Continuous-Time Stochastic Mirror Descent on a Network: Variance Reduction, Consensus, Convergence

Maxim Raginsky and Jake Bouvrie

***Abstract*— The method of Mirror Descent (MD), originally proposed by Nemirovski and Yudin in the late 1970s, has recently seen a major resurgence in the fields of large-scale optimization and machine learning. In a nutshell, MD is a primal-dual method that can be adapted to the geometry of the optimization problem at hand through the choice of a suitable strongly convex potential function. We study a stochastic, continuous-time variant of MD performed by a network of coupled noisy agents (processors). The overall dynamics is described by a system of stochastic differential equations, coupled linearly through the network Laplacian. We address the impact of the network topology (encoded in the spectrum of the Laplacian) on the speed of convergence of the "mean-field" component to the optimum. We show that this convergence is particularly rapid whenever the potential function can be chosen in such a way that the resulting mean-field dynamics in the dual space follows an Ornstein–Uhlenbeck process.**

I. Introduction

Large-scale optimization problems, involving anywhere up to millions of variables, are becoming ubiquitous in control, machine learning, communications, signal/image processing, and other areas of science and engineering. This development has highlighted the importance of *structure* in the problem formulation as an enabler of efficiently implementable optimization schemes. In particular, most of the recent progress in the above-mentioned fields can be traced to the recognition that many (if not all) optimization problems of interest can be cast as *convex programs* of the form

$$\min\{f(x) : x \in \mathsf{X}\}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, and the problem domain $\mathsf{X} \subseteq \mathbb{R}^n$ is a closed convex set. Thus, in principle, one can take advantage of powerful polynomial-time convex programming schemes, such as interior-point methods [1]. However, the per-iteration computational complexity of such methods scales nonlinearly with the problem dimension n, at least as $O(n^2)$ and typically as $O(n^3)$, unless the problem at hand has a very "nice" (e.g., sparse) structure [1]. This unfortunate feature of "fancy" convex programming methods rules out their use whenever n is already on the order of tens of thousands. In this large-scale regime, simpler iterative methods with linear per-iteration complexity become more attractive.

Perhaps the simplest and the best-known of such methods is *projected subgradient descent* [2], [3], which starts with some initial point $x_0 \in \mathsf{X}$ and iteratively generates the points

$$x_{k+1} = \Pi_{\mathsf{X}}\left(x_k - \gamma_k f'(x_k)\right), \tag{1}$$

where $\{\gamma_k\}_{k=0}^{\infty}$ is a sequence of positive step sizes, $f'(x_k)$ is an arbitrary subgradient[1] of f at $x_k \in \mathsf{X}$, and Π_{X} is the Euclidean projection onto X, i.e., $\Pi_{\mathsf{X}}(y) = \arg\min_{x\in\mathsf{X}}\{\|y - x\|_2\}$ where $\|\cdot\|_2$ denotes the Euclidean (ℓ_2) norm on $\mathbb{R}^n$. Provided this projection is efficiently computable, the above method is easy to implement. However, its main drawback is slow convergence. For instance, when X is compact and the objective f is Lipschitz on X with constant L, the updates $\{x_k\}$ generated by (1) with properly tuned step sizes satisfy

$$f(x_k) - \min_{x\in\mathsf{X}} f(x) \le \text{const} \cdot \frac{LD_{\mathsf{X}}}{\sqrt{k}}, \tag{2}$$

where $D_{\mathsf{X}} \triangleq \sup_{x,x'\in\mathsf{X}} \|x - x'\|_2$ is the ℓ_2 diameter of X. Moreover, this rate of convergence cannot be improved [5]. Thus, when X is an ℓ_2 ball, its diameter [and hence the convergence rate in (2)] is independent of n; however, in the case of a cube (ℓ_∞ ball) $\mathsf{X} = [-1, 1]^n$, we have $D_{\mathsf{X}} = 2\sqrt{n}$, so the convergence rate deteriorates as n grows.

This phenomenon is due to the fact that the subgradient scheme (1) is inextricably tied to the Euclidean geometry of $\mathbb{R}^n$ through the projection operator Π_{X}. The so-called method of Mirror Descent (MD), proposed in the late 1970s by Nemirovski and Yudin [5] (cf. also [6], [7]) is a substantial generalization of (1) that can be flexibly tailored to the geometry of X, potentially resulting in a much better dependence of the rate of convergence on the problem dimension n. As an example, MD has been applied to a tomographic reconstruction problem in medical imaging [8], which can be cast as a convex optimization problem over the standard simplex in $\mathbb{R}^n$, giving the convergence rate of $O\left(\sqrt{\frac{\ln n}{k}}\right)$ and thus outperforming the Euclidean subgradient method (1) by a factor of $\sqrt{n/\ln n}$. Moreover, MD has been recently demonstrated to be a "universal" scheme (in the sense of achieving optimal convergence rates) for online convex optimization problems arising in machine learning [9].

M. Raginsky is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA; maxim@illinois.edu. Research supported by NSF under grant CCF–1017564.

J. Bouvrie is with the Department of Mathematics, Duke University, Durham, NC 27708, USA; jvb@math.duke.edu. JB acknowledges support under contracts FA8650-11-1-7150 SUB#7-3130298 (DARPA), SUB#113054 G002745 (Washington State U.) and IIS-08-03293 (NSF), all to M. Maggioni.

[1] A subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at a given point x in the domain of f is any vector $g \in \mathbb{R}^n$, such that $f(y) \ge f(x) + \langle g, x - y\rangle$ for all $y \in \mathbb{R}^n$. The set of all subgradients of f at x is called the *subdifferential* of f at x and denoted by $\partial f(x)$ [4].

The original motivation for MD, given by Nemirovski and Yudin [5], can be roughly summarized as follows. Suppose that the convex objective function $f : \mathbb{R}^n \to \mathbb{R}$ is smooth, and consider the gradient flow

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\nabla f(x(t)), \qquad t \geq 0 \tag{3}$$

with some initial condition $x(0) \in \mathsf{X}$. Let $x^* \in \arg\min_{\mathsf{X}} f$ be any minimizer of f on X. It is a well-known fact that $V_t(x^*) = \frac{1}{2}\|x(t) - x^*\|_2^2$ is a *Lyapunov function* for (3), i.e., $\mathrm{d}V_t(x^*)/\,\mathrm{d}t \leq 0$ for all $t \geq 0$. From this it can be shown that $f(x(t)) \to f(x^*) \equiv \min_{x\in\mathsf{X}} f(x)$ as $t \to \infty$. Moreover, the subgradient scheme (1) can be viewed as a discretization of (3). The main insight behind MD is that it is possible to proceed in the opposite direction: first choose an appropriate Lyapunov function for a given X, and then obtain the corresponding MD recursion as a discretization of a certain continuous-time process that involves the gradient of f and has the specified Lyapunov function.

This continuous-time process (described in detail in Section II-A) was introduced by Nemirovski and Yudin merely as a heuristic device for motivating the discrete-time MD scheme. Indeed, to the best of our knowledge, all studies of MD to date have dealt with the discrete-time formulation. However, the continuous-time MD dynamics may be of interest in its own right (e.g., in the context of control systems [10], or wherever sensor signals and noise sources may be inherently continuous quantities). Indeed, the behavior of many discrete-time approaches to cooperative optimization in the presence of noise may be characterized by studying continuously perturbed dynamical systems, the discretizations of which provide practical algorithms.

