|\mathcal{A}_\delta}(\eta_l) \leq \exp\left(\Lambda(\alpha_l\eta_l; x^\star) + r_0(\lambda, \delta)\right)\Gamma_l(\eta_{l-1}), \tag{68}$$

where $r_0(\lambda, \delta) = L_\Lambda a^2\|\lambda\|^2\delta^2 + a\|\lambda\|\overline{h}(\delta)$.

C.2 Case 2: $x \in \mathcal{A}^{\mathrm{c}}_{l,\delta}$

. By strong convexity and Lipschitz smoothness of f in Assumption 1, for each $l \geq 1$, the following holds:

$$\|X_l - g(X_l) - x^\star\| \leq \gamma_l\|X_l - x^\star\|, \tag{69}$$

$$\leq \overline{\gamma}\|X_l - x^\star\| \tag{70}$$

where $\gamma_l = (1 - 2\alpha_l\mu + \alpha_l^2L^2)^{1/2}$, see Appendix A for the proof, and $\overline{\gamma} = \sup\{\gamma_l : l = 1, 2, ...\}$; it is easy to verify that $\overline{\gamma} = \max\{1, \sqrt{(1 - a\mu)^2 + a^2(L^2 - a^2)}\}$.

For an arbitrary $x \in \mathbb{R}^d$, we have:

$$\begin{aligned}\Gamma_{l+1|l}(\eta_l; x) &= \exp\left(\Lambda(\alpha_l\eta_l; x) + \eta_l^\top(x - \alpha_l g(x) - x^\star)\right)\\ &\leq \exp\left(\frac{C_1\alpha_l^2\|\eta_l\|^2}{2}\right)\exp\left(\overline{\gamma}\|\eta_l\|\|x - x^\star\|\right), \qquad (71)\\ &\leq \exp\left(\frac{C_1a^2\|\lambda\|^2}{2}\right)\exp\left(\overline{\gamma}(l + k_0)\|\lambda\|\|x - x^\star\|\right), \qquad (72)\end{aligned}$$

where in (71) we used the assumption that Z_k is sub-Gaussian, Assumption 5, for the first term, together with (69) and Cauchy-Schwartz, for the second term, while in (72) we exploited (61). Recalling the induced measure ν_l, we now have

$$\Gamma_{l+1|\mathcal{A}^{\mathrm{c}}_{l,\delta}}(\eta_l) \leq \exp\left(\frac{C_1a^2\|\lambda\|^2}{2}\right)\int_{z\geq\delta} e^{(l+k_0)\overline{\gamma}\|\lambda\|z}\nu_l(dz). \tag{73}$$

The idea of analysing the "tail" term $\Gamma_{l+1|\mathcal{A}^{\mathrm{c}}_{l,\delta}}(\eta_l)$ is the following: by Lemma 5, we know that the probability density ν_l at a given point z behaves roughly as $e^{-(l+k_0)Bz^2}$. If δ is sufficiently large, then, for all $z \geq \delta$, the negative exponential rate of the measure $\nu_l(z)$ is in absolute terms higher than the exponent $(l + k_0)\overline{\gamma}\|\lambda\|z$. Integrating by parts, we obtain that for $\delta = \frac{2\overline{\gamma}\|\lambda\|}{B}$, the integral on the right hand-side of (73) is upper bounded by a constant K. Thus:

$$\Gamma_{l+1|\mathcal{A}^{\mathrm{c}}_{l,\delta}}(\eta_l) \leq K\exp\left(\frac{C_1a^2\|\lambda\|^2}{2}\right). \tag{74}$$

Combining with (68) and recalling (65),

$$\Gamma_{l+1}(\eta_l) \leq \exp\left(\Lambda(\alpha_l\eta_l; x^\star) + r(\lambda)\right)\Gamma_l(\eta_{l-1}) + K\exp\left(\frac{C_1a^2\|\lambda\|^2}{2}\right),$$

where $r(\lambda) = r_0\left(\lambda, \frac{2\overline{\gamma}\|\lambda\|}{B}\right)$. Iterating the preceding recursion, where we exploit the nonnegativity of Λ, property 3. from Lemma 3, we obtain:

$$\begin{aligned}\Gamma_{k+1}(k\lambda) &\leq \exp\left(\sum_{l=1}^k\left(\Lambda(\alpha_l\eta_l; x^\star) + r(\lambda)\right)\right)\Gamma_1(\alpha_1\eta_1)\\ &+ K\exp\left(\frac{C_1a^2\|\lambda\|^2}{2}\right)\sum_{l=1}^k e^{\sum_{j=l}^k(\Lambda(\alpha_j\eta_j; x^\star) + r(\lambda))}\\ &\leq (k+1)K\exp\left(\frac{C_1a^2\|\lambda\|^2}{2} + a\|\lambda\|\|X_1 - x^\star\|\right)\\ &\quad\times e^{\sum_{l=1}^k(\Lambda(\alpha_l\eta_l; x^\star) + r(\lambda))}. \qquad (75)\end{aligned}$$

