STOCHASTIC MIRROR DESCENT FOR CONVEX OPTIMIZATION WITH CONSENSUS CONSTRAINTS

A. BOROVYKH, N. KANTAS, P. PARPAS, G. A. PAVLIOTIS

Abstract. The mirror descent algorithm is known to be effective in applications where it is beneficial to adapt the mirror map to the underlying geometry of the optimization model. However, the effect of mirror maps on the geometry of distributed optimization problems has not been previously addressed. In this paper we propose and study exact distributed mirror descent algorithms in continuous-time under additive noise and present the settings that enable linear convergence rates. Our analysis draws motivation from the augmented Lagrangian and its relation to gradient tracking. To further explore the benefits of mirror maps in a distributed setting we present a preconditioned variant of our algorithm with an additional mirror map over the Lagrangian dual variables. This allows our method to adapt to the geometry of the consensus manifold and leads to faster convergence. We illustrate the performance of the algorithms in convex settings both with and without constraints. We also explore their performance numerically in a non-convex application with neural networks.

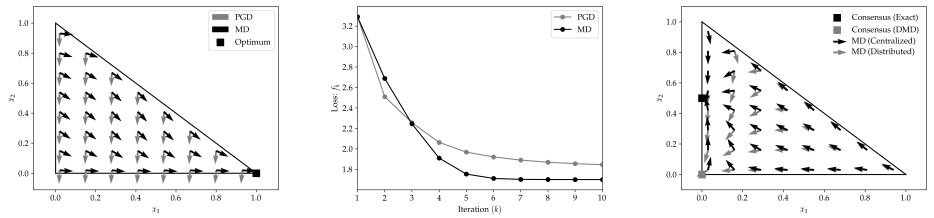
Key words. Distributed optimization, mirror descent, pre-conditioning, interacting particles, stochastic optimization.

AMS subject classifications. 65K10, 68Q87, 60G07, 68W15

1. Introduction.

The choice of mirror map has a significant impact on both the theoretical and numerical performance of the Mirror Descent (MD) algorithm [4, 9]. With an appropriate choice of the mirror map, MD captures the geometry of the optimization model more faithfully than other first-order methods. We illustrate this point in Figure 1a by plotting the vector fields generated by MD (using the negative entropy function as the mirror map) and Projected Gradient Descent (PGD) (with Euclidean projection) for a strongly convex quadratic optimization problem over the three-dimensional simplex. It is clear from Figure 1a that the PGD vector field points in the correct direction towards the unique minimum. But as soon as the PGD vector field hits the boundary, then the algorithm slows down considerably. The slowdown is due to the fact that the gradient always points towards the direction of steepest descent *for the objective function irrespective of the constraints*. When PGD hits the boundary, then the steepest descent direction is no longer appropriate for the problem's geometry. When MD hits the boundary of the feasible region, it glides across the boundary and towards the solution. This observation is reflected in the numerical performance of the two algorithms. In Figure 1b we indeed see that PGD initially makes good progress towards the solution but then stalls. MD, on the other hand, is slower in the first two iterations but converges to the optimal solution much faster. This phenomenon is not only present in problems with constraints but is also relevant in unconstrained problems, especially for ill-conditioned problems, and inverse optimization problems that have a sparsity inducing norm in the objective function. For example, in unconstrained ill-conditioned problems, the gradient descent method performs no preconditioning, whereas mirror descent uses the Hessian of the mirror map as a preconditioner (see Section 3.1).

There exists no theoretical or algorithmic framework to explain how to compute an optimal mirror map for a given problem. However, mirror maps for some particular classes of problems are well known (see Appendix A.1, [4], [9]). Mirror descent, and especially the effect of the choice of the mirror map for *distributed* optimization problems has received much less attention (see Section 1.1 for related work). Distributed optimization problems, even when otherwise unconstrained, have to satisfy a consensus



(a) Vector Fields of Centralized PGD/MD (b) PGD/MD iterations (c) Vector Fields of Distributed PGD/MD

Fig. 1: Vector fields for Projected Gradient Descent (PGD) and Mirror Descent (MD) for a quadratic function over the three dimensional simplex (we plot the two dimensional projection). MD uses the negative entropy function as the mirror map, and PGD performs the projection using the ℓ_2 norm.

constraint, and existing algorithms do not capture the geometry of the consensus manifold. Motivated by the attractive features of the mirror descent algorithm described above, we attempt to answer the question: *Does there exist a distributed variant of Mirror Descent that can accurately capture the geometry of distributed optimization models?* To answer this question, we study distributed algorithms for the following optimization model,

$$\min_{x_i \in \mathcal{X}} \sum_{i=1}^{N} f_i(x^i), \quad \text{s.t. } x^i = x^j \quad \forall (i,j) \in E. \tag{1.1}$$

The $\{x^i\}_{i=1}^N$ with indices $i = 1, ..., N$ denote the computational nodes or particles, as we refer to them in previous work [6]. These communicate through a strongly connected, weighted, undirected graph $\mathcal{G} := (V, E, A)$; where V represents the nodes of the graph, E its edges and A is the adjacency matrix. Each particle has access to its own objective function $f_i : \mathbb{R}^d \to \mathbb{R}$, and constraint set $\mathcal{X} \subset \mathbb{R}^d$.

For the purposes of motivating the results of our work consider the following natural generalization of Distributed Mirror Descent (DMD),

$$\partial_t \nabla \Phi(x_t^i) = -\nabla f_i(x_t^i) - \sum_{j=1}^{N} A_{ij}(x_t^i - x_t^j) \qquad i = 1, \dots, N, \tag{1.2}$$

where Φ is the mirror map, and A_{ij} is the weight of edge (i, j). We call a distributed algorithm *exact if it converges to a solution that it is both optimal and satisfies the consensus constraint.* In Figure 1c we plot the vector field generated by (1.2) on a quadratic optimization model with $N = 2$ over the three dimensional simplex, and the centralized MD algorithm. The centralized MD algorithm for (1.1) substitutes the constraints in the objective function and follows the dynamics below,

$$\partial_t \nabla \Phi(x_t) = -\sum_{j=1}^{N} \nabla f_i(x_t).$$

As expected, the two algorithms generate different vector fields. What is more concerning is that the Distributed Mirror Descent in (1.2) does not converge to

Reference	Mirror	Rate	Exact	Noise	Step-size
Liang et al. [21]	No	Linear	Yes	No	-
[22, 16, 23, 43]	No	N/A	Yes	No	-
Sun et al. [38]	Yes	N/A	Yes	No	-
This work	Yes	Linear	Yes	Yes	-
Shi et al. [36]	No	Linear	Yes	No	Constant
Qu & Li [30]	No	Linear	Yes	No	Constant
Jakovetic et al. [19]	No	Sub-Linear	No	No	Diminishing
Pu & Nedic [29]	No	Linear	Yes	Yes	Constant
Ram et. al. [32]	No	N/A	Yes/No	Yes	Diminish./Const.
Duchi et al. [13]	Yes	Sub-Linear	Yes/No	Yes	Diminish./Const.
Nedic et al. [27]	Yes	N/A	Yes	Yes	Diminishing
Shahrampour et al. [33]	Yes	Sub-Linear	Yes	Yes	Diminishing

Table 1: Overview of convergence rates for different types of algorithms. Exact refers to whether or not the algorithm achieves exact consensus, and mirror refers to whether or not the algorithm allows for mirror maps. Continuous time methods are marked as - in the step size entry.

the unique solution of the problem. This observation is not surprising given that Distributed Gradient Descent (DGD) (unless suitable modifications are made to the algorithm) also fails to converge to the exact solution of distributed optimization problems [42]. The second question we seek to address in this paper is: *How should the dynamics of distributed mirror descent be modified, so that convergence to the exact solution is guaranteed?* These guarantees are meant to hold for determistic dynamics, but in this paper we will also consider the more general case of stochastic dynamics with additive noise, where the noise is added to account for corrupted gradient information, data sub-sampling (as is the case in stochastic gradient descent) or errors due to the network, such as communication channels being corrupted.

1.1. Previous work. Distributed optimization has a variety of applications. Removing the existence of a central server and having the nodes communicate in a decentralized manner can remove both computational bottlenecks and privacy risks. A classic reference for distributed optimization is [5], and more recent applications in statistical learning are described in [8]. The authors in [10] also describe several interesting applications. A variant of distributed optimization known as federated learning [25] was proposed recently for solving optimization problems in which the data is stored across a very large number devices for privacy purposes.

The literature on distributed optimization algorithms is vast. Since this paper focuses on distributed first-order algorithms for convex optimization models, we will focus on this class of algorithms. Two algorithmic techniques can be used to develop exact distributed optimization algorithms. The first technique uses diminishing step-sizes, and the second one relies on gradient tracking. Gradient tracking is closely related to augmented Lagrangian methods (see Section 3 for more details). Algorithms

with diminishing step-sizes tend to be very slow in practice, so recent literature focuses on using constant step-sizes. The algorithm we propose in this paper, and its variants, are developed in continuous time. The works of (among others) [16, 22, 38, 23, 43] also analyze decentralized optimization schemes in continuous time. The works of [13, 27, 33, 32] focus on an analysis of the distributed mirror descent algorithm. In Table 1, we summarize selected related works that show how this paper fits within the existing literature. Current works on exact distributed algorithms, with a fixed step-size, are only based on gradient or sub-gradient descent. Table 1 also lists earlier proposed exact distributed algorithms in discrete time, which rely on the slower mechanism of diminishing step-sizes to achieve exact convergence. Moreover, the mirror maps in the existing literature are used only to model accurately the geometry of the separable constraints and not the consensus constraint that is the distinguishing feature of this paper. Finally, we note mirror descent dynamics are related to Riemannian descent as presented in e.g. [2, 11] and preconditioning [17, 1, 34].

1.2. Main results and contributions. Our results are based on a continuous-time analysis of stochastic mirror descent dynamics. Our contributions can be summarized as follows,

- In Section 4.1 we show that without strong assumptions on the minimizers of each f_i, the classic distributed stochastic mirror descent formulation with constant noise achieves exponential convergence to a neighborhood around a different point than the optimum and that the size of this neighborhood cannot be reduced using the mirror map or reducing noise.
- To address the inexactness of the conventional DMD algorithm we propose an exact variant called Exact Interacting Stochastic Mirror Descent (EISMD), that is able to converge exponentially fast to a much smaller neighborhood than the conventional distributed mirror descent (Section 4, Proposition 4.4).
- We propose a preconditioned version of EISMD, which adapts the mirror map based on the geometry of the consensus manifold resulting in even faster convergence (Proposition 4.6).
- In Section 5 we illustrate in detail the performance of our algorithms in constrained and unconstrained convex optimization problems.