The objective of the present paper is to study a noisy variant of continuous-time MD, described by an Itô stochastic differential equation (SDE) [11]. In particular, we show that the favorable convergence properties of noiseless continuous-time MD are adversely affected by the addition of a white-noise perturbation (we do not consider a small-noise limit here). This observation then motivates a *distributed* implementation of MD by a network of noisy agents (processors), coupled linearly through the Laplacian of the network graph [12]. This coupling serves two purposes: (1) it helps reduce the noise variance; and (2) provided the underlying network graph is connected, the agents converge to consensus. We show that this convergence is particularly rapid whenever the problem structure is such that the "mean field" (i.e., the average of the agents' trajectories) evolves according to an Ornstein–Uhlenbeck process. Of course, in this linear regime noise reduction can be achieved simply by averaging of the agents' trajectories, without the need for any additional inter-agent coupling. This coupling, however, leads to a fully decentralized design: since the agents converge to consensus, we can track any given agent to obtain an accurate approximation to the optimum, without having to introduce a dedicated averaging unit.

A. Notation

For any two vectors $v, w \in \mathbb{R}^n$, $\langle v, w\rangle$ will denote their standard (Euclidean) inner product. Given an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^n$, we will denote by $B_{\|\cdot\|}$ the corresponding unit ball, i.e., $B_{\|\cdot\|} \triangleq \{v \in \mathbb{R}^n : \|v\| \leq 1\}$. The *dual norm* $\|\cdot\|_*$ is defined by $\|z\|_* \triangleq \sup\left\{\langle z, v\rangle : v \in B_{\|\cdot\|}\right\}$. Any pair of dual norms $\|\cdot\|, \|\cdot\|_*$ satisfies *Hölder's inequality*, $|\langle v, z\rangle| \leq \|v\| \cdot \|z\|_*$. In n dimensions, $\mathbf{1}_n$ will stand for the vector of all ones, while I_n will be the $n \times n$ identity matrix. We will denote by $\{W_t\}$ (possibly with additional sub- and superscripts) the standard one-dimensional Wiener process; similarly, $\{B_t\}$ will denote the standard n-dimensional Wiener process [11].

II. The method of Mirror Descent: preliminaries

We start by describing the discrete-time implementation of MD. Let $\mathsf{X} \subset \mathbb{R}^n$ be a compact convex decision set, and let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^n$. The structure of MD hinges on the concept of a *distance-generating function* (also referred to as the *potential function*) and its induced *Bregman divergence* [13]:

Definition 1 (Distance-generating function). *A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is a* distance-generating function (DGF) *with modulus $\alpha > 0$ w.r.t. $\|\cdot\|$, provided it has the following properties:*

- *It is convex and continuous on X.*
- *The set $\mathsf{X}^\circ = \{x \in \mathsf{X} : \partial\psi(x) \neq \varnothing\}$ is convex (in fact, it always contains the relative interior of X [7]).*
- *Restricted to X°, ψ is C^1 and* strongly convex *with parameter $\alpha > 0$, i.e., for all $x, x' \in \mathsf{X}$*

$$\psi(x') \geq \psi(x) + \langle\nabla\psi(x), x' - x\rangle + \frac{\alpha}{2}\|x' - x\|^2.$$

Definition 2 (Bregman divergence). *Let ψ be a DGF on X. Then the* Bregman divergence *induced by ψ is the function $D_\psi : \mathsf{X} \times \mathsf{X}^\circ \to \mathbb{R}$ defined by*

$$D_\psi(x, x') \triangleq \psi(x) - \psi(x') - \langle\nabla\psi(x'), x - x'\rangle.$$

As an example, take $\psi(x) = \frac{1}{2}\|x\|_2^2$ and any compact and convex X. Then ψ is a DGF w.r.t. $\|\cdot\|_2$ with $\alpha = 1$ and $\mathsf{X}^\circ = \mathsf{X}$, and $D_\psi(x, x') = \frac{1}{2}\|x - x'\|_2^2$. As another example, take $\mathsf{X} = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1; x \succeq 0\}$, the unit simplex in $\mathbb{R}^n$, and let $\|\cdot\| = \|\cdot\|_1$. Then the (negative) *entropy function* $\psi(x) = \sum_{i=1}^n x_i \ln x_i$ is a DGF w.r.t. $\|\cdot\|_1$ with $\alpha = 1$, $\mathsf{X}^\circ = \{x \in \mathsf{X} : x \succ 0\}$, and

$$D_\psi(x, x') = D_{\|\cdot\|_1}(x, x') = \sum_{i=1}^n x_i \ln\frac{x_i}{x'_i},$$

the relative entropy between $x \in \mathsf{X}$ and $x' \in \mathsf{X}^\circ$.

We say that a DGF $\psi : \mathsf{X} \to \mathbb{R}$ is *admissible* if we can efficiently compute its *Legendre–Fenchel conjugate* [4]

$$\psi^*(z) \triangleq \max_{x\in\mathsf{X}}\left\{\langle x, z\rangle - \psi(x)\right\}.$$

Given an admissible DGF and a sequence $\{\gamma_k\}_{k=0}^\infty$ of positive step sizes, we can now write down the generic MD update: starting with some z_0 such that $\|\nabla\psi^*(z_0)\| < \infty$

(a good choice is $z_0 = \arg\min_{\mathsf{X}} \psi$), generate the sequence $\{x_k\}_{k=0}^{\infty}$ via

$$x_k = \nabla\psi^*(z_k), \qquad z_{k+1} = \nabla\psi(x_k) - \gamma_k f'(x_k). \qquad (4)$$

This structure is what gives the MD method its name: the current point x_k is mapped to its dual-space "mirror image" $z_k = \nabla\psi(x_k)$, updated to z_{k+1} by stepping in the direction of the negative subgradient $-f'(x_k)$, and then mapped back to $x_{k+1} = \nabla\psi^*(z_{k+1})$ in the primal space X. The rate of convergence of MD can then be bounded as follows: Define the *ψ-diameter* of X,

$$D_{\psi,\mathsf{X}} \triangleq \sup_{x \in \mathsf{X}, x' \in \mathsf{X}^\circ} \sqrt{2D_\psi(x,x')}.$$

Then, provided f is L-Lipschitz on X w.r.t. $\|\cdot\|$, i.e., $|f(x) - f(x')| \le L\|x - x'\|$ for all $x, x' \in \mathsf{X}$, the MD updates (4) with suitably tuned step sizes $\gamma_k > 0$ satisfy

$$f(x_k) - \min_{x \in \mathsf{X}} f(x) \le \text{const} \cdot \frac{D_{\psi,\mathsf{X}} L}{\sqrt{\alpha k}},$$

and this convergence rate is optimal [5]. Thus, if we can find a suitable DGF ψ, so that $D^2_{\psi,\mathsf{X}}$ is either independent of n or (at least) sublinear in n, then the dimension dependence of the convergence rate of MD will be better than that of the Euclidean subgradient descent (1).

A. MD in continuous time

To give the intuition behind MD, Nemirovski and Yudin [5] focused on the case when the objective function f is convex and smooth (say, C^1), and considered the ODE

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = h(z(t)), \qquad t \ge 0 \qquad (5)$$

where $h(z) \triangleq -\nabla f(\nabla\psi^*(z))$. [Note that for $\psi(x) = \frac{1}{2}\|x\|_2^2$, (5) reduces to the gradient flow (3).] If we let $x(t) = \nabla\psi^*(z(t))$, then a suitable discretization of (5) gives the following simplified version of MD:

$$x_k = \nabla\psi^*(z_k), \qquad z_{k+1} = z_k - \gamma_k f'(x_k).$$

Let us assume that the DGF ψ is chosen in such a way that its conjugate ψ^* is C^1 and that the gradient $\nabla\psi^*$ maps $\mathbb{R}^n$ onto X°. This will be the case, for example, if ψ is *steep*, i.e., if for any sequence $\{x_n\}$ of points in X° converging to a point on the boundary of X, $\|\nabla\psi(x_n)\| \to \infty$ as $n \to \infty$. (In fact, when ψ is steep, the gradient mappings $\nabla\psi : \mathsf{X}^\circ \to \mathbb{R}$ and $\nabla\psi^* : \mathbb{R}^n \to \mathsf{X}^\circ$ are inverses of one another.) We can now establish the following basic property of (5):