Taking the limit, dividing by k, and taking the $\limsup$

$$\limsup_{k\to+\infty} \frac{1}{k}\log\Gamma_{k+1}(k\lambda) \leq$$
$$r(\lambda) + \limsup_{k\to+\infty} \frac{1}{k}\sum_{l=1}^{k}\Lambda(\alpha_l\eta_l; x^\star). \tag{76}$$

To complete the proof of the lemma, we next show that

$$\lim_{k\to+\infty} \frac{1}{k}\sum_{l=1}^{k}\Lambda(\alpha_l\eta_l; x^\star) = \int_0^1 \Lambda(aQD(\theta)Q^\top\lambda; x^\star)d\theta. \tag{77}$$

C.3 Proof of (77)

Introduce step-wise constant function $s_k : [0,1] \mapsto \mathbb{R}$, defined by

$$s_k(\theta) = \begin{cases} \Lambda(\alpha_l\eta_l; x^\star), & \text{for } \frac{l-1}{k} < \theta \leq \frac{l}{k} \\ 0, & \text{for } \theta = 0 \end{cases}. \tag{78}$$

It is easy to verify that the integral of s_k over $[0,1]$ equals the desired sum in the right hand-side of (76), i.e.,

$$\int_0^1 s_k(\theta) = \frac{1}{k}\sum_{l=1}^{k}\Lambda(\alpha_l\eta_l; x^\star). \tag{79}$$

We next show that

$$\lim_{k\to+\infty} s_k(\theta) = \Lambda(aQD(\theta)Q^\top\lambda; x^\star), \tag{80}$$

where $D(\theta)$ is as defined in the claim of the theorem. To show the preceding limit, note that, for each $\theta \in (0,1]$,

$$s_k(\theta) = \Lambda(k\alpha_{l_k}QD_{k,l_k}Q^\top\lambda; x^\star) \tag{81}$$

where $[D_{k,l_k}]_{ii} = \beta_{k,l_k}(\rho_i, 0)$, l_k is the index of the interval in the definition of s_k to which θ belongs, $l_k = \min\{l = 1,..,k : \theta \leq \frac{l}{k}\}$, and ρ_i is, we recall, the i-th eigenvalue of $H^\star$.

Using the bounds from Lemma 2, it is easy to establish the by sandwiching argument that

$$\lim_{k\to\infty} k\alpha_{l_k}\beta_{k,l_k}(\rho_i, 0) = a\theta^{a\rho_i - 1}. \tag{82}$$

The limit in (80) now follows by the continuity of $\Lambda(\cdot; x^\star)$, which follows by convexity of $\Lambda(\cdot; x^\star)$, Lemma 3.

Using the fact that s_k can be uniformly bounded for all k and $\theta \in [0,1]$, we can exchange the order of the limit and the integral, to obtain:

$$\lim_{k\to+\infty}\int_0^1 s_k(\theta)d\theta = \int_0^1 \lim_{k\to+\infty} s_k(\theta)d\theta = \int_0^1 \Lambda(aQD(\theta)Q^\top\lambda; x^\star), \tag{83}$$

establishing the claim in (77). This completes the proof of Lemma 6.

D DERIVATIONS FROM REMARKS

D.1 Derivations from Remark 7

Consider function $\overline{\Psi}(\lambda) = \Psi^\star(\lambda) + r(\lambda)$ in Lemma 6. We derive here a lower bound on rate function $\overline{I}$ in Theorem 1 that does not explicitly depend on $H(x^\star)$. In view of the fact that $\overline{I}$ is the Fenchel-Legendre transform of $\overline{\Psi}$, a lower bound on $\overline{I}$ is readily obtained by deriving an upper bound on $\overline{\Psi}(\lambda)$. Note that $r(\lambda)$ does not explicitly depend on $H(x^\star)$, hence we only need to derive an upper bound on $\Psi^\star$. By Assumption 5, we have, for any $\theta \in [0,1]$, that $\Lambda(aQD(\theta)Q^\top\lambda; x^\star) \leq \frac{C_1\, a^2}{2}\lambda^\top(QDQ^\top)^2\lambda \leq \frac{C_1\, a^2}{2}\|\lambda\|^2\,\|D(\theta)\|^2$, where we recall that $\|\cdot\|$ denotes the 2-norm of its vector or matrix argument. Next, note that $\|D(\theta)\| \leq \theta^{a\,\mu-1}$, for all $\theta \in [0,1]$, because all eigenvalues ρ_i's of $H(x^\star)$ belong to the interval $[\mu, L]$. Therefore, we obtain:

$$\Psi^\star(\lambda) \leq \frac{C_1\, a^2}{2}\|\lambda\|^2\int_0^1 \theta^{2a\,\mu-2}d\theta = \frac{C_1\, a^2}{2(2\,a\mu-1)}\|\lambda\|^2.$$

D.2 The case of random initial iterate X_1

Recall that, by definition, $\Gamma_1(\lambda) = \mathbb{E}\left[e^{\lambda^\top (X_1 - x^\star)}\right]$. When X_1 is random, Γ_1, as a function of λ, is therefore the log-moment generating function of $X_1 - x^\star$. Provided its domain is $\mathbb{R}^d$, all arguments in the proof of Theorem 1 remain the same. In particular, in eq. (75), the factor $e^{\|\lambda\| \|X_1 - x^\star\|}$ would be replaced by a (finite-valued) function (of λ), and the subsequent results would be unaltered; a similar comment applies for the statement and the proof of Lemma 5.

E PROOF OF THEOREM 2

It is easy to show that for the assumed quadratic form, the iterates X_k have the following representation:

$$X_{k+1} = A_{k0} X_1 + \sum_{l=1}^{k} \alpha_l A_{k,l+1} Z_l, \tag{84}$$

where $A_{k,l} = \prod_{j=l}^{k} (I - \alpha_j H)$. By the assumption that the noise realizations at different times are independent and with a constant distribution, we obtain:

$$\Gamma_{k+1}(\lambda) = e^{\lambda^\top X_1} e^{\sum_{l=1}^{k} \Lambda(\alpha_l A_{k,l+1} \lambda)}. \tag{85}$$

The proof now follows from Lemma 4 and the limit established in 77.

F NUMERICAL RESULTS

We now illustrate the achieved results through a numerical simulation. We consider a strongly convex quadratic cost function $f : \mathbb{R}^d \to \mathbb{R}$, defined by $f(x) = \frac{1}{2} x^\top A x + bx$, $d = 10$, where the symmetric $d \times d$ matrix A and the $d \times 1$ vector b are generated randomly. Specifically, we generate the entries of b mutually independently, according to the standard normal distribution. The matrix A is generated as follows. We let $A = Q \Lambda Q^\top$, where Q is the matrix whose columns are the orthonormal eigenvectors of matrix $(B + B^\top)/2$, and the entries of B are drawn mutually independently from the standard normal distribution; the matrix Λ is the diagonal matrix whose diagonal entries are drawn from the uniform distribution on the interval $[1, 2]$. Clearly, the optimal solution for the problem equals $x^\star = A^{-1} b$.

We consider the gradient noise that is generated in an i.i.d. manner over iterations and over the gradient noise vector elements, independently from the solution iterate sequence. Two different noise distributions per gradient noise entry are considered, such that the per-entry noise variance is kept equal for the two distributions, equal to σ^2. In this way, we evaluate the effects of higher order moments on the performance of SGD. The first distribution is zero-mean Gaussian with variance σ^2. The second distribution is the zero-mean Laplacian with the same variance. We set $\sigma^2 = 0.04$.