2. Preliminaries. In this section we fix our notation, state our main assumptions and establish some useful technical lemmas that will be used later.

2.1. Notation. We use $\otimes$ to denote the Kronecker product, I_d the d-dimensional identity matrix and $\mathbf{1}_d$ denotes the d-dimensional vector of ones. $\mathrm{Diag}(a)$ with $a \in \mathbb{R}^d$ denotes a matrix with diagonal elements $[a_1, ..., a_d]$. We use A to denote the $N \times N$ weighted adjacency matrix associated with a graph $\mathcal{G} = (V, E)$. The graph Laplacian is given by $L := \mathrm{Diag}(A\mathbf{1}_N) - A$ and we use the following notation $\mathcal{L} := L \otimes I_d$ with $\mathcal{L} \in \mathbb{R}^{Nd \times Nd}$ to denote the vectorized version of the graph Laplacian. We use $\langle x, y\rangle = x^\top y$ for the standard dot product, and $\langle x, y\rangle_Q = \langle x, Qy\rangle = x^\top Qy$ for the Q-inner product, for some positive definite matrix Q. We use $A \succeq B$ to denote a partial matrix ordering meaning $A - B \succeq 0$. We assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set. We use $\mathcal{D}$ to denote an open set such that $\mathcal{X} \subset \mathrm{cl}(\mathcal{D})$. The set $\mathcal{D}$ will be used to denote the domain of the mirror maps of the mirror descent algorithm. We use $\mathcal{X}^\star$ to denote the dual space of $\mathcal{X}$. The normal cone of $\mathcal{X}$ is defined as $N_{\mathcal{X}}(x) = \{z \in \mathcal{X}^\star \mid \langle z, y - x\rangle \leq 0 \ \ \forall y \in \mathcal{X}\}$.

Given an arbitrary norm $||\cdot||$ on $\mathbb{R}^d$, we will define $B_{\|\cdot\|} := \{v \in \mathbb{R}^d : \|v\| \leq 1\}$. The dual norm $\|\cdot\|_*$ is defined as $\|z\|_* := \sup\{\langle z, v\rangle : v \in B_{\|\cdot\|}\}$. If A is a matrix then $\|A\|_2$ denotes its spectral norm and we assume that the dual norm is compatible

with the spectral norm, i.e. $\|Az\|_* \leq \|A\|_2 \, \|z\|_*$. We will make use of the following generalized Cauchy inequality,

$$|\langle v, w\rangle| \leq \|v\|_* \|w\| \quad \forall w \in \mathcal{X}, \; v \in \mathcal{X}^\star.$$

Since $0 \leq (\|v\|_* - \|w\|))^2 = \|v\|_*^2 + \|w\|^2 - 2\|v\|_*\|w\|$, we also have,

$$\langle v, w\rangle \leq \frac{1}{2}\|v\|_*^2 + \frac{1}{2}\|w\|^2. \tag{2.1}$$

A function g is said to be L-Lipschitz continuous with respect to a norm $\|\cdot\|$ if $\|g(x) - g(y)\| \leq L\|x - y\|, \; \forall x, y \in \mathcal{X}$. The Bregman divergence associated with a convex, differentiable function $g : \mathcal{X} \to \mathbb{R}$ is defined as follows,

$$D_g(x, y) = g(x) - g(y) - \langle \nabla g(y), x - y\rangle.$$

If the second-order derivative of g exists it furthermore holds,

$$\nabla_x D_g(x, y) = \nabla g(x) - \nabla g(y), \;\; \nabla_y D_g(x, y) = \nabla^2 g(y)(y - x). \tag{2.2}$$

The aggregate cost function will be written as $f(\mathbf{x}) = \sum_{i=1}^N f_i(x^i)$, where $\mathbf{x} = [x^{1^T}, \ldots, x^{N^T}]^T$ denotes the stacked vector of particles and each $x_i \in \mathcal{X}$. We will use $(X^\star, \Lambda^\star)$ to denote the set of primal-dual variables that satisfy the first order optimality conditions for (1.1). Unless specified otherwise gradient vectors ∇f are taken with respect to the joint particle vector $\mathbf{x}$ following the usual conventions and the same applies for Hessian matrices.

REMARK 2.1. *The space of the Lagrange multipliers for the consensus constraint, $\boldsymbol{\lambda} \in \Lambda \subset \mathbb{R}^{Nd}$, will play an important role in the definition of the algorithms below. We note that the norm associated with $\boldsymbol{\lambda} \in \Lambda$ will not necessarily be the same as the one used for the primal variables $\mathbf{x} \in \mathcal{X}^N$. We will however use the same notation: $\|\cdot\|$, and its dual $\|\cdot\|_*$ for both spaces, and it will be clear from context which norm is being used. For $\mathbf{w} = [\mathbf{x}^\top, \; \boldsymbol{\lambda}^\top]^\top$ we will use the following mixed norm convention $\|\mathbf{w}\| = \|\mathbf{x}\| + \|\boldsymbol{\lambda}\|$, with the understanding that the two norms could be different. For example, the norm in $\mathcal{X}^N$ could be the ℓ_1, and in Λ the Q-norm (for some positive definite matrix Q) so that, $\|\mathbf{w}\| = \|\mathbf{x}\|_1 + \|\boldsymbol{\lambda}\|_Q$.*

2.2. Assumptions and Definitions.

2.2.1. Optimality Conditions and Model Assumptions. The consensus constraint in (1.1) is satisfied if and only if $\mathcal{L}\mathbf{x} = 0$, where $\mathcal{L}$ denotes the vectorized graph Laplacian. Therefore the optimality conditions for (1.1) are as follows,

$$-\nabla f(\mathbf{x}^\star) - \mathcal{L}\boldsymbol{\lambda}^\star \in N_{\mathcal{X}}(\mathbf{x}^\star).$$

If the solution of (1.1) is in the interior of $\mathcal{X}$, and if f is convex, then we must have that $N_{\mathcal{X}}(\mathbf{x}^\star) = \{\mathbf{0}\}$ for any $\mathbf{x}^\star \in X^\star$. Because the focus of this paper is on the effect of the consensus constraint and its impact on the dynamics of the algorithm, we will assume that the optimal solution of (1.1) is in the interior of $\mathcal{X}$. Because the consensus constraint couples all the particles together, its impact on the algorithm's convergence is far less understood than dealing with separable constraints on $\mathcal{X}$. For certain applications, especially in machine learning, the assumption that the solution lies in the interior of the feasible set holds (e.g. [26, 37]). The extension to the general case requires some minor technical modifications to our convergence analysis similar to [26]. We gather our assumptions so far below.

ASSUMPTION 1. *Each f_i in (1.1) is convex and twice differentiable. The elements in $X^\star$ are in the interior of $\mathcal{X}$.*

Derivatives of f are required so that we can apply Itô's formula. Since we assume that the function is convex this assumption could be relaxed (see [26] Proposition C.2), but the assumption is kept here for simplicity and brevity. We proceed with some standard convexity and smoothness definitions.

DEFINITION 2.2. *We say that $f : \mathcal{X}^N \to \mathbb{R}$ is μ-strongly convex w.r.t. some norm $||\cdot||$ provided that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \geq \mu\|\mathbf{x}-\mathbf{y}\|$. Similarly, a function f is L_f-smooth w.r.t. some norm $\|\cdot\|$ when $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L_f\|\mathbf{x}-\mathbf{y}\|$.*

Some of our results will use the notion of relative strong convexity and smoothness. We refer the reader to [24] for more properties and [12] for the stochastic case. Below we present some definitions and properties that will be useful later on.

DEFINITION 2.3 (Relative strong convexity). *A function $g : \mathcal{X}^N \to \mathbb{R}$ is μ-strongly convex with respect to some convex function h if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}^N$ the following holds,*

$$g(\mathbf{x}) \geq g(\mathbf{y}) + \nabla g(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \mu D_h(\mathbf{x}, \mathbf{y}).$$

Or equivalently, $\langle \mathbf{x}-\mathbf{y}, \nabla g(\mathbf{x}\rangle - \nabla g(\mathbf{y})) \geq \mu\langle \mathbf{x}-\mathbf{y}, \nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\rangle$.

DEFINITION 2.4 (Relative smoothness). *A function $g : \mathcal{X}^N \to \mathbb{R}$ is α-smooth with respect to some function h if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}^N$ the following holds,*

$$g(\mathbf{x}) \leq g(\mathbf{y}) + \nabla g(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \alpha D_h(\mathbf{x}, \mathbf{y}).$$

Or equivalently, $\langle (\mathbf{x}-\mathbf{y}, \nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\rangle \leq \alpha\langle \mathbf{x}-\mathbf{y}, \nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\rangle$.

If we assume that g is μ-strongly convex and α-smooth with respect to h it holds,

$$\mu D_h(\mathbf{x}, \mathbf{y}) \leq D_g(\mathbf{x}, \mathbf{y}) \leq \alpha D_h(\mathbf{x}, \mathbf{y}).$$

We adopt the following definition for the convex conjugate of a relatively strong convex function.