Proposition 1. *For any $z^* \in \mathbb{R}^n$ such that $x^* = \nabla\psi^*(z^*) \in \arg\min_{\mathsf{X}} f$, define*

$$V_t(z^*) \triangleq \psi^*(z(t)) - \psi^*(z^*) - \langle \nabla\psi^*(z^*), z(t) - z^* \rangle.$$

Then $V_t(z^*)$ is a Lyapunov function for the dynamics (5)*, i.e., $\mathrm{d}V_t(z^*)/\,\mathrm{d}t \le 0$ along the trajectory $\{z(t)\}_{t \ge 0}$.*

Remark 1. Note that $V_t(z^*)$ is itself a Bregman divergence $D_{\psi^*}(z(t), z^*)$ induced by ψ^*. Since ψ^* is convex, $V_t(z^*) \ge 0$. Moreover, because ψ is steep, $V_t(z^*) = D_\psi(x^*, x(t))$. $\diamond$

Proof. Direct calculation:

$$\begin{aligned} \frac{\mathrm{d}V_t(z)}{\mathrm{d}t} &= \left\langle \nabla\psi^*(z(t)) - \nabla\psi^*(z^*), \frac{\mathrm{d}z(t)}{\mathrm{d}t} \right\rangle \\ &= \langle x(t) - x^*, h(z(t)) \rangle \\ &= \langle x^* - x(t), \nabla f(x(t)) \rangle \\ &\le f(x^*) - f(x(t)) \le 0, \end{aligned}$$

where the fourth step uses the convexity of f, while the one before uses the fact that $x(t) = \nabla\psi^*(z(t)) \in \mathsf{X}$. □

In fact, we can estimate the rate of convergence:

Proposition 2. *For any $T > 0$,*

$$\inf_{0 \le t \le T} f(x(t)) - \min_{\mathsf{X}} f \le \frac{D^2_{\psi,\mathsf{X}}}{2T} \qquad (6)$$

and, with $\overline{x}_T \triangleq \frac{1}{T}\int_0^T x(t)\,\mathrm{d}t$,

$$f(\overline{x}_T) - \min_{\mathsf{X}} f \le \frac{D^2_{\psi,\mathsf{X}}}{2T}. \qquad (7)$$

Proof. Starting from $\mathrm{d}V_t(z^*)/\,\mathrm{d}t = \langle x^* - x(t), \nabla f(x(t)) \rangle$, integrate from $t = 0$ to $t = T$:

$$\begin{aligned} V_T(z^*) &= V_0(z^*) + \int_0^T \langle x^* - x(t), \nabla f(x(t)) \rangle\,\mathrm{d}t \\ &\le V_0(z^*) + \int_0^T [f(x^*) - f(x(t))]\,\mathrm{d}t, \end{aligned}$$

where the second step uses the convexity of f. Now rearrange and divide by T to get

$$\frac{1}{T}\int_0^T [f(x(t)) - f(x^*)]\,\mathrm{d}t \le \frac{V_0(z^*) - V_T(z^*)}{T} \le \frac{D^2_{\psi,\mathsf{X}}}{2T},$$

where the last step is due to the fact that $V_t(z^*) \ge 0$ for all $t \ge 0$, and (cf. Remark 1) that $V_0(z^*) = D_\psi(x^*, x(0)) \le D^2_{\psi,\mathsf{X}}/2$. The bound (6) now follows immediately; (7) is a consequence of Jensen's inequality. □

III. Continuous-time stochastic MD

As already pointed out in the Introduction, the continuous-time dynamics (5) was used in [5] merely to provide intuition for the discrete-time MD scheme (4). However, given the superior convergence properties of (5) listed in Proposition 2, continuous-time implementations of MD may be of interest in their own right. Motivated by this observation, let us consider a *noisy* version of MD, given by the Itô SDE

$$\mathrm{d}Z_t = h(Z_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t, \qquad t \ge 0 \qquad (8)$$

where, as before, $h(z) = -\nabla f(\nabla\psi^*(z))$ for a given C^1 convex function f. The corresponding primal-space updates are given by $X_t = \nabla\psi^*(Z_t)$. Our assumptions on ψ then guarantee that $X_t \in \mathsf{X}$ for all t.

Remark 2. It should be pointed out that (8) is *not* to be interpreted as a limiting case of the discrete-time stochastic approximation scheme

$$X_k = \nabla\psi^*(Z_k), \quad Z_{k+1} = Z_k + \gamma[h(Z_k) + \sigma\xi_k],$$

where $\{\xi_k\} \overset{\text{i.i.d.}}{\sim} N(0, I_n)$. This corresponds to the setting in which the gradient information at each time step is corrupted by an independent white Gaussian disturbance. By contrast, the "correct" discretization of (8) is given by

$$X_k = \nabla\psi^*(Z_k), \quad Z_{k+1} = Z_k + \gamma h(Z_k) + \sqrt{\gamma}\sigma\xi_k.$$

The difference is due to the fact that (8) describes the situation in which white Gaussian background noise is superimposed upon the noiseless MD dynamics (5). ◇

Let us analyze the convergence properties of (8). To that end, we have the following:

Proposition 3. *Consider the same setting as in Proposition 1 and assume, moreover, that ψ^* is C^2 and that $\|\Delta\psi^*\|_\infty < \infty$, where $\Delta = \nabla\cdot\nabla$ denotes the Laplace operator acting on $C^2(\mathbb{R}^n)$. Then, for any deterministic initial condition $Z_0 = z_0$ and any $T > 0$ we have*

$$\begin{aligned} V_T(z^*) &\le \frac{D^2_{\psi,\mathsf{X}}}{2} + \int_0^T \left[\min_{\mathsf{X}} f - f(X_t)\right] \mathrm{d}t \\ &\quad + \frac{\sigma^2 T}{2}\|\Delta\psi^*\|_\infty + \sigma\int_0^T \|X_t - x^*\|_2 \, \mathrm{d}W_t. \qquad (9) \end{aligned}$$

Remark 3. The uniform boundedness of $\Delta\psi^*$ is not a very restrictive requirement. Indeed, from the fact that ψ is strongly convex with constant α, it can be shown that the gradient of ψ^* is Lipschitz-continuous: $\|\nabla\psi^*(z) - \nabla\psi^*(z')\| \le \alpha^{-1}\|z - z'\|_*$ for any $z, z' \in \mathbb{R}^n$ [4, Theorem 4.2.1]. Hence, if ψ^* is C^2, this means that the operator norm of the Hessian $\nabla^2\psi^*$ is uniformly bounded. Since $\Delta\psi^*(z) = \mathrm{Tr}\,\nabla^2\psi^*(z)$, the uniform boundedness of $\Delta\psi^*$ follows. ◇

Proof. By definition,

$$\begin{aligned} V_t(z^*) &= \psi^*(Z_t) - \psi^*(z^*) - \langle\nabla\psi^*(z^*), Z_t - z^*\rangle \\ &= \psi^*(Z_t) - \psi^*(z^*) - \langle x^*, Z_t - z^*\rangle. \end{aligned}$$

Applying Itô's formula to the function $Z_t \mapsto \psi^*(Z_t)$, we get

$$\begin{aligned} \mathrm{d}V_t(z^*) &= \left[\langle X_t - x^*, h(Z_t)\rangle + \frac{\sigma^2}{2}\Delta\psi^*(Z_t)\right]\mathrm{d}t \\ &\quad + \sigma\|X_t - x\|_2 \, \mathrm{d}W_t, \end{aligned}$$

where we have used the fact that, for any $v \in \mathbb{R}^n$, $\langle v, B_t\rangle = \|v\|_2 W_t$ in law. Integrating, we obtain

$$\begin{aligned} V_T(z^*) &= V_0(z^*) + \int_0^T \langle X_t - x^*, h(Z_t)\rangle \, \mathrm{d}t \\ &\quad + \frac{\sigma^2}{2}\int_0^T \Delta\psi^*(Z_t)\,\mathrm{d}t + \sigma\int_0^T \|X_t - x^*\|_2 \, \mathrm{d}W_t. \end{aligned} \qquad (10)$$