We numerically estimate, via Monte Carlo simulations, the probability $P(\|X_k - x^\star\| > \delta)$ along iterations $k = 1, 2, ...$ We denote the corresponding numerical estimate by p_k. Two different values of δ are considered, $\delta = 0.3$, and $\delta = 0.03$. For each Monte Carlo run, X_1 is set to the zero vector. For the numerical example here, $\|x^\star\| = 2.342$, and hence $\delta = 0.3$ corresponds to the relative error level $\delta / \|x^\star\| \approx 0.13$, while $\delta = 0.03$ corresponds to $\delta / \|x^\star\| \approx 0.013$. Figure 1 plots p_k versus iteration counter k (in linear scale for the horizontal axis, and $\log_{10}$-scale for the vertical axis) for the Gaussian noise case (blue line) and the Laplacian noise case (red line). The top Figure is for $\delta = 0.3$, and the bottom Figure is for $\delta = 0.03$. We can see that, for a large value of δ, the two curves are very different: the Laplacian gradient noise case leads to a worse performance. This is because, for large δ, the argument λ of the LMGF Λ that corresponds to the minimizer in the rate function value $I^\star$ is large (see Theorem 1), and hence higher order polynomial coefficients ($\sim \lambda^3$ and higher) play a significant role. As the higher order moments of the Gaussian and Laplace distributions are very different (equal to zero for the Gaussian and strictly positive for the Laplacian), the result is the different large deviations performance (worse for the Laplacian case) as seen in Figure 1, top. On the other hand, for a small value of δ (bottom, Figure 1), the argument λ of the LMGF that corresponds to the minimizer in the rate function expression $I^\star$ is small, and hence only the first two order polynomial coefficients of Λ play a significant role. As the two distributions here are both zero mean and have equal variance (hence having equal first and second order moments), the large deviation performance for the two noises matches, as seen in Figure 1, bottom. This behavior is in accordance with the theory derived.

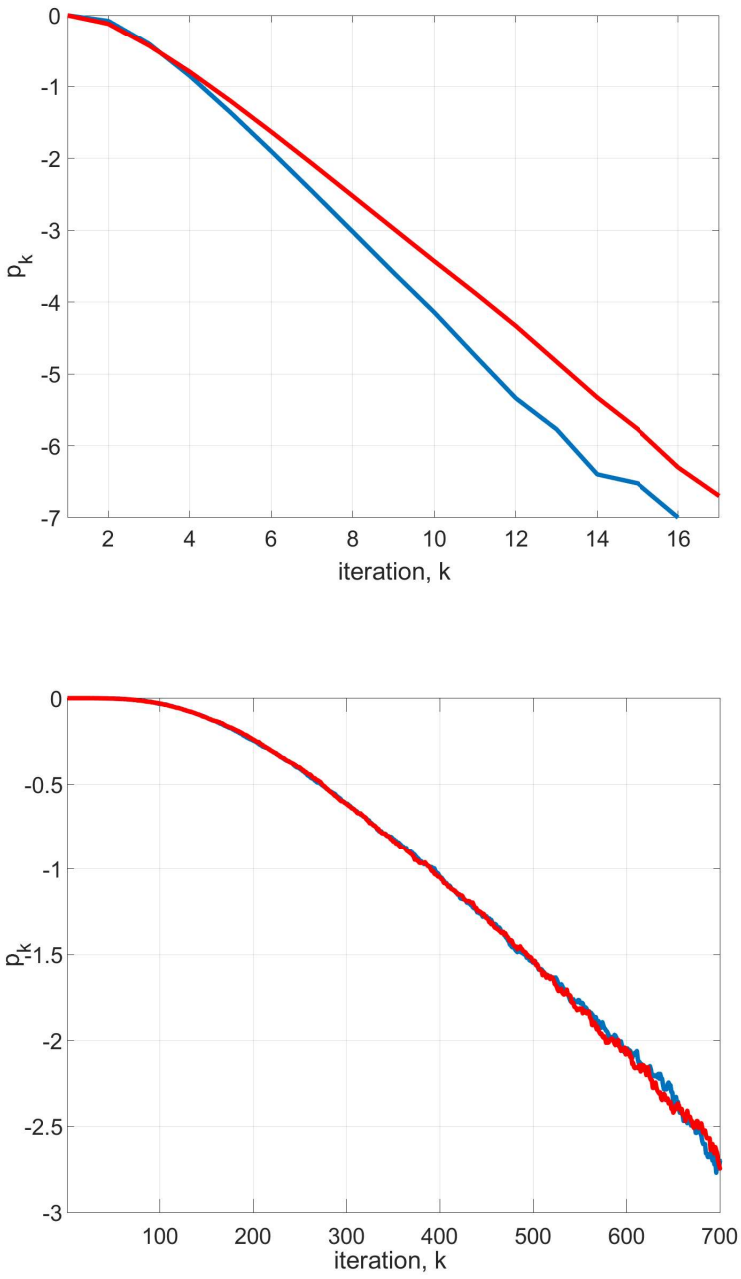

Figure 1: Monte Carlo estimate of $P\left(\|X_k - x^\star\| > \delta\right)$ along iterations $k = 1, 2, ...$ for SGD with Gaussian (blue line) and Laplacian (red line) gradient noise with equal per-entry variance $\sigma^2 = 0.04$. Top Figure: $\delta = 0.3$; Bottom Figure: $\delta = 0.03$.