DEFINITION 2.5 (Convex conjugate). *Let $g : \mathcal{X}^N \to \mathbb{R}$ be a μ-strongly convex function with respect to some h. Then $g^*(\mathbf{z}) := \max_{\mathbf{x}\in\mathcal{X}^N} \langle \mathbf{z}^T, \mathbf{x}\rangle - g(\mathbf{x})$ is its Legendre-Fenchel convex conjugate. When g is differentiable, we also have $\nabla g^*(\mathbf{z}) := \arg\max_{\mathbf{x}\in\mathcal{X}^N} \langle \mathbf{z}^T, \mathbf{x}\rangle - g(\mathbf{x})$. and $\nabla g \circ \nabla g^*(\mathbf{z}) = \mathbf{z}$.*

2.2.2. Network Assumptions. We first state our assumptions on the network topology.

ASSUMPTION 2. *The graph $\mathcal{G}$ is connected, undirected and the adjacency matrix A is doubly stochastic.*

These assumptions imply that the graph Laplacian $\mathcal{L}$ is a real symmetric matrix with nonnegative eigenvalues. We will denote the pseudo-inverse of $\mathcal{L}$ by $\mathcal{L}^+$ such that,

$$\mathcal{L}\mathcal{L}^+\mathcal{L} = \mathcal{L}. \tag{2.3}$$

We will use the following definition of the β-regularized Laplacian [10],

$$\mathcal{L}_\beta = \mathcal{L} + \frac{\beta}{N}\mathbf{1}_N\mathbf{1}_N^\top \otimes I_d. \tag{2.4}$$

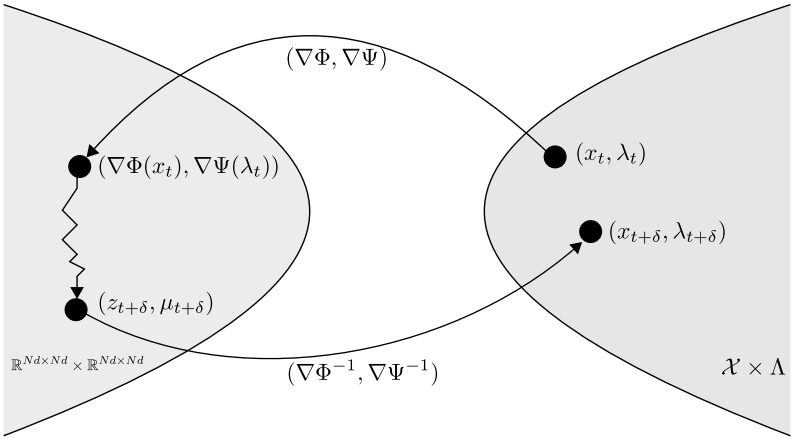


Fig. 2: Stochastic Mirror Descent with two mirror maps. Φ maps the primal variables to the dual space, and Ψ maps the Lagrangian dual variables associated with the consensus constraint to the algebraic dual of the Lagrange multipliers.

Note that the β-regularized Laplacian is positive definite. We define the Rayleigh quotient associated with the β-regularized Laplacian as follows,

$$\kappa_{\beta,N} = \max_{\mathbf{d}_x \in \mathbb{R}^{Nd}} \frac{\|\mathcal{L}_\beta \mathbf{d}_x\|_2^2}{\|\mathbf{d}_x\|_2^2}. \tag{2.5}$$

It holds that [10, p.103],

$$\mathcal{L}_\beta^{-1} = \mathcal{L}^+ + \frac{1}{\beta N}\mathbf{1}_N\mathbf{1}_N^\top \otimes I_d \succeq \mathcal{L}^+, \tag{2.6}$$

where the latter inequality follows from the fact that $\mathbf{1}_d\mathbf{1}_d^\top \otimes I_N$ is positive semidefinite.

LEMMA 2.6. *Let Assumption 2 hold and suppose that $\kappa_{\beta,N}$ is as defined in (2.5) then,*

$$\langle \mathbf{x}, \mathcal{L}\mathbf{x}\rangle \geq \frac{1}{\kappa_{\beta,N}}\|\mathcal{L}\mathbf{x}\|_2^2.$$

Proof. Using the definition of the pseudo-inverse in (2.3) and its relationship with the inverse of the regularized Laplacian in (2.6) we obtain,

$$\begin{aligned}\langle \mathbf{x}, \mathcal{L}\mathbf{x}\rangle &= \langle \mathbf{x}, \mathcal{L}\mathcal{L}^+\mathcal{L}\mathbf{x}\rangle = \langle \mathcal{L}\mathbf{x}, (\mathcal{L}_\beta^{-1} - \frac{1}{\beta N}\mathbf{1}_d\mathbf{1}_d^\top \otimes I_N)\mathcal{L}\mathbf{x}\rangle \\ &= \langle \mathcal{L}\mathbf{x}, \mathcal{L}_\beta^{-1}\mathcal{L}\mathbf{x}\rangle,\end{aligned}$$

where in the last equality we used the fact that $(\mathbf{1}_d\mathbf{1}_d^\top \otimes I_N)\mathcal{L} = 0$. Since $\mathcal{L}_{\beta,N} \preceq \kappa_{\beta,N} I$ then $\mathcal{L}_{\beta,N}^{-1} \succeq \kappa_{\beta,N}^{-1} I$ and the result follows. ∎

2.2.3. Mirror Maps. The role of the mirror map, $\Phi : \mathcal{D} \to \mathbb{R}$, in the Mirror Descent algorithm is to transform the primal $x \in \mathcal{X}$ variables to the dual space $\nabla\Phi(x) \subset \mathbb{R}^d$. The dual variables will be denoted by z, i.e $\nabla\Phi(x) = z$. In the algorithm

proposed in this paper we will use two mirror maps. The first mirror map Φ, is used to transform the primal variables x. The second mirror map Ψ, is used to transform the Lagrangian dual variables λ associated with the consensus constraint in (1.1). The algebraic dual variables will be denoted by μ, i.e. $\nabla\Psi(\lambda) = \mu$. When no confusion arises between Lagrangian and algebraic dual variables we will refer to them simply as dual variables. Figure 2 explains the main steps in mirror descent with the two maps. At time-step t the primal-dual pair (x_t, λ_t) is mapped to $(z_t, \mu_t) = (\nabla\Phi(x_t), \nabla\Psi(\lambda_t))$. The algorithm then follows the stochastic dynamics specified in Section 3. For example, a variant of the proposed scheme performs a gradient descent on the Augmented Lagrangian w.r.t the primal x variables, and a dual ascent w.r.t the Lagrangian dual variables λ (see Section 3 for a detailed explanation). The inverse $(\nabla\Phi^{-1}, \nabla\Psi^{-1})$ maps the algebraic duals back to the primal space $\mathcal{X} \times \Lambda$. The proposed algorithm, and its variants, are described in Section 3. Below we state our related assumptions:

ASSUMPTION 3 (Mirror map). *$\Phi : \mathcal{D} \to \mathbb{R}$ is twice differentiable, μ_Φ-strongly convex and L_Φ-smooth w.r.t. some norm $||\cdot||$. The same holds for $\Psi : \Lambda \to \mathbb{R}$ with constants μ_Ψ,L_Ψ. respectively. We furthermore make the additional assumption $\nabla\Phi^*(\mathbb{R}^d) = \mathcal{X}$, $\nabla\Psi^*(\mathbb{R}^d) = \Lambda$, Φ^* is L_{Φ^*}-smooth and assume uniform boundedness of the Laplacians of Φ^*, Ψ^* such that $||\Delta\Phi^*||_\infty, ||\Delta\Psi^*||_\infty < \infty$.*

The assumption that $\nabla\Phi^*$ maps directly to $\mathcal{X}$ (and similarly for Ψ) avoids the need for projections. Extending our results without this assumption is possible by following a route similar to [26].

A useful property of the Bregman divergence induced by mirror maps that satisfy our assumptions is the following,

$$D_{\Phi^*}(z, z') = D_\Phi(x', x), \tag{2.7}$$

where $z = \nabla\Phi(x)$ and $z' = \nabla\Phi(x')$. For $x, y, z \in \mathbb{R}^d$ we have the triangle property for Bregman divergences (see Lemma 9.11 in [4])

$$\langle x - y, \nabla\Phi(z) - \nabla\Phi(y)\rangle = D_\Phi(x, y) + D_\Phi(y, z) - D_\Phi(x, z). \tag{2.8}$$

We also make use of the following property,

$$D_f(\mathbf{x}, \mathbf{x}') \leq \alpha(\Phi) D_\Phi(\mathbf{x}', \mathbf{x}) \tag{2.9}$$

where $\alpha(\Phi) = \frac{L_f L_\Phi}{\mu_\Phi}$. This property follows from the relative smoothness assumption combined with the strong convexity and Lipschitz assumption on Φ,

$$D_f(\mathbf{x}, \mathbf{x}') \leq L_f D_\Phi(\mathbf{x}, \mathbf{x}') \leq \frac{L_f L_\Phi}{2}\|\mathbf{x}' - \mathbf{x}\|^2 \leq \frac{2L_f L_\Phi}{2\mu_\Phi} D_\Phi(\mathbf{x}', \mathbf{x}),$$

where with slight abuse we denote $\Phi(\mathbf{x}) = \sum_{i=1}^N \Phi(x^i)$. We will use the following Rayleigh quotient,

$$\kappa_N = \max_{\mathbf{d}_x \in \mathbb{R}^{Nd}, \mathbf{d}_\lambda \in \mathbb{R}^{Nd}} \max\left\{ \frac{\|\mathcal{L}^{\frac{1}{2}}\mathbf{d}_x\|_2^2}{\|\mathbf{d}_x\|^2}, \frac{\|\mathcal{L}^{\frac{1}{2}}\mathbf{d}_\lambda\|_2^2}{\|\mathbf{d}_\lambda\|^2} \right\}. \tag{2.10}$$

Note that the norms for $\mathbf{d}_x$ and $\mathbf{d}_\lambda$ in the definition above may be different (see Remark 2.1). We will also need the following generalized Rayleigh quotient,

$$\kappa_g = \inf_{\mathbf{x}, \mathbf{d}_x, \mathbf{d}_\lambda \in \mathbb{R}^{Nd}} \frac{\|A(\mathbf{x})[\mathbf{d}_x^T, \mathbf{d}_\lambda^T]^\top\|^2_{\nabla^2\Phi^*(\mathbf{z})}}{\|\mathbf{d}_x\|^2 + \|\mathbf{d}_\lambda\|^2}, \tag{2.11}$$

where $A(\mathbf{x}) = [\nabla^2 f(\mathbf{x}) + \mathcal{L}, \mathcal{L}] \in \mathbb{R}^{Nd \times 2Nd}$. If strong convexity is assumed then κ_g is strictly positive. This fact is not obvious since $A(\mathbf{x})$ is not a square matrix, and the norm used in the definition of (2.11) is not standard, we therefore provide a short proof below.