By convexity, $\min_{\mathsf{X}} f = f(x^*) \ge f(X_t) + \langle\nabla f(X_t), x^* - X_t\rangle = \langle h(Z_t), X_t - x^*\rangle$. Using this and the fact that $V_0(z^*) \le D^2_{\psi,\mathsf{X}}/2$ in (10), we get (9). □

The bound (9) translates into the following estimates:

Proposition 4. *Under the same assumptions as above,*

$$\begin{aligned} \mathbb{E}\left\{\inf_{0\le t\le T} f(X_t) - \min_{\mathsf{X}} f\right\} &\le \frac{D^2_{\psi,\mathsf{X}}}{2T} + \frac{\sigma^2}{2}\|\Delta\psi^*\|_\infty \\ \mathbb{E}\left\{f\left(\frac{1}{T}\int_0^T X_t \, \mathrm{d}t\right) - \min_{\mathsf{X}} f\right\} &\le \frac{D^2_{\psi,\mathsf{X}}}{2T} + \frac{\sigma^2}{2}\|\Delta\psi^*\|_\infty. \end{aligned}$$

Proof. We proceed in the same way as in the proof of Proposition 2, except that now we also use the fact that the process $\{\|X_t - x^*\|_2\}_{t\ge 0}$ is adapted to the filtration $\{\sigma(X_t : 0 \le s \le t)\}_{t\ge 0}$, so the expectation of the Itô integral in (9) is zero. □

IV. Distributed stochastic MD

Compared to the noiseless set-up, one unpleasant feature of the noisy MD dynamics (8) is that the value of the objective f either at the "best" point in $\{X_t\}_{t=0}^T$ or at the time average $(1/T)\int_0^T X_t \, \mathrm{d}t$ does not converge to the minimum value of f as $T \to \infty$; instead, the gap to optimalitly is bounded from above by a quantity proportional to the noise variance σ^2. Hence, it is of interest to devise a way to reduce the noise level.

One such way is to introduce redundancy and coupling. Specifically, consider a network consisting of N agents (processors) implementing the MD dynamics (8) with the same objective function f, but with possibly different initial conditions at $t = 0$. Since the dynamics in (8) are generally nonlinear, we cannot expect to reduce the effect of the noise simply by averaging the agents' trajectories. However, if the agents are also coupled to one another, and if the coupling is sufficiently strong, then they will attempt to converge to consensus. In this way, the noise variance in the "mean field" (i.e., the average of the agents' trajectories) will be reduced by a factor of N. Moreover, because the agents are converging to a consensus, we do not need to explicitly measure and average their trajectories in order to read off the mean field — we can instead simply track any given agent, thus achieving a fully decentralized design. (Network architectures of this sort, implementing noisy gradient flows with nonlinear saturation effects, can be used to model the effects of synchronization and redundancy on the learning and decision-making processes in the brain [14].)

We now set up our model. For $i = 1, \ldots, N$, let $\{Z^i_t\}_{t\ge 0}$ denote the trajectory of the ith agent, described by the SDE

$$\mathrm{d}Z^i_t = \left[h(Z^i_t) + \sum_{j=1}^n A_{ij}(Z^j_t - Z^i_t)\right]\mathrm{d}t + \sigma\,\mathrm{d}B^i_t.$$

Here, $\{B^1_t\}, \ldots, \{B^N_t\}$ are N independent copies of $\{B_t\}$, and $A = \{A_{ij}\}_{i,j=1}^N$ is an $N \times N$ symmetric matrix of nonnegative coupling weights. If we consider a weighted undirected graph $G = (V, E, A)$ with $V = \{1, \ldots, N\}$ and $E = \{(i,j) : A_{ij} > 0\}$, and introduce the (unnormalized) *graph Laplacian* $\mathscr{L} \triangleq \mathrm{diag}(A\mathbf{1}_N) - A$ of G [12], then we can rewrite the network dynamics more compactly as

$$\mathrm{d}\boldsymbol{Z}_t = [h(\boldsymbol{Z}_t) - (\mathscr{L} \otimes I_n)\boldsymbol{Z}_t]\,\mathrm{d}t + \sigma\,\mathrm{d}\boldsymbol{B}_t,$$

where we have defined

$$\begin{aligned}\boldsymbol{Z}_t &= \big((Z_t^1)^T,\ldots,(Z_t^N)^T\big)^T\\ h(\boldsymbol{Z}_t) &= \big(h^T(Z_t^1),\ldots,h^T(Z_t^N)\big)^T\\ \boldsymbol{B}_t &= \big((B_t^1)^T,\ldots,(B_t^N)^T\big)^T\end{aligned}$$

and $\otimes$ denotes the Kronecker product of matrices. We assume throughout that the network graph G is connected. This implies [12] that the Laplacian $\mathscr{L}$ has the eigenvalues

$$\lambda_1(\mathscr{L}) = 0 < \lambda_1(\mathscr{L}) \le \lambda_2(\mathscr{L}) \le \ldots \le \lambda_N(\mathscr{L}).$$

We will denote by $\underline{\lambda}$ the smallest nonzero eigenvalue of $\mathscr{L}$.

Given $z^* = \nabla\psi^*(x^*)$ with any $x^* \in \arg\min_{\mathsf{X}} f$, we are interested in the evolution of $V_t^i(z^*) \triangleq D_{\psi^*}(Z_t^i, z^*)$ for each $i \in \{1,\ldots,N\}$. To facilitate the analysis, we follow the approach of [14] and decompose the trajectory $\{\boldsymbol{Z}_t\}$ into the mean-field and the fluctuation components. Specifically, let us define the *consensus subspace* $\mathcal{C} = \{\boldsymbol{1}_N \otimes z : z \in \mathbb{R}^n\}$ of $\mathbb{R}^N \otimes \mathbb{R}^n$, and let $\boldsymbol{P} \triangleq \frac{1}{N}(\boldsymbol{1}_N\boldsymbol{1}_N^T \otimes I_n)$ be the linear projection operator onto $\mathcal{C}$. We define the mean field $\overline{Z}_t \in \mathbb{R}^n$ and the fluctuation $\widetilde{\boldsymbol{Z}}_t \in \mathbb{R}^N \otimes \mathbb{R}^n$ via $\boldsymbol{1}_N \otimes \overline{Z}_t = \boldsymbol{P}\boldsymbol{Z}_t$ and $\widetilde{\boldsymbol{Z}}_t = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Z}_t$. (Here, $\boldsymbol{I} = I_N \otimes I_n$ denotes the identity matrix on $\mathbb{R}^N \otimes \mathbb{R}^n$.) Then $\boldsymbol{Z}_t = \widetilde{\boldsymbol{Z}}_t + \boldsymbol{1}_N \otimes \overline{Z}_t$, and

$$\overline{Z}_t = \frac{1}{N}\sum_{t=1}^N Z_t^i,\ \widetilde{Z}_t^i = Z_t^i - \overline{Z}_t,\ \frac{1}{N}\sum_{i=1}^N \widetilde{Z}_t^i = 0, \qquad (11)$$

where $\widetilde{\boldsymbol{Z}}_t = \big((\widetilde{Z}_t^1)^T,\ldots,(\widetilde{Z}_t^N)^T\big)^T$. We start with the following basic lemma:

Lemma 1. *Define $\overline{V}_t(z^*) \triangleq D_{\psi^*}(\overline{Z}_t, z^*)$. If ψ is a DGF w.r.t. a norm $\|\cdot\|$ with modulus $\alpha > 0$, then*

$$V_t^i(z^*) \le \overline{V}_t(z^*) + \frac{1}{2\alpha}\|\widetilde{Z}_t^i\|_*^2 + \langle \nabla\psi^*(\overline{Z}_t) - \nabla\psi^*(z^*), \widetilde{Z}_t^i\rangle \qquad (12)$$

and

$$\overline{V}_t(z^*) \le \frac{1}{N}\sum_{i=1}^N V_t^i(z^*) \le \overline{V}_t(z^*) + \frac{1}{2\alpha N}\sum_{i=1}^N \|\widetilde{Z}_t^i\|_*^2. \qquad (13)$$

Proof. From definitions, we have

$$V_t^i(z^*) = \psi^*(Z_t^i) - \psi^*(z^*) - \langle\nabla\psi^*(z^*), Z_t^i - z\rangle. \qquad (14)$$

Since ψ is, by hypothesis, strongly convex with parameter $\alpha > 0$, the gradient $\nabla\psi^*$ is Lipschitz with constant $1/\alpha$ [4, Theorem 4.2.1] (cf. also Remark 3). This, in turn, implies (cf. Lemma 1.2.3 in [3]), for all $z, \eta \in \mathbb{R}^n$,

$$\psi^*(z+\eta) \le \psi^*(z) + \langle\nabla\psi^*(z), \eta\rangle + \frac{1}{2\alpha}\|\eta\|_*^2. \qquad (15)$$

Because $Z_t^i = \widetilde{Z}_t^i + \overline{Z}_t$, we can use (15) to write

$$\psi^*(Z_t^i) \le \psi^*(\overline{Z}_t) + \langle\nabla\psi^*(\overline{Z}_t), \widetilde{Z}_t^i\rangle + \frac{1}{2\alpha}\|\widetilde{Z}_t^i\|_*^2$$

Substituting this into (14) and simplifying, we obtain

$$\begin{aligned}V_t^i(z^*) \le{}& \underbrace{\psi^*(\overline{Z}_t) - \psi^*(z^*) - \langle\nabla\psi^*(z^*), \overline{Z}_t - z^*\rangle}_{=\overline{V}_t(z^*)}\\ &+ \frac{1}{2\alpha}\|\widetilde{Z}_t^i\|_*^2 + \langle\nabla\psi^*(\overline{Z}_t) - \nabla\psi^*(z^*), \widetilde{Z}_t^i\rangle,\end{aligned}$$

which is (12). Summing over i and dividing by N gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N V_t^i(z^*) &\le \overline{V}_t(z^*) + \frac{1}{2\alpha N}\sum_{i=1}^N \|\widetilde{Z}_t^i\|_*^2\\ &\quad + \left\langle \nabla\psi^*(\overline{Z}_t) - \nabla\psi^*(z), \frac{1}{n}\sum_{i=1}^n \widetilde{Z}_t^i \right\rangle\\ &= \overline{V}_t(z^*) + \frac{1}{2\alpha N}\sum_{i=1}^N \|\widetilde{Z}_t^i\|_*^2,\end{aligned}$$

where in the last step we have used (11). This gives the second inequality in (13). The first inequality follows from Jensen's inequality and the convexity of ψ^*. □

To apply the lemma, we need to track the evolution of the mean field $\overline{Z}_t$ and the fluctuations $\widetilde{\boldsymbol{Z}}_t$. For the former,

$$\mathrm{d}\overline{Z}_t = \overline{h}(\boldsymbol{Z}_t)\,\mathrm{d}t + \sigma\,\mathrm{d}\overline{B}_t,$$

where $\overline{h}(\boldsymbol{Z}_t) \triangleq \boldsymbol{1}_N^T\boldsymbol{P}h(\boldsymbol{Z}_t)$, $\overline{B}_t \triangleq \boldsymbol{1}_N^T\boldsymbol{P}\boldsymbol{B}_t$, and we have used the fact that $\boldsymbol{1}_N^T\mathscr{L} = 0$ (as an aside, $\mathrm{d}\overline{B}_t$ is equal to $(1/\sqrt{N})\,\mathrm{d}B_t$ in law). For the latter,

$$\mathrm{d}\widetilde{Z}_t^i = \big[h(Z_t^i) - \overline{h}(\boldsymbol{Z}_t) - (\mathscr{L}_i \otimes I_n)\boldsymbol{Z}_t\big]\,\mathrm{d}t + \sigma\,\mathrm{d}\widetilde{B}_t^i,$$

where $\mathscr{L}_i$ denotes the ith row of $\mathscr{L}$, and $\mathrm{d}\widetilde{B}_t^i \triangleq \mathrm{d}B_t^i - \mathrm{d}\overline{B}_t$ (which is equal to $\sqrt{(N-1)/N}\,\mathrm{d}B_t$ in law). Bounding the mean-field term $\overline{V}_t(z^*)$ is relatively easy:

Lemma 2. *Under the same assumptions as in Proposition 3, for any $T > 0$ we have the bound*

$$\begin{aligned}\overline{V}_T(z^*) \le{}& \frac{D_{\psi,\mathsf{X}}^2}{2} + \frac{\sigma^2 T}{2N}\|\Delta\psi^*\|_\infty\\ &+ \int_0^T \langle\nabla\psi^*(\overline{Z}_t) - x^*, \overline{h}(\boldsymbol{Z}_t)\rangle\,\mathrm{d}t\\ &+ \frac{\sigma}{\sqrt{N}}\int_0^T \|\nabla\psi^*(\overline{Z}_t) - x^*\|_2\,\mathrm{d}W_t. \qquad (16)\end{aligned}$$

Proof. We follow the same steps as in the proof of Proposition 3. Specifically, Itô's formula gives

$$\begin{aligned}\mathrm{d}\overline{V}_t(z^*) ={}& \left[\langle\nabla\psi^*(\overline{Z}_t) - x^*, \overline{h}(\boldsymbol{Z}_t)\rangle + \frac{\sigma^2}{2N}\Delta\psi^*(\overline{Z}_t)\right]\mathrm{d}t\\ &+ \frac{\sigma}{\sqrt{N}}\|\nabla\psi^*(\overline{Z}_t) - x^*\|_2\,\mathrm{d}W_t.\end{aligned}$$

Integrating and upper-bounding the terms involving $\overline{V}_0(z^*)$ and $\Delta\psi^*$ as before, we get (16). □

By contrast, bounding the squared dual norm $\|\widetilde{Z}_t^i\|_*^2$ for each $i \in \{1,\ldots,N\}$ is tricky: in general, the function $z \mapsto \|z\|_*^2$ is not differentiable (let alone C^2), so we cannot apply Itô's formula directly. Instead, to prove the lemma below we use the fact that $\|z\|_*^2 = \sup\{\langle v, z\rangle^2 : v \in B_{\|\cdot\|}\}$. For each

$v \in \mathbb{R}^n$, the function $v \mapsto g_v(z) \triangleq \langle v, z\rangle^2$ is C^2. Thus, we can apply Itô's lemma to each v separately, and then take the supremum over $B_{\|\cdot\|}$.

Lemma 3. *For any $T > 0$*

$$
\begin{aligned}
&\|\widetilde{Z}^i_T\|^2_* \le e^{-2\underline{\lambda} T}\|\widetilde{Z}^i_0\|^2_* + \frac{\sigma^2 D^2_{\|\cdot\|}}{8\underline{\lambda}} \\
&+ 2 \sup_{v \in B_{\|\cdot\|}} \int_0^T e^{2\underline{\lambda}(t-T)} \langle v, \widetilde{Z}^i_t\rangle \langle v, h(Z^i_t) - \overline{h}(\boldsymbol{Z}_t)\rangle \, \mathrm{d}t \\
&+ \sigma D_{\|\cdot\|}\sqrt{\frac{N-1}{N}} \sup_{v \in B_{\|\cdot\|}} \int_0^T e^{2\underline{\lambda}(t-T)} \langle v, \widetilde{Z}^i_t\rangle \, \mathrm{d}W_t, \qquad (17)
\end{aligned}
$$

where $D_{\|\cdot\|} = 2\sup\{\|v\|_2 : v \in B_{\|\cdot\|}\}$ is the Euclidean (ℓ_2) diameter of $B_{\|\cdot\|}$.