LEMMA 2.7. *Suppose Assumptions 1-3 hold and that f is relatively strongly convex with respect to Φ, then κ_g defined in (2.11) is positive.*

Proof. We note that $A(\mathbf{x})$ can be obtained by removing the last Nd columns and rows of the following matrix,

$$B(\mathbf{x}) := \begin{bmatrix} \nabla^2 f(\mathbf{x}) + \mathcal{L} & \mathcal{L} \\ -\mathcal{L} & 0 \end{bmatrix}.$$

Let $\mathbf{d} = [\mathbf{d}_x, \mathbf{d}_\lambda]^\top$ and note that $\langle \mathbf{d}, B(\mathbf{x})\mathbf{d} \rangle = \mathbf{d}_x^\top \nabla^2 f(\mathbf{x}) \mathbf{d}_x$. It follows from the relative strong convexity assumption that $B(\mathbf{x}) \succeq \mu_\Phi \nabla^2 \Phi(\mathbf{x})$ and therefore $\|B(\mathbf{x})\|^2_{\nabla^2 \Phi(\mathbf{z})^{-1}} \succ 0$. Since $A(\mathbf{x})$ can be obtained by removing the last Nd columns and rows of $B(\mathbf{x})$ the result follows from the interlacing theorem for singular values, see e.g. Theorem 3.1.3 in [18]. ∎

Lastly, we will need the following result.

LEMMA 2.8. *Suppose that Assumptions 1-3. Then for an arbitrary optimal primal dual pair $(\mathbf{x}^\star, \boldsymbol{\lambda}^*)$ we have*

$$\begin{aligned} \|\nabla f(\mathbf{x}) + \mathcal{L}\boldsymbol{\lambda} + \mathcal{L}\mathbf{x}\|_{\nabla \Phi^*(\mathbf{z})} &\geq \frac{2\kappa_g}{\hat{\mu}} \left(\sum_{i=1}^N D_\Phi(x^\star, x^i) + D_\Psi(\lambda^\star, \lambda^i) \right) \\ &= \frac{2\kappa_g}{\hat{\mu}} \left(\sum_{i=1}^N D_{\Phi^*}(z^i, z^\star) + D_{\Psi^*}(\mu^i, \mu^\star) \right), \end{aligned}$$

where $\hat{\mu} = \min\{\mu_\Phi, \mu_\Psi\}$

Proof. Since f is twice differentiable there exists an $\mathbf{y}$ on the line segment joining $\mathbf{x}$ and $\mathbf{x}^\star$ such that $\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star) = \langle \nabla^2 f(\mathbf{y}), \mathbf{x} - \mathbf{x}^\star \rangle$. We then have,

$$\begin{aligned} \|\nabla f(\mathbf{x}) + \mathcal{L}\boldsymbol{\lambda} + \mathcal{L}\mathbf{x}\|^2_{\nabla^2 \Phi^*(\mathbf{z})} &= \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star) + \mathcal{L}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*) + \mathcal{L}(\mathbf{x} - \mathbf{x}^*)\|^2_{\nabla^2 \Phi^*(\mathbf{z})} \\ &= \|A(\mathbf{y})[\mathbf{x}_t - \mathbf{x}^*, \boldsymbol{\lambda} - \boldsymbol{\lambda}^*]^T\|^2_{\nabla^2 \Phi^*(\mathbf{z})} \\ &\geq \kappa_g \left(\|\mathbf{x}^* - \mathbf{x}_t\|^2 + \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}]\|^2 \right) \\ &\geq \frac{2\kappa_g}{\hat{\mu}} \left(\sum_{i=1}^N D_\Phi(x^\star, x^i) + D_\Psi(\lambda^\star, \lambda^i) \right). \end{aligned}$$

We use (2.7) to obtain the bound in terms of the (algebraic) dual variables. ∎

3. Distributed Stochastic Mirror Descent: Exact and Preconditioned Dynamics. In this section we introduce different variants of distributed MD algorithms. We adopt a dynamical systems point of view for our analysis. Numerical realizations of the proposed schemes are discussed in Section 5. For an introduction to the original mirror descent algorithm we refer the interested reader to [4, Ch. 9].

Interacting Stochastic Mirror Descent (ISMD). The starting point of our analysis is the Interacting Stochastic Mirror Descent (ISMD) algorithm. This algorithm was proposed in [31] but its convergence was only established in the linear case. The case

where all the functions are identical (i.e. $f_1 = \ldots = f_N$) and strongly convex was analyzed in [6]. The discrete time version of the algorithm for the general convex case was discussed in [13], but exact convergence was only established under a diminishing step-size strategy. In the continuous time setting the dynamics of ISMD are as follows,

$$dz_t^i = -\eta \nabla f_i(x_t^i)dt + \epsilon \sum_{j=1}^{N} A_{ij}(z_t^j - z_t^i)dt + \sigma dB_t^i, \quad x_t^i = \nabla \Phi^*(z_t^i), \tag{3.1}$$

for particles $i = 1, ..., N$, and where B_t^i are independent Brownian motions. The matrix $A = \{A_{ij}\}_{i,j=1}^N$ is an $N \times N$ doubly-stochastic matrix representing the interaction weights and η, ϵ are tuning constants representing the learning rate and interaction strength, respectively. For simplicity in most of the subsequent analysis we set $\eta = \epsilon = 1$, but it is straightforward to extend the results for arbitrary values of η and ϵ. In the context of modern large scale applications, we note that understanding convergence under the presence of noise is often motivated from computational considerations such as when sub-sampling the gradient f or the interaction graph when N is large.

Using the graph Laplacian, we can rewrite the evolution in vector form as

$$d\mathbf{z}_t = \left(-\nabla f(\mathbf{x}_t) - \mathcal{L}\mathbf{z}_t\right) dt + \sigma d\mathbf{B}_t, \;\; \mathbf{x}_t = \nabla \Phi^*(\mathbf{z}_t), \tag{3.2}$$

where $\mathbf{B}_t := ((B_t^1)^T, ..., (B_t^N)^T)^T$. In the case where the mirror map above is the ℓ_2 norm it is known that even in the deterministic case the dynamics in (3.2) will not converge to the exact solution (see [36, 35]). In Section 4.1 we show that in general the dynamics in (3.2) also fails to converge to the exact solution of (1.1) and identify that this can occur only under additional assumptions. This motivates proposing a different dynamics below.

Exact Interacting Stochastic Mirror Descent (EISMD). To address the limitations of the ISMD algorithm discussed above we propose the following,

$$\begin{aligned} d\mathbf{z}_t &= -\nabla f(\mathbf{x}_t)dt - \mathcal{L}\mathbf{x}_t dt - \mathcal{L}\boldsymbol{\lambda}_t dt + \sigma d\mathbf{B}_t, \\ d\boldsymbol{\lambda}_t &= \mathcal{L}\mathbf{x}_t dt, \end{aligned} \tag{3.3}$$

with $\mathbf{x}_t = \nabla \Phi^*(\mathbf{z}_t)$ and initial conditions $\boldsymbol{\lambda}_0 = 0$, $\mathbf{z}_0 = \nabla \Phi(\mathbf{x}_0)$. The idea behind this method is to add historical feedback into the algorithm through the integral $\int_0^t \mathcal{L}\mathbf{x}_s ds$. At optimality this will cancel out the gradient term $\nabla f(\mathbf{x}_t)$. In Section 3.1 we show that the drift term in EISMD is related to the Augmented Lagrangian. We exploit this connection in the theoretical analysis in Section 4. Compared to previous works considered in Table 1 this algorithm integrates past information into the dynamics through the integral term and is applicable to the mirror descent framework. For the case where $\sigma = 0$, (3.3) has been considered in [38]. Here we extend the ideas in [38] to allow f being only convex, adding Brownian noise and considering general preconditioning.

Exact Preconditioned Interacting Stochastic Mirror Descent (EPISMD). A potential limitation of ISMD in (3.3) is that the mirror map Φ only captures the geometry of the primal space $\mathcal{X} \subset \mathbb{R}^d$. Even if $\mathcal{X} = \mathbb{R}^d$, our optimization problem is still constrained to the consensus manifold $\mathcal{X}^C$. In order to incorporate information from the consensus constraint we introduce a second mirror map Ψ that acts on the Lagrangian dual variable (see Figure 8). The preconditioned dynamics (EPISMD) is given by,

$$\begin{aligned} d\mathbf{z}_t &= -\nabla f(\mathbf{x}_t)dt - \mathcal{L}\mathbf{x}_t dt - \mathcal{L}\boldsymbol{\lambda}_t dt + \sigma d\mathbf{B}_t, \\ d\boldsymbol{\mu}_t &= \mathcal{L}\mathbf{x}_t dt, \;\; \boldsymbol{\lambda}_t = \nabla \Psi^*(\boldsymbol{\mu}_t), \end{aligned} \tag{3.4}$$

where $\boldsymbol{\mu}$ is the mirrored version of the $\boldsymbol{\lambda}$ variable using the mirror map Ψ. As we will show in Section 3.1, this algorithm is related to preconditioning $\boldsymbol{\lambda}_t$ and we will later show numerically that it can lead to faster convergence.

3.1. Preconditioning and the Augmented Lagrangian. The Augmented Lagrangian for the standard gradient descent setting is well-known (see e.g. [28, 40, 16]). The Alternating Direction Method of Multipliers (ADMM) is based on an Augmented Lagrangian with a Bregman divergence [39, 41] and Riemannian primal-dual methods over the Augmented Lagrangian were considered in [2] (see also [15] for a continuous time analysis of ADMM). Below we discuss the relationship between the different variants of the proposed methods.

Consider the Augmented Lagrangian,

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \langle \mathcal{L}\mathbf{x}, \boldsymbol{\lambda} \rangle + \frac{1}{2}\|\mathcal{L}^{\frac{1}{2}}\mathbf{x}\|_2^2. \tag{3.5}$$

The Augmented Lagrangian Method (ALM) proceeds by a descent step in the primal variables and an ascent step in the Lagrangian dual variables. When Bregman divergence is used to define the ALM, then (in discrete time) the iterates are given by the following,

$$\begin{aligned} \mathbf{x}_{t+\Delta t} &= \operatorname*{argmin}_{\mathbf{d}} \ \Delta t \langle \nabla_{\mathbf{x}} L(\mathbf{x}_t, \boldsymbol{\lambda}_t), \mathbf{d} \rangle + D_\Phi(\mathbf{d}, \mathbf{x}_t), \\ \boldsymbol{\lambda}_{t+\Delta t} &= \operatorname*{argmax}_{\mathbf{d}} \ \Delta t \langle \nabla_{\boldsymbol{\lambda}} L(\mathbf{x}_t, \boldsymbol{\lambda}_t), \mathbf{d} \rangle - \frac{1}{2}\|\mathbf{d} - \boldsymbol{\lambda}_t\|^2. \end{aligned} \tag{3.6}$$

Writing down the optimality conditions of the two subproblems above and taking the limit $\Delta t \to 0$, we obtain the deterministic version of (3.3).