Remark 4. Since all norms on $\mathbb{R}^n$ are equivalent, we could have simply bounded the squared ℓ^2 norm $\|\widetilde{Z}^i_t\|^2_2$ and then used the fact that (due to norm equivalence) $\|\widetilde{Z}^i_t\|^2_* \le K\|\widetilde{Z}^i_t\|^2_2$ for some $K > 0$ that depends on $\|\cdot\|$. However, the constant K may actually grow with n — for instance, if $\|\cdot\| = \|\cdot\|_\infty$, then $\|\cdot\|_* = \|\cdot\|_1$ and $K = n$. ◇

Proof. For a given $v \in \mathbb{R}^n$, let $Y^i_{v,t} = g_v(\widetilde{Z}^i_t) \equiv \langle v, \widetilde{Z}^i_t\rangle^2$. Then Itô's formula gives

$$
\begin{aligned}
\mathrm{d}Y^i_{v,t} &= 2\langle v, \widetilde{Z}^i_t\rangle \, \mathrm{d}\langle v, \widetilde{Z}^i_t\rangle + \frac{\sigma^2(N-1)}{N}\|v\|^2_2 \, \mathrm{d}t \\
&= \left[2\langle v, \widetilde{Z}^i_t\rangle\langle v, h(Z^i_t) - \overline{h}(\boldsymbol{Z}_t)\rangle + \frac{\sigma^2(N-1)}{N}\|v\|^2_2\right] \mathrm{d}t \\
&\qquad - 2\langle v, \widetilde{Z}^i_t\rangle\langle v, (\mathscr{L}_i \otimes I_n)\boldsymbol{Z}_t\rangle \, \mathrm{d}t \\
&\qquad + 2\sigma\sqrt{\frac{N-1}{N}}\langle v, \widetilde{Z}^i_t\rangle \|v\|_2 \, \mathrm{d}W_t
\end{aligned}
$$

Let us add $2\underline{\lambda} Y^i_{v,t}\,\mathrm{d}t = 2\underline{\lambda}\langle v, \widetilde{Z}^i_t\rangle^2\,\mathrm{d}t$ to both sides of this equation. We thus obtain

$$
\begin{aligned}
&\mathrm{d}Y^i_{v,t} + 2\underline{\lambda} Y^i_{v,t}\,\mathrm{d}t \\
&= \left[2\langle v, \widetilde{Z}^i_t\rangle\langle v, h(Z^i_t) - \overline{h}(\boldsymbol{Z}_t)\rangle + \frac{\sigma^2(N-1)}{N}\|v\|^2_2\right] \mathrm{d}t \\
&\qquad 2\left[\underline{\lambda}\langle v, \widetilde{Z}^i_t\rangle^2 - \langle v, \widetilde{Z}^i_t\rangle\langle v, (\mathscr{L}_i \otimes I_n)\boldsymbol{Z}_t\rangle\right] \mathrm{d}t \\
&\qquad + 2\sigma\sqrt{\frac{N-1}{N}}\langle v, \widetilde{Z}^i_t\rangle\|v\|_2 \, \mathrm{d}W_t \\
&= e^{-2\underline{\lambda} t}\, \mathrm{d}\left(Y^i_{v,t} e^{2\underline{\lambda} t}\right),
\end{aligned}
$$

where the last step follows from the fact that $e^{2\underline{\lambda} t}\,\mathrm{d}Y^i_{v,t} + 2\underline{\lambda} e^{2\underline{\lambda} t} Y^i_{v,t}\,\mathrm{d}t$ is the total Itô derivative of $Y^i_{v,t}e^{2\underline{\lambda} t}$. Integrating then gives

$$
\begin{aligned}
&Y^i_{v,T} = e^{-2\underline{\lambda} T} Y^i_{v,0} + \frac{\sigma^2(N-1)}{2N\underline{\lambda}}(1 - e^{-2\underline{\lambda} T})\|v\|^2_2 \\
&+ 2\int_0^T e^{2\underline{\lambda}(t-T)}\langle v, \widetilde{Z}^i_t\rangle\langle v, h(Z^i_t) - \overline{h}(\boldsymbol{Z}_t)\rangle \, \mathrm{d}t \\
&+ 2\int_0^T e^{2\underline{\lambda}(t-T)}\left[\underline{\lambda}\langle v, \widetilde{Z}^i_t\rangle^2 - \langle v, \widetilde{Z}^i_t\rangle\langle v, (\mathscr{L}_i \otimes I_n)\boldsymbol{Z}_t\rangle\right] \mathrm{d}t \\
&+ 2\sigma\sqrt{\frac{N-1}{N}}\|v\|_2 \int_0^T e^{2\underline{\lambda}(t-T)}\langle v, \widetilde{Z}^i_t\rangle \, \mathrm{d}W_t \qquad (18)
\end{aligned}
$$

Now let $w = (e_i e_i^T \otimes I_n)(\boldsymbol{1}_N \otimes v)$, where e_i is the ith canonical basis vector in $\mathbb{R}^N$. Then

$$
\begin{aligned}
\langle v, \widetilde{Z}^i_t\rangle\langle v, (\mathscr{L}_i \otimes I_n)\boldsymbol{Z}_t\rangle &= \langle w, \widetilde{\boldsymbol{Z}}_t\rangle\langle(\mathscr{L} \otimes I_n)\boldsymbol{Z}_t, w\rangle \\
&= \langle w, \widetilde{\boldsymbol{Z}}_t\rangle\langle(\mathscr{L} \otimes I_n)\widetilde{\boldsymbol{Z}}_t, w\rangle \\
&\ge \underline{\lambda}\langle w, \widetilde{\boldsymbol{Z}}_t\rangle^2 \\
&= \underline{\lambda}\langle v, \widetilde{Z}^i_t\rangle^2,
\end{aligned}
$$

where the second step uses the fact that $\mathscr{L}\boldsymbol{1}_N = 0$, and that $\widetilde{\boldsymbol{Z}}_t$ is orthogonal to any element of the consensus subspace $\mathcal{C}$. The inequality that follows is due to the fact that $\underline{\lambda}$ is the smallest eigenvalue of $\mathscr{L} \otimes I_n$ on the orthogonal complement of $\mathcal{C}$. Therefore, the third term in (18) is nonpositive. Using this fact, together with the definition of $\|\cdot\|_*$, we get (17). □

In analogy to classical results on the linear agreement dynamics [12, Ch. 3], Lemma 3 shows the impact of the network topology on the speed with which the agents reach consensus, as well as on the size of the squared dual norm of the fluctuations $\widetilde{Z}^i_t$, $i = 1, \dots, N$. In particular, the density of the network connections affects the value of the smallest nonzero eigenvalue (also known as the Fiedler eigenvalue) of $\mathscr{L}$. For example, if the nonzero entries of A are all equal to some constant $\kappa > 0$, then the value of $\underline{\lambda}$ can be as large as $N\kappa$ (when G is the complete graph) or as small as κ (when G is the star graph). We point out that it is possible to use quorum sensing [15] to implement an effective all-to-all connectivity using only a linear number of connections.

V. Case study: composite objectives

In order to apply Lemmas 2 and 3, further assumptions on the structure of the problem are needed. In this section, we will analyze a certain type of optimization problems, for which a particularly simple characterization is available. Specifically, we assume that the objective f is of the form

$$
f(x) = \langle a, x\rangle + b + \psi(x) \qquad (19)
$$

where $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ are given, and where ψ is a steep DGF on X with w.r.t. a norm $\|\cdot\|$ with modulus $\alpha > 0$. Objective functions of this form can arise in many scenarios. For example, in the context of statistical estimation, given a large number m of independent samples $Y_1, \dots, Y_m$ from an unknown probability distribution on $\mathbb{R}^k$, we may wish to fit an *exponential family model* [16] to this set of data. Given a σ-finite base measure μ on $\mathbb{R}^k$, an exponential family of distributions is given by densities of the form

$$
p_x(y) = q(y)\exp\Big\{\langle x, \mathsf{T}(y)\rangle - \psi(x)\Big\}, \qquad x \in \mathsf{X}
$$

where q is some reference probability density w.r.t. μ, $\mathsf{T} : \mathbb{R}^k \to \mathbb{R}^n$ is a Borel mapping known as the *sufficient statistic*, the set

$$
\mathsf{X} \triangleq \left\{x \in \mathbb{R}^n : \int_{\mathbb{R}^k} q(y) e^{\langle x, \mathsf{T}(y)\rangle}\mu(\mathrm{d}y) < +\infty\right\}
$$

is called the *natural parameter space*, and the function $\psi(x) = \log\int_{\mathbb{R}^k} q(y)e^{\langle x, \mathsf{T}(y)\rangle}\mu(\mathrm{d}y), x \in \mathsf{X}$ is referred to as the *log-partition function*. The parameter-fitting problem is then to minimize the empirical negative log-likelihood

$$
\begin{aligned}
f(x) &\triangleq -\frac{1}{m}\sum_{j=1}^{m} \log p_x(Y_j) \\
&= -\frac{1}{m}\left[\left\langle x, \sum_{j=1}^{m} \mathsf{T}(Y_j)\right\rangle + \sum_{j=1}^{m} \log q(Y_j)\right] + \psi(x) \qquad (20)
\end{aligned}
$$

over the natural parameter space X. From the theory of exponential families, we know that the natural parameter space X is closed and convex, and the log-partition function ψ is C^∞ and steep on X [16]. If, in addition, X is compact and ψ is strongly convex w.r.t. some norm $\|\cdot\|$, then the objective function in (20) is precisely of the form (19).

The very structure of the problem suggests that we should use the MD method with the DGF ψ. In this case, the basic Itô SDE describing the MD process is particularly simple. Indeed, because ψ is steep, the functions $\nabla\psi$ and $\nabla\psi^*$ are inverses of each other. From this, we get $h(z) = -\nabla f(\nabla\psi^*(z)) = -[a + \nabla\psi(\nabla\psi^*(z))] = -(a+z)$, so we can write (8) as

$$
\mathrm{d}Z_t = -(a + Z_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t.
$$

This is a Langevin equation whose solution is given by the n-dimensional Ornstein–Uhlenbeck process

$$
Z_T = e^{-T} Z_0 - a(1 - e^{-T}) + \sigma \int_0^T e^{t-T}\,\mathrm{d}B_t
$$

Turning to the distributed N-agent set-up, we write down the SDEs for the mean field $\overline{Z}_t$ and the fluctuations $\widetilde{\boldsymbol{Z}}_t$:

$$
\mathrm{d}\overline{Z}_t = -(a + \overline{Z}_t)\,\mathrm{d}t + \frac{\sigma}{\sqrt{N}}\,\mathrm{d}B_t \qquad (21)
$$

$$
\mathrm{d}\widetilde{\boldsymbol{Z}}_t = -\left[(I_N + \mathscr{L}) \otimes I_n\right] \widetilde{\boldsymbol{Z}}_t + \sigma\sqrt{\frac{N-1}{N}}\,\mathrm{d}\boldsymbol{B}_t. \qquad (22)
$$

We can now analyze the behavior of our distributed MD scheme:

Theorem 1. *Let $\overline{X}_t = \nabla\psi^*(\overline{Z}_t)$. Then for any $T > 0$*

$$
\begin{aligned}
\overline{V}_T(z^*) \le \frac{D^2_{\psi,\mathsf{X}}}{2} &+ \frac{\sigma^2 T}{2N}\|\Delta\psi^*\|_\infty \\
&+ \int_0^T \left[f(x^*) - f(\overline{X}_t)\right]\mathrm{d}t \\
&+ \frac{\sigma}{\sqrt{N}}\int_0^T \|\overline{X}_t - x^*\|_2\,\mathrm{d}B_t \qquad (23)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\overline{V}_T(z^*) \le \frac{D^2_{\psi,\mathsf{X}}}{2} &+ \frac{\sigma^2 T}{2N}\|\Delta\psi^*\|_\infty \\
&+ \mathbb{E}\left\{\int_0^T \left[f(x^*) - f(\overline{X}_t)\right]\mathrm{d}t\right\} \qquad (24)
\end{aligned}
$$

Proof. Since $h(z) = -(a+z)$, $\overline{h}(\boldsymbol{Z}_t) = h(\overline{Z}_t)$. Hence, for the second term on the right-hand side of (16) we have

$$
\begin{aligned}
&\langle \nabla\psi^*(\overline{Z}_t) - x^*, \overline{h}(\boldsymbol{Z}_t)\rangle \\
&\quad = \langle \nabla\psi^*(\overline{Z}_t) - x^*, h(\overline{Z}_t)\rangle \\
&\quad = -\langle \nabla\psi^*(\overline{Z}_t) - x^*, \nabla f(\nabla\psi^*(\overline{Z}_t))\rangle \\
&\quad = -\langle \overline{X}_t - x^*, \nabla f(\overline{X}_t)\rangle \\
&\quad \le f(x^*) - f(\overline{X}_t),
\end{aligned}
$$

where the last step is by convexity of f. Substituting this into (16), we get (23). Taking expectations of both sides and using the fact that the Itô term has mean zero, we get (24). □

In order to state our next result pertaining to the squared dual norm $\|\widetilde{Z}^i_t\|^2_*$, we need to introduce the so-called γ_2 *functional* [17]. Let (S, ρ) be a metric space. An *admissible sequence* for S is a collection $\{S_k\}_{k\ge 0}$ of finite subsets of S, such that $|S_0| = 1$ and, for every $k \ge 1$, $|S_k| = 2^{2^k}$. Then the γ_2 functional of (S,ρ) is defined by

$$
\gamma_2(S,\rho) \triangleq \inf \sup_{s\in S} \sum_{k=0}^{\infty} 2^{k/2}\rho(s, S_k),
$$

where the infimum is over all admissible sequences, and, for each k, $\rho(s, S_k) \triangleq \inf_{s'\in S_k}\rho(s,s')$ is the minimum distance from $s \in S$ to S_k. In particular, we will be interested in the γ_2 functionals of norm balls in $\mathbb{R}^n$. Thus, for a norm $\|\cdot\|$, we will use the shorthand $\gamma_2(B_{\|\cdot\|})$ for $\gamma_2(S,\rho)$ with $S = B_{\|\cdot\|}$ and the metric ρ induced by $\|\cdot\|$. Exact expressions for the γ_2 functional of general convex bodies are not easy to obtain, but we point out that this quantity is closely related to entropy numbers. We can now state the following:

Theorem 2. *For any $T > 0$,*

$$
\begin{aligned}
\|\widetilde{Z}^i_T\|^2_* \le e^{-2\underline{\lambda}T}\|\widetilde{Z}^i_0\|^2_* &+ \frac{\sigma^2 D^2_{\|\cdot\|}}{8\underline{\lambda}} \\
+ \sigma D_{\|\cdot\|}\sqrt{\frac{N-1}{N}} &\sup_{v\in B_{\|\cdot\|}} \int_0^T e^{2\underline{\lambda}(t-T)}\langle v, \widetilde{Z}^i_t\rangle\,\mathrm{d}W_t. \qquad (25)
\end{aligned}
$$

Moreover, there exists a positive constant $M > 0$ that depends only on the choice of $\|\cdot\|$, such that for any deterministic initial condition $Z_0 = z_0$,

$$
\mathbb{E}\|\widetilde{Z}^i_T\|^2_* \le e^{-2\underline{\lambda}T}\|\widetilde{z}^i_0\|^2_* + \frac{\sigma^2 D^2_{\|\cdot\|}}{8\underline{\lambda}} + M D_{\|\cdot\|}\sigma^2\sqrt{\frac{n}{\underline{\lambda}}}\gamma_2(B_{\|\cdot\|}) \qquad (26)
$$

Proof. Since $h(Z^i_t) - \overline{h}(\boldsymbol{Z}_t) = h(Z^i_t) - h(\overline{Z}_t) = \overline{Z}_t - Z^i_t = -\widetilde{Z}^i_t$, the second term on the right-hand side of (17) is nonpositive. This gives (25).