Alternatively, we can rewrite the dynamics of ISMD in (3.2) in terms of $\mathbf{x}_t$ drawing a connection with preconditioned or Riemannian gradient descent [26]. Note that $\nabla\Phi(\nabla\Phi^*(\mathbf{z})) = \mathbf{z}$. Differentiating this w.r.t. z we obtain $\nabla^2\Phi^*(\mathbf{z})\nabla^2\Phi(\nabla\Phi^*(\mathbf{z})) = I_{dN}$. Therefore,

$$\nabla^2\Phi^*(\mathbf{z}) = \nabla^2\Phi(\mathbf{x})^{-1}. \tag{3.7}$$

By applying Itô's lemma to $\nabla\Phi^*(\mathbf{z}_t)$, using the definition of the Bregman divergence and properties (2.2) and (3.7),

$$\begin{aligned} dx_t^i = &- \nabla^2\Phi(x_t^i)^{-1}\nabla f_i(x_t^i)dt - \nabla^2\Phi(x_t^i)^{-1}\textstyle\sum_{j=1}^N A_{ij}(z_t^i - z_t^j)dt \\ &+ \frac{1}{2}\sigma^2\nabla^2(\nabla\Phi^*(z_t^i))dt + \sigma\nabla^2\Phi(x_t^i)^{-1}dB_t^i. \end{aligned}$$

From this expression we see that mirror descent is a preconditioned algorithm where the choice of preconditioner is determined through the function Φ. Preconditioned dynamics have been studied in previous work to improve communication complexity in a distributed setting [17] or speed up mixing rates of the dynamics [20]. The additional drift term with the third-order gradient of the mirror map arises as a correction term due to the nonlinearity of the mirror map. A similar scheme can be derived for EISMD.

Preconditioning of the Interaction. The choice of mirror map can be used to precondition the dynamics and accelerate convergence. Now we show, based on the primal-dual interpretation, that the dynamics in (3.4) can also be cast into the preconditioned setting, where preconditioning is done on the $\boldsymbol{\lambda}_t$ variable. Motivated

by the augmented Lagrangian and preconditioning, as opposed to the dynamics in (3.6), one could redefine the dual variable dynamics as,

$$\boldsymbol{\lambda}_{t+\Delta t} = \operatorname*{argmax}_{\mathbf{d}} \Delta t \langle \nabla_{\boldsymbol{\lambda}} L(\mathbf{x}_t, \boldsymbol{\lambda}_t), \mathbf{d} \rangle - D_{\Psi}(\mathbf{d}, \boldsymbol{\lambda}).$$

The first order optimality conditions of the problem above are,

$$\nabla \Psi(\boldsymbol{\lambda}_{t+\Delta t}) = \nabla \Psi(\boldsymbol{\lambda}_t) + \Delta t \mathcal{L} \mathbf{x}_t.$$

Taking $\Delta t \to 0$ we obtain,

$$d\boldsymbol{\lambda}_t = \nabla^2 \Psi(\boldsymbol{\lambda}_t)^{-1} \mathcal{L} \mathbf{x}_t dt, \tag{3.8}$$

where we used $\nabla^2 \Psi^*(\boldsymbol{\mu}) = \nabla^2 \Psi(\boldsymbol{\lambda})^{-1}$ as in (3.7). Therefore, the method in (3.4) allows for additional flexibility due to the preconditioning of the dual variable.

4. Convergence Analysis. In this section we present a convergence analysis for the exact interacting mirror descent algorithm. EISMD with a strongly convex objective is able to converge *exponentially* fast to an area of the optimum, however the size of this area can be made arbitrarily small by decreasing σ.

4.1. When first-order optimization fails. Our first result shows that if there exists an $x^\star$ such that $\nabla f_i(x^\star) = 0$ for all $i = 1, ..., N$, exact consensus can be obtained for ISMD.

LEMMA 4.1. *Let Assumptions 1-3 hold. Consider the dynamics in (3.2) with $\sigma = 0$. If*

$$\bigcap_{i=1}^{N} \{\nabla f_i(x) = 0\} \neq \emptyset, \tag{4.1}$$

then $\lim_{t\to\infty} x_t^i = x^\star$.

Proof. Let x_0 be the initial point of the algorithm, and let $z^\star$ be an optimal (dual) point closest to $z_0 = \nabla \Phi(x_0)$ with respect to the divergence generated by Φ^*,

$$z^\star = \arg\min_{z \in Z^\star} D_{\Phi^*}(z, z_0)$$

where $Z^\star = \{z \mid z = \nabla \Phi(x), \exists x \in \mathcal{X} : \nabla f_i(x) = 0, i = 1, \dots, N\}$. By assumption (4.1), $Z^\star$ is not empty. With a slight abuse of notation we let $x^\star = \nabla \Phi^*(z^\star)$ and note that $(x^\star, z^\star)$ is an equilibrium point for (3.2) (for a strongly convex function it is also the unique equilibrium point, but here we only assume convexity of f).

Define the Lyapunov candidate function $V_t = \sum_{i=1}^N D_{\Phi^\star}(z_t^i, z^\star)$. Then we obtain,

$$dV_t = \sum_{i=1}^{N} (x^\star - x_t^i)^T \nabla f_i(x_t^i) dt + \sum_{i=1}^{N} (x_t^i - x^\star)^T \sum_{j=1}^{N} A_{ij} (z_t^j - z_t^i) dt.$$

Under convexity of f and optimality at $x^\star$ we have

$$\sum_{i=1}^{N} (x^\star - x_t^i)^T \nabla f_i(x_t^i) \leq \sum_{i=1}^{N} (f_i(x^\star) - f_i(x_t^i)) \leq 0.$$

By the triangle equality of the Bregman divergence in (2.8),

$$\begin{aligned}(x_t^i - x^\star)^T(z_t^j - z_t^i) &= -(x^\star - x_t^i)^T(z_t^j - z_t^i)\\ &= -(\nabla\Phi^*(z^\star) - \nabla\Phi^*(z_t^i))^T(z_t^j - z_t^i)\\ &= -D_{\Phi^*}(z_t^j, z_t^i) - D_{\Phi^*}(z_t^i, z^\star) + D_{\Phi^*}(z_t^j, z^\star).\end{aligned}$$

Then,

$$\begin{aligned}(4.2)\quad &\sum_{i=1}^N\sum_{j=1}^N A_{ij}(x_t^i - x^\star)^T(z_t^j - z_t^i)\\ &= \sum_{i=1}^N\sum_{j=1}^N A_{ij}\left(-D_{\Phi^*}(z_t^j, z_t^i) - D_{\Phi^*}(z_t^i, z^\star) + D_{\Phi^*}(z_t^j, z^\star)\right) \le 0,\end{aligned}$$

where we have used $A_{ij} \ge 0$, $\sum_{i=1}^N\sum_{j=1}^N D_{\Phi^*}(z_t^i, z^\star) = \sum_{i=1}^N\sum_{j=1}^N D_{\Phi^*}(z_t^j, z^\star)$, and $D_{\Phi^*}(z_t^j, z_t^i) \ge 0$. Since $V_t > 0$ for $\mathbf{z} \neq \mathbf{1}_N \otimes z^\star$, $V_t = 0$ when $\mathbf{z} = \mathbf{1}_N \otimes z^\star$ and $dV_t \le 0$ with equality only at $\mathbf{z} = \mathbf{1}_N \otimes z^\star$ we conclude that V_t is a Lyapunov function for $\mathbf{z}_t$. Since $D_{\Phi^*}(z_t^i, z^\star) = D_\Phi(x^\star, x_t^i)$ the statement follows. ∎

The Lemma above can be extended to an if and only if statement based on the arguments of [35, Theorem 1], but precise details lie beyond the scope of this paper. If (4.1) is violated, even with the right choice of mirror map, achieving exact consensus is not possible. In general imposing $x^\star$ to satisfy (4.1) is quite restrictive as $\nabla f(\mathbf{x}) = \sum_{i=1}^N \nabla f_i(x^\star) = 0$ does not necessarily imply $\nabla f_i(x^\star) = 0$ for all $i = 1, ..., N$. The crucial point to realise here is that if and only if (4.1) holds then $(\mathbf{x}^\star, \mathbf{z}^\star)$ will also be the minimizer of $f(\mathbf{x}) + \frac{1}{2}\mathbf{z}^T\mathcal{L}\mathbf{z}$; see [35, Lemma 7] for details. As a result one can establish consensus at equilibrium and V_t will approach zero at large t. If $\mathbf{x}^\star$ does not satisfy (4.1) and one has just $\nabla f(\mathbf{x}^\star) = 0$, the arguments above can be used to establish exponential but *approximate* convergence for (3.2).

PROPOSITION 4.2 (Approximate convergence of (3.2)). *Let Assumptions 1-3 hold and assume that f is μ_f-strongly convex w.r.t. Φ. Let $\mathbf{x}^\dagger = \arg\min\{f(\mathbf{x}) + \frac{1}{2}\nabla\Phi(\mathbf{x})^T\mathcal{L}\nabla\Phi(\mathbf{x})\}$ with $\mathbf{x}^\dagger = 1_N \otimes x^\dagger$ and $V_t = \frac{1}{N}\sum_{i=1}^N D_{\Phi^*}(z_t^i, z^\dagger)$, where z_t^i obeys the dynamics of (3.2). Then*

$$\mathbb{E}\left[V_t\right] \le e^{-\mu_f t}\frac{1}{N}\sum_{i=1}^N D_{\Phi^*}(z_0^i, z^\dagger) + \frac{\sigma^2}{2\mu_f}||\Delta\Phi^*||_\infty + \frac{C_f}{\mu_f},$$

where $C_f \ge 0$ is a constant depending on f.

The proof and details are in Appendix A.2. While the relative strong convexity of the objective function can speed up convergence, only approximate convergence can be obtained even with $\sigma = 0$. In this setup the additional preconditioning via the mirror map does not facilitate exact convergence nor consensus. When (4.1) holds the arguments above can be used to show that $C_f = 0$ thus achieving exact convergence.

4.2. Exact Interacting Stochastic Mirror Descent Analysis. In this section we show that the EISMD algorithm in (3.3) allow us to converge close to the optimum and this convergence is exact when $\sigma = 0$.

We note that $x^i = x^j$ for $(i,j) \in E$ if and only if $\mathcal{L}\mathbf{x} = 0$, therefore the problem in (1.1) can be written as,

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}^N} \quad & f(\mathbf{x}) + \frac{1}{2}\|\mathcal{L}^{\frac{1}{2}}\mathbf{x}\|_2^2 \\ \text{s.t.} \quad & \mathcal{L}\mathbf{x} = 0. \end{aligned}$$

The application of the Karush–Kuhn–Tucker (KKT) conditions to the problem above implies that if $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = (\mathbf{1}_N \otimes x^\star, \mathbf{1}_N \otimes \lambda^\star)$ is an arbitrary point that satisfies the first order optimality conditions for (1.1), then when $\sigma = 0$, $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ is also an equilibrium point of (3.3). The connection of the dynamics of (3.3) and the augmented Lagrangian is key to the convergence analysis below.

The analysis of the algorithm in (3.3) is based on the following Lyapunov function,

$$V(\mathbf{x}, \boldsymbol{\lambda}) = c(V_1(\mathbf{x}) + V_2(\boldsymbol{\lambda})) + V_3(\mathbf{x}, \boldsymbol{\lambda}), \tag{4.3}$$

where,

$$V_1(\mathbf{x}) = \sum_{i=1}^N D_\Phi(x^\star, x^i), \quad V_2(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2,$$

$$V_3(\mathbf{x}, \boldsymbol{\lambda}) = D_f(\mathbf{x}, \mathbf{x}^\star) + \langle \mathbf{x} - \mathbf{x}^\star, \mathcal{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star)\rangle + \frac{1}{2}\|\mathcal{L}^{\frac{1}{2}}\mathbf{x}\|_2^2,$$

and $c \geq \underline{c}$ with $\underline{c} > 0$ to be specified below for different contexts. In the case that f is only convex and thus multiple minimizers might exist, then we define the optimal primal dual pair, $(x^\star, \lambda^\star)$ to be the one that is closest to the initial conditions,

$$(x^\star, \lambda^\star) = \arg \min_{x,\lambda \in (X^\star, \Lambda^\star)} D_\Phi(x, x_0) + \|\lambda - \lambda_0\|_2^2.$$

Below we establish upper and lower bounds for (4.3) that will be useful later on.

LEMMA 4.3. *Let Assumptions 1-3 hold. Then (4.3) satisfies the following,*

i $V(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = 0$.

ii.a *If* $c \geq \max\{\kappa_N/\mu_\Phi, \kappa_N\}$ *then*

$$V(\mathbf{x}, \boldsymbol{\lambda}) \geq \frac{1}{2}(\mu_\Phi c - \kappa_N)\|\mathbf{x} - \mathbf{x}^\star\|^2 + \frac{1}{2}(c - \kappa_N)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2 \geq 0.$$

iii.a *Let* $\hat{\mu} = \min\{\mu_\Phi, 2\}$. *Then,*

$$V(\mathbf{x}, \boldsymbol{\lambda}) \leq \left(c + \frac{3\kappa_N + 2\alpha(\Phi)}{\hat{\mu}}\right)\left(\sum_{i=1}^N D_\Phi(x^\star, x_i) + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2\right), \tag{4.4}$$

where $\alpha(\Phi) = L_f L_\Phi/\mu_\Phi$ *was defined in (2.9), and* κ_N *in (2.10).*

If, in addition, f *is* μ_f*-strongly convex relative to* Φ *then:*

ii.b *For any* $c \geq \max\{((\kappa_N - \mu_f L_\Phi)/\mu_\Phi, \kappa_N\}$,

$$V(\mathbf{x}, \boldsymbol{\lambda}) \geq \frac{1}{2}(\mu_\Phi c + \mu_f L_\Phi - \kappa_N)\|\mathbf{x} - \mathbf{x}^\star\|^2 + \frac{1}{2}(c - \kappa_N)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2 \geq 0.$$

Proof. Property (i) is obvious. For (ii.a) we bound V_1 using the strong convexity of Φ:

$$V_1(\mathbf{x}) = \sum_{i=1}^N D_\Phi(x^\star, x) \geq \frac{\mu_\Phi}{2}\|x^\star - x\|^2.$$

We note that the convexity of f implies that $D_f(\mathbf{x}, \mathbf{x}^\star) \geq 0$, and we bound V_3 as follows,

$$V_3(\mathbf{x}, \boldsymbol{\lambda}) \geq \langle \mathbf{x} - \mathbf{x}^\star, \mathcal{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star) \rangle \geq -\frac{1}{2}(\|\mathcal{L}^{\frac{1}{2}}\mathbf{x} - \mathbf{x}^\star\|_2^2 + \|\mathcal{L}^{\frac{1}{2}}\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2)$$
$$\geq -\frac{\kappa_N}{2}(\|\mathbf{x} - \mathbf{x}^\star\|^2 + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2)$$

where in the second inequality we used (2.1) and in the third one (2.10).

If, in addition, f is strongly convex relative to Φ then,

$$\sum_{i=1}^N D_f(x^\star, x^i) \geq \mu_f \sum_{i=1}^N D_\Phi(x^\star, x^i) \geq \frac{\mu_f L_\Phi}{2}\|\mathbf{x}^\star - \mathbf{x}\|^2$$

Using the preceding inequality to bound V_3 we obtain the bound (ii.b).

For the upper bound in (iii.a) we bound the the first term in V_3 using the symmetry bound in (2.9),

$$\sum_{i=1}^N D_f(x^\star, x_i) \leq \alpha(\Phi) \sum_{i=1}^N D_\Phi(x^\star, x_i).$$

For the second term in V_3 we use (2.1) again and for any $\gamma > 0$,

$$\begin{aligned}
\frac{1}{2}\langle \mathbf{x} - \mathbf{x}^\star, \mathcal{L}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star) \rangle &\leq \frac{\kappa_N}{2}(\gamma\|\mathbf{x} - \mathbf{x}^\star\|^2 + \frac{1}{\gamma}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2) \\
&\leq \frac{\kappa_N \gamma}{\mu_\Phi} \sum_{i=1}^N D_\Phi(x^\star, x^i) + \frac{\kappa_N}{2\gamma}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2 \\
&\leq \frac{\kappa_N}{\hat{\mu}}(\gamma \sum_{i=1}^N D_\Phi(x^\star, x^i) + \frac{1}{\gamma}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2) \\
&\leq \frac{\kappa_N}{\hat{\mu}} \sum_{i=1}^N D_\Phi(x^\star, x^i) + \frac{\kappa_N + \alpha(\Phi)}{\hat{\mu}}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2^2,
\end{aligned}$$

where in the second inequality we used the relative strong convexity of Φ and for the last inequality we set $\gamma = \frac{\kappa_N}{\kappa_N + \alpha(\Phi)} \leq 1$. Finally, for the last term in V_3 we use the bound from (2.10), the strong convexity of Φ and the definition of $\hat{\mu}$,

$$\|\mathcal{L}^{\frac{1}{2}}\mathbf{x}\|_2^2 = \|\mathcal{L}^{\frac{1}{2}}(\mathbf{x} - \mathbf{x}^\star)\|_2^2 \leq \kappa_N\|\mathbf{x} - \mathbf{x}^\star\|^2 \leq \frac{2\kappa_N}{\hat{\mu}} \sum_{i=1}^N D_\Phi(x^\star, x^i)$$

Using the upper bounds for the three terms in V_3 we obtain the bound in (4.4). ∎

We then have the following convergence result for the dynamics in (3.3).

PROPOSITION 4.4 (Convergence of the dynamics in (3.3)). *Let Assumptions 1-3 hold and assume $\kappa_g > 0$. Consider the dynamics in (3.3) and V_t as defined in (4.3). Let $\hat{\mu} = \min\{\mu_\Phi, 2\}$. Then it holds,*

$$\mathbb{E}[V_T] \leq e^{-rT}\mathbb{E}[V_0] + \int_0^T e^{-r(T-s)} M(\mathbf{x}_s) ds \tag{4.5}$$

where $V_T := V(\mathbf{x}_T, \boldsymbol{\lambda}_T)$,

$$r = \frac{2\kappa_g}{c\hat{\mu} + 2\alpha(\Phi) + 3\kappa_N},$$

$$M(\mathbf{x}) = c\frac{\sigma^2}{2}(\mathrm{tr}(C_1(\mathbf{x})) + \langle \mathbf{x} - \mathbf{x}^\star, \Delta \cdot \nabla\Phi^*(\mathbf{z})\rangle) + \frac{\sigma^2}{2}\mathrm{tr}(C_2(\mathbf{x}))$$

$$C_1(\mathbf{x}) = \sum_{i=1}^{N} \nabla^2\Phi(x^i)^{-1}\nabla^2_{xx}D_\Phi(x^\star, x^i)\nabla^2\Phi(x^i)^{-1}$$

$$C_2(\mathbf{x}) = \nabla^2\Phi(\mathbf{x})^{-1}(\nabla^2 f(x) + \mathcal{L})\nabla^2\Phi(\mathbf{x})^{-1},$$

and $c \geq 2\kappa_{\beta,N}$.

Proof. Let c be as in Lemma 4.3 to ensure non-negativity of the Lyapunov function. Since $x^i = \nabla\Phi^*(z^i)$ it follows from Itô's Lemma that,

$$dx^i_t = \nabla^2\Phi^*(z^i)dz^i_t + \frac{1}{2}\sigma^2\Delta \cdot \nabla\Phi^*(z^i)dt,$$

where the j^{th} element of the Itô correction term is $[\Delta \cdot \nabla\Phi^*(z^i)]_j = \sum_{k=1}^{d} \partial^2_{kk}\partial_j\Phi^*(z^i)$. For ease of exposition we define the following terms,

$$dM^1_t = \frac{\sigma^2}{2}(\mathrm{tr}(C_1(\mathbf{x}_t)) + \langle \mathbf{x}_t - \mathbf{x}^\star, \Delta \cdot \nabla\Phi^*(\mathbf{z}_t)\rangle)dt + \langle \mathbf{x}_t - \mathbf{x}^\star, \sigma d\mathbf{B}_t\rangle,$$

$$dM^2_t = \frac{\sigma^2}{2}\mathrm{tr}(C_2(\mathbf{x}_t))dt + \sigma\langle \nabla_x L(\mathbf{x}_t, \boldsymbol{\lambda}_t), \nabla^2\Phi^*(\mathbf{z}_t)d\mathbf{B}_t\rangle.$$

Using the equilibrium points and the fact that $\nabla^2\Phi^*(z)\nabla^2\Phi(x) = I$, we obtain,

$$\begin{aligned} d(V^1_t + V^2_t) &\leq -\langle \mathbf{x}_t - \mathbf{x}^\star, \mathcal{L}(\mathbf{x}_t - \mathbf{x}^\star)\rangle dt + dM^1_t \\ &\leq -\frac{1}{\kappa_{\beta,N}}\|\mathcal{L}(\mathbf{x}_t - \mathbf{x}^\star)\|_2^2 + dM^1_t \end{aligned} \tag{4.6}$$

where for the first inequality we used the convexity of f and the symmetry of $\mathcal{L}$ and for the second inequality we used Lemma 2.6. We also have using L from (3.5),

$$\begin{aligned} dV^3_t &= \langle \nabla_{\mathbf{x}} L(\mathbf{x}_t, \boldsymbol{\lambda}_t), d\mathbf{x}_t\rangle + \frac{\sigma^2}{2}\mathrm{tr}(C_2(\mathbf{x}_t, \boldsymbol{\lambda}_t))dt + \langle \mathcal{L}(\mathbf{x}_t - \mathbf{x}^\star), d\boldsymbol{\lambda}_t\rangle \\ &= \left(-\|\nabla_x L(\mathbf{x}_t, \boldsymbol{\lambda}_t)\|^2_{\nabla^2\Phi^*(\mathbf{z}_t)} + \|\mathcal{L}\mathbf{x}_t\|_2^2\right)dt + dM^2_t \\ &\leq -\frac{2\kappa_g}{\hat{\mu}}\left(\sum_{i=1}^{N} D_{\Phi^*}(z^i_t, z^\star) + \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\|_2^2\right)dt + \|\mathcal{L}\mathbf{x}_t\|_2^2 dt + dM^2_t, \end{aligned}$$

where in the first line we used the optimality conditions $\nabla f(\mathbf{x}^\star) + \mathcal{L}\boldsymbol{\lambda}^\star$, $\mathcal{L}\mathbf{x}^\star = 0$, and to obtain the last inequality we used Lemma 2.8 with $\Psi = \frac{1}{2}\|\cdot\|_2^2$. If in addition $c \geq 2\kappa_{\beta,N}$ and using the bound in (4.6) we obtain,

$$\begin{aligned} dV_t &\leq -\frac{2\kappa_g}{\hat{\mu}}\left(\sum_{i=1}^{N} D_{\Phi^*}(z^i_t, z^\star) + \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star\|_2^2\right)dt - \|\mathcal{L}\mathbf{x}_t\|_2^2 dt + cdM^1_t + dM^2_t \\ &\leq -\frac{2\kappa_g}{c\hat{\mu} + 2\alpha(\Phi) + 3\kappa_N}V(\mathbf{x}_t, \boldsymbol{\lambda}_t)dt + cdM^1_t + dM^2_t, \end{aligned}$$

where in the last inequality we used (4.4) from Lemma 4.3. Finally, taking expectations, integrating and using Gronwall's lemma we obtain (4.5). ∎

It is clear that V is a stochastic Lyapunov function and exact convergence can be achieved using $\sigma = 0$. We note that if $\kappa_g \geq 0$ then the result above implies that $dV_t \leq 0$, but we may not have an exponential convergence rate. In Lemma 2.7 we showed that the strong convexity of the objective function implies that $\kappa_g > 0$ and we note that the reverse is not true.

4.3. Convergence with preconditioned interaction. The motivation behind the EPISMD algorithm in (3.4) is that in both unconstrained and constrained settings, additional speedup can be obtained by *preconditioning* the dual variable $\boldsymbol{\lambda}$. The use of the mirror map Ψ results in additional flexibility in the convergence rate; furthermore it allows to work with the Bregman divergence as the Lyapunov function. We observe this additional flexibility through the term $\hat{\mu}$, which is given by $\hat{\mu} = \min(\mu_\Phi, \mu_\Psi)$ so that the proper choice of mirror map Ψ can additionally improve the convergence.

Consider V_t as in (4.3) but with,

$$V_t^2(\boldsymbol{\lambda}_t) = \sum_{i=1}^{N} D_\Psi(\lambda^\star, \lambda_t^i). \tag{4.7}$$

As before we will change the definition of the optimal point the algorithm will converge to as follows,

$$(x^\star, \lambda^\star) = \arg \min_{x,\lambda \in (X^\star, \Lambda^\star)} D_\Phi(x, x_0) + D_\psi(\lambda, \lambda_0).$$

The convergence of (3.4) can be obtained using slight modifications of the proof of Lemma 4.3.

LEMMA 4.5. *Let Assumptions 1-3 hold. Then V_t with V_t^2 as in (4.7) satisfies Lemma 4.3 (**i**) and,*

ii.a *If $c \geq \max\{\kappa_N/\mu_\Phi, \kappa_N/\mu_\Psi\}$,*

$$V(\mathbf{x}, \boldsymbol{\lambda}) \geq \frac{1}{2}(\mu_\Phi c - \kappa_N)||\mathbf{x} - \mathbf{x}^\star||^2 + \frac{1}{2}(\mu_\Psi c - \kappa_N)||\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star||_2^2 \geq 0.$$

iii.a *Let $\hat{\mu} = \min\{\mu_\Phi, \mu_\Psi\}$. Then,*

$$V(\mathbf{x}, \boldsymbol{\lambda}) \leq \left(c + \frac{3\kappa_N + 2\alpha(\Phi)}{\hat{\mu}}\right)\left(\sum_{i=1}^{N} D_\Phi(x^\star, x^i) + \sum_{i=1}^{N} D_\Psi(\lambda^\star, \lambda^i)\right). \tag{4.8}$$

If in addition f is μ_f-strongly convex relative to Φ then:

ii.b *For any $c \geq \max\{((\kappa_N - \mu_f L_\Phi)/\mu_\Phi, \kappa_N/\mu_\Psi\}$,*

$$V(\mathbf{x}, \boldsymbol{\lambda}) \geq \frac{1}{2}(\mu_\Phi c + \mu_f L_\Phi - \kappa_N)||\mathbf{x} - \mathbf{x}^\star||^2 + \frac{1}{2}(\mu_\Psi c - \kappa_N)||\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star||_2^2 \geq 0.$$

The following convergence then holds.

PROPOSITION 4.6 (Convergence of the preconditioned dynamics in (3.4)). *Let Assumptions 1-3 hold and $\kappa_g > 0$. Consider the dynamics in (3.4). Let the Lyapunov function V_t be defined as in (4.3) with V_t^2 as in (4.7). Then the result from Proposition 4.4 holds with $\hat{\mu} = \min\{\mu_\Phi, \mu_\Psi\}$ and $c \geq 2\kappa_{\beta,N}/\mu_\Psi$.*

Proof. Let c be as in Lemma 4.5 to ensure non-negativity of the Lyapunov function. We follow similar steps as in the proof of Proposition 4.4. Observe that, assuming that the Bregman divergence is differentiable in the second variable and using (3.8),

$$\begin{aligned} dV_t^2 &= \langle -\nabla^2\Psi(\boldsymbol{\lambda}_t)(\boldsymbol{\lambda}^\star - \boldsymbol{\lambda}_t), \nabla^2\Psi(\boldsymbol{\lambda}_t)^{-1}\mathcal{L}\mathbf{x}_t\rangle \\ &= (\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star)^T\mathcal{L}(\mathbf{x}_t - \mathbf{x}^\star)dt. \end{aligned}$$

Then, a modification of Lemma 2.6 can be derived; using the positive semi-definiteness of $\nabla^2\Psi$ we can derive $\frac{\mu_\Psi}{\kappa_{\beta,N}}\nabla^2\Psi(\boldsymbol{\lambda})^{-1} \preceq \mathcal{L}_\beta^{-1}$ and obtain,

$$\langle \mathbf{x}_t, \mathcal{L}\mathbf{x}_t\rangle \geq \frac{\mu_\Psi}{\kappa_{\beta,N}}\langle \mathcal{L}\mathbf{x}_t, \nabla^2\Psi(\boldsymbol{\lambda}_t)^{-1}\mathcal{L}\mathbf{x}_t\rangle.$$

Then,

$$d(cV_t^1 + cV_t^2) \leq -\frac{c\mu_\Psi}{\kappa_{\beta,N}}||\mathcal{L}\mathbf{x}_t||^2_{\nabla^2\Psi(\boldsymbol{\lambda}_t)^{-1}}dt + dM_t^1. \tag{4.9}$$

Furthermore using (3.8),

$$dV_t^3 = \left(-||\nabla_{\mathbf{x}}L(\mathbf{x}_t, \boldsymbol{\lambda}_t)||^2_{\nabla^2\Phi^*(\mathbf{z}_t)} + ||\mathcal{L}\mathbf{x}_t||_{\nabla^2\Psi(\boldsymbol{\lambda}_t)^{-1}}\right)dt + dM_t^2$$

Consequently apply Lemma 2.8, the bound in (4.9) and use the additional assumption that $c \geq 2\kappa_{\beta,N}/\mu_\Psi$ to obtain

$$\begin{aligned} dV_t \leq &-\frac{2\kappa_g}{\hat{\mu}}\left(\sum_{i=1}^N D_{\Phi^*}(z_t^i, z^\star) + \sum_{i=1}^N D_\Psi(\lambda_t^i, \lambda^\star)\right)dt - \|\mathcal{L}\mathbf{x}_t\|^2_{\nabla^2\Psi(\boldsymbol{\lambda})^{-1}}dt \\ &+ cdM_t^1 + dM_t^2. \end{aligned}$$

Lastly use (4.8) and we complete the proof by integrating and taking expected values.□

The convergence of $\boldsymbol{\lambda}_t$ is affected by how well-conditioned the interaction graph and objective function are. In the proof of Proposition 4.6 we observe that the use of the mirror map Ψ results in an improved rate of convergence. In particular, when $\frac{\mu_\Phi}{\mu_\Psi} < 1$ we end up with r in Proposition 4.6 being larger than that seen in Proposition 4.4.

Choosing the preconditioner. For many interesting applications good mirror maps are known, and the advantages of mirror descent over gradient descent are well understood. Unfortunately, it is not clear how to select a good mirror map for the space of Lagrange multipliers. However, if we restrict the mirror map to be quadratic, we postulate that a positive definite approximation of the Hessian of the dual function will work well in practice, e.g. as seen in the results of Section 5. We argue that the extra computation associated with approximating the Hessian of the dual function could be justified in the scenario where the Laplacian matrix is ill-conditioned. Below we briefly outline the derivation of the Hessian for the dual function in the deterministic setting and when f is strongly convex. In particular, the (negative) dual function $q(\boldsymbol{\lambda})$ is defined as follows,

$$q(\boldsymbol{\lambda}) = \max_{\mathbf{x}\in\mathcal{X}^N}\left\{-(f(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathcal{L}\mathbf{x}\rangle)\right\}, \tag{4.10}$$

and let $\mathbf{x}(\boldsymbol{\lambda})$ be a maximizer of (4.10). We know that for the strongly convex case, $\mathbf{x}(\boldsymbol{\lambda})$ is unique and the gradient of the dual function is given by ([3, Theorem 6.3.3]),

$$\nabla_{\boldsymbol{\lambda}}q(\boldsymbol{\lambda}) = -\mathcal{L}\mathbf{x}(\boldsymbol{\lambda}).$$

Applying the KKT conditions to (4.10) and differentiating with respect to λ we also have that,

$$\nabla^2 f(\mathbf{x}(\boldsymbol{\lambda}))\frac{d\mathbf{x}(\boldsymbol{\lambda})}{d\boldsymbol{\lambda}} + \mathcal{L} = 0.$$

Using the preceding equation we obtain the following expression for the Hessian of the dual function,

$$\nabla^2 q(\boldsymbol{\lambda}) = -\mathcal{L}\frac{d\mathbf{x}(\boldsymbol{\lambda})}{d\boldsymbol{\lambda}} = \mathcal{L}\nabla^2 f(\mathbf{x}(\boldsymbol{\lambda}))^{-1}\mathcal{L}.$$

Unfortunately, even if f is strongly convex, the dual function in (4.10) is only convex (and not strongly convex). So we cannot directly use the Hessian of the dual function above, instead we proposed to use,

$$\nabla^2 \Psi(\boldsymbol{\lambda}) = \mathcal{L}_\beta \nabla^2 f(\mathbf{x}(\boldsymbol{\lambda}))^{-1}\mathcal{L}_\beta,$$

where $\mathcal{L}_\beta$ is the β-regularized Laplacian defined in (2.4). Note that we can avoid inverting the Hessian of f at every iteration and instead work with the convex conjugate of Ψ which occurs a one-off cost of diagonalizing the regularized Laplacian,

$$\nabla^2 \Psi^*(\boldsymbol{\lambda}) = \mathcal{L}_\beta^{-1} \nabla^2 f(\mathbf{x}(\boldsymbol{\lambda}))\mathcal{L}_\beta^{-1}.$$

The argument above could be made more precise and extended to more general settings. We report promising numerical experiments with this choice for the mirror map in Section 5.5.

5. Numerical results. Here we study the performance of the proposed algorithms in three problems:

- A **An unconstrained ill conditioned linear system**. We set the local cost functions as $f_i(x) = \frac{1}{2}||Q_i x - b_i||_2^2$. Unless mentioned otherwise $\mathcal{X} = \mathbb{R}^d$ with $d = 200$ and $Q_i \in \mathbb{R}^{20\times 200}$ is a random matrix with condition number 15 and $b_i \sim \mathcal{N}(0, I_{20})$. For the mirror map we set $\Phi(x) = \frac{1}{2}||x||^2$. We remark that in this setting the only *constraint* in the problem comes from the consensus.
- B **A constrained linear system.** The setup is similar to the unconstrained linear system but we set $\mathcal{X} = \Delta_d$, the d-dimensional simplex. The mirror map is set to be the negative entropy function $\Phi(x) = \sum_{j=1}^d [x]_j \log([x]_j)$ such that $[z]_i = 1 + \log([x]_j)$ (and $[z]_i = 0$ if $[x]_j = 0$) and $[x]_i = e^{[z]_i - 1}$ and the mapping onto the simplex is done using the normalization of this negative entropy mirror map. Here we denote with $[x]_i$ the i-th element of vector x.
- C **A neural network.** Motivated by federated learning applications [25] we consider the training of a neural network with one hidden layer and 30 nodes per layer using a ReLU activation, a softmax output and the cross-entropy loss. The training data is the FashionMNIST data. Each particle i has access to 10 of these samples and the cross-entropy loss over this subset defines each f_i. We will assume the solution lies in the constraint set $\mathcal{X} = \Delta_d$ (e.g. looking for a sparse solution) and use the negative entropy mirror map.

For the connectivity graph we define three options: **(i)** a cyclic graph with each node connected to the previous and next node, **(ii)** an Erdos-Rényi graph, or **(iii)** a barbell graph. For the dynamics we use a standard Euler discretization with $\Delta t = 0.01$ and 50,000 epochs. We implement ISMD, EISMD and EPISMD dynamics that include

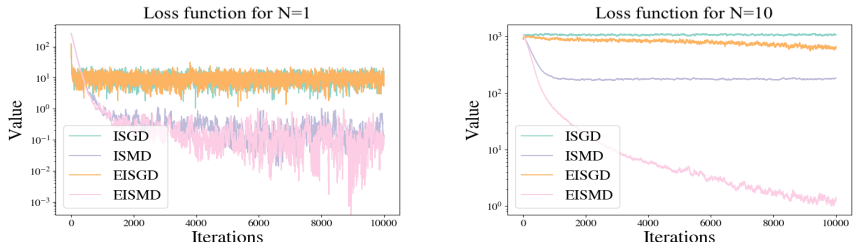


Fig. 3: Loss function for problem (B) with cyclic graph (i). Left is centralized ($N = 1$), right distributed ($N = 10$). ISGD, EISGD and ISMD, EISMD use (3.2), (3.3) with $\Phi(x) = \frac{1}{2}||x||^2$ (i.e. $\mathbf{x}_t = \mathbf{z}_t$) and the entropic mirror map, resp. projected to Δ_d.

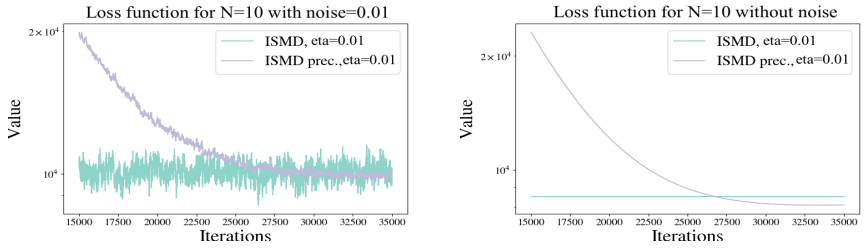


Fig. 4: Loss functions for Problem (A) and graph (i) with condition number set to 100, $N = 10$ and preconditioned with a local objective function for different learning rates (eta in legend). Left $\sigma = 0.01$, right $\sigma = 0$.

hyperparameters ϵ, η for the interaction strength and learning rate; see Appendix B.1 for more details.

5.1. The effect of the mirror map. We first compare distributed mirror descent with projected gradient descent for problem (B) using graph (i) to showcase the benefits of the entropic mirror map in simplex constrained systems. In Figure 3 we present the results for ISMD and EISMD for $N = 1$ (centralized implementation) and $N = 10$ (distributed). The benefits of the exact dynamics in (3.3) and the mirror map are clear in both cases.

5.2. The limits of the mirror map for ISMD. As shown in Section 3.1, the mirror map is equivalent to a preconditioner in the primal dynamics. In problem (A) with $\mathcal{X} = \mathbb{R}^d$, if the mirror map is chosen to be the local objective function then ISMD results in a local Newton-like algorithm. Here we will explore the effectiveness of using such a local preconditioner and set $\nabla^2\Phi(x^i) = Q_i^T Q_i$. Figure 4 shows the convergence results for the problem A with condition number increased to 100, $d = 20$ and the cyclic interaction graph (i). The distributed case when preconditioned with the local Hessian does not result in full convergence. Thus the mirror map alone cannot facilitate convergence. The results are meant as motivation for using EISMD. We note that other solutions are possible, e.g. extending [34] to provide an approximation of the full preconditioner, but this will increase communication at each step.

5.3. The effects of the exact algorithm. Here we present a detailed analysis of the benefits of EISMD compared to ISMD using both the ℓ_2 error between the

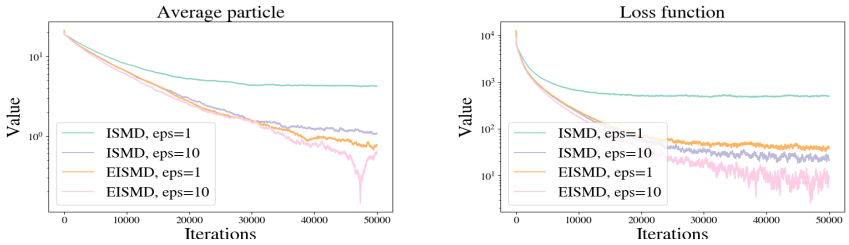


Fig. 5: ISMD and EISMD for Problem (A) and graph (i) for $\eta = 0.01$, $N = 10$, $\sigma = 0.1$ and $\epsilon = 1, 10$ (eps in legend). Left panel shows $||x_t^b - x_t^w||^2$ with b, w the best and worst particle index at each time and right panel shows cost functions.

best and worst performing particle and the loss function computed over all particles. Figure 5 shows results for problem (A) and graph (i). In all cases a higher interaction strength allows to converge closer to the optimum and EISMD performs significantly better then ISMD. Figure 6 shows similar results for problem (B). In the top panels EISMD is clearly more effective than ISMD. The lower plots of Figure 6 show that in the presence of noise a high interaction in ISMD is able to mitigate the