We now turn to (26). The challenge is to bound the expectation of the supremum of the Itô integral

$$
I_v \triangleq \int_0^T e^{2\underline{\lambda}(t-T)}\langle v, \widetilde{Z}^i_t\rangle\,\mathrm{d}W_t.
$$

in (25) over $v \in B_{\|\cdot\|}$. With a deterministic initial condition $Z_0 = z_0$, $\widetilde{\boldsymbol{Z}}_t$ is a Gaussian process adapted to the natural filtration $\{\sigma(\widetilde{\boldsymbol{Z}}_s : 0 \le s \le t)\}_{t\ge 0}$. Therefore, $\{I_v :$

$v \in B_{\|\cdot\|}\}$ is a zero-mean Gaussian process indexed by $B_{\|\cdot\|}$. Thus, we can utilize the well-known *generic chaining* technique [17] for bounding the expectations of suprema of Gaussian processes. To that end, define the metric

$$\rho(v,v') \triangleq \sqrt{\mathbb{E}|I_v - I_{v'}|^2}$$

on $B_{\|\cdot\|}$. Then (cf. Theorem 2.1 in [17])

$$\mathbb{E}\left\{\sup_{v \in B_{\|\cdot\|}} I_v\right\} \le C_0 \gamma_2(B_{\|\cdot\|}, \rho),$$

where $C_0 > 0$ is a universal constant. Now,

$$\begin{aligned}\rho(v,v') &= \sqrt{\mathbb{E}\left|\int_0^T e^{2\underline{\lambda}(t-T)} \langle v - v', \widetilde{Z}^i_t\rangle \, \mathrm{d}W_t\right|^2} \\ &= \sqrt{\mathbb{E}\int_0^T e^{4\underline{\lambda}(t-T)} \langle v - v', \widetilde{Z}^i_t\rangle^2 \, \mathrm{d}t} \\ &\le \|v - v'\| \sqrt{\mathbb{E}\int_0^T e^{4\underline{\lambda}(t-T)} \|\widetilde{Z}^i_t\|_*^2 \, \mathrm{d}t} \\ &\le \sqrt{K}\|v - v'\| \sqrt{\int_0^T e^{4\underline{\lambda}(t-T)} \mathbb{E}\|\widetilde{Z}^i_t\|_2^2 \, \mathrm{d}t} \qquad (27)\end{aligned}$$

where the second step is by the Itô isometry, the next step uses Hölder's inequality, and the last step uses equivalence of norms $\|\cdot\|_*$ and $\|\cdot\|_2$ (cf. Remark 4 for the definition of the constant K). Since $\{\widetilde{\boldsymbol{Z}}_t\}$ is an Ornstein–Uhlenbeck process, cf. (22), we can bound the square root term in (27) by $C_1\sqrt{\frac{\sigma^2 n(N-1)}{N\underline{\lambda}}} \le C_1\sqrt{\frac{\sigma^2 n}{\underline{\lambda}}}$, where the constant $C_1 > 0$ is independent of n, N, T, or σ. This implies that

$$\gamma_2(B_{\|\cdot\|}, \rho) \le C_1 \sigma \sqrt{\frac{Kn}{\underline{\lambda}}} \gamma_2(B_{\|\cdot\|}).$$

From this, we get (26) with $M = C_0 C_1 \sqrt{K}$. □

The above results can then be used to track the evolution of the time-averaged optimization error $(1/T)\int_0^T [f(X^i_t) - \min_{\mathsf{X}} f]\,\mathrm{d}t$ for any $i = 1, \dots, N$, where $X^i_t = \nabla\psi^*(Z^i_t) \in \mathsf{X}$. Indeed, from Theorem 1 we have

$$\frac{1}{T}\mathbb{E}\left\{\int_0^T \left[f(\overline{X}_t) - \min_{\mathsf{X}} f\right] \mathrm{d}t\right\} \le \frac{D^2_{\psi,\mathsf{X}}}{2T} + \frac{\sigma^2}{2N}\|\Delta\psi^*\|_\infty$$

If ψ is Lipschitz with constant L, then

$$\begin{aligned} f(X^i_t) &\le (\|a\|_* + L)\|X^i_t - \overline{X}_t\| + f(\overline{X}_t) \\ &= (\|a\|_* + L)\|\nabla\psi^*(Z^i_t) - \nabla\psi^*(\overline{Z}_t)\| + f(\overline{X}_t) \\ &\le \frac{\|a\|_* + L}{\alpha}\|\widetilde{Z}^i_t\|_* + f(\overline{X}_t), \end{aligned}$$

where the last step uses the Lipschitz property of $\nabla\psi^*$. Therefore,

$$\begin{aligned} &\frac{1}{T}\mathbb{E}\left\{\int_0^T \left[f(X^i_t) - \min_{\mathsf{X}} f\right] \mathrm{d}t\right\} \\ &\quad \le \frac{\|a\|_* + L}{\alpha T}\mathbb{E}\left\{\int_0^T \|\widetilde{Z}^i_t\|_* \, \mathrm{d}t\right\} + \frac{D^2_{\psi,\mathsf{X}}}{2T} + \frac{\sigma^2}{2N}\|\Delta\psi^*\|_\infty \\ &\quad \le \frac{\|a\|_* + L}{\alpha\sqrt{T}}\sqrt{\int_0^T \mathbb{E}\|\widetilde{Z}^i_t\|_*^2 \, \mathrm{d}t} + \frac{D^2_{\psi,\mathsf{X}}}{2T} + \frac{\sigma^2}{2N}\|\Delta\psi^*\|_\infty, \end{aligned}$$

where the last step follows after two uses of Jensen's inequality. We can now use Theorem 2 to bound the integral in the last line (details are omitted for lack of space).

VI. Conclusion

We have analyzed continuous-time stochastic MD methods, including a robust decentralized implementation. Our treatment can be extended in multiple ways, including time- and agent-dependent objectives. Another important direction for future work is to tighten the bounds on the squared dual norm of the fluctuation part of the trajectory.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambrdige Univ. Press, 2004.

[2] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[3] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

[4] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer, 2001.

[5] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley, 1983.

[6] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Res. Lett.*, vol. 31, pp. 167–175, 2003.

[7] A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 29, no. 4, pp. 1574–1609, 2009.

[8] A. Ben-Tal, T. Margalit, and A. Nemirovski, "The ordered subsets mirror descent optimization method with applications to tomography," *SIAM J. Optim.*, vol. 12, no. 1, pp. 79–108, 2001.

[9] N. Srebro, K. Sridharan, and A. Tewari, "On the universality of online mirror descent," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 2645–2653.

[10] M. Raginsky, A. Rakhlin, and S. Yüksel, "Online convex programming and regularization in adaptive control," in *Proc. 49th IEEE Conf. on Decision and Control*, Atlanta, GA, December 2010, pp. 1957–1962.

[11] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. Springer, 2003.

[12] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.

[13] L. M. Bregman, "The relaxation method for finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. and Math. Phys.*, vol. 7, pp. 200–217, 1967.

[14] J. Bouvrie and J.-J. Slotine, "Synchronization and redundancy: implications for robustness of neural learning and decision making," *Neural Comput.*, vol. 23, pp. 2915–2941, 2011.

[15] A. Taylor, M. R. Tinsley, F. Wang, Z. Huang, and K. Showalter, "Dynamical quorum sensing and synchronization in large populations of chemical oscillators," *Science*, vol. 323, no. 5914, pp. 614–617, 2009.

[16] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[17] M. Talagrand, *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer, 2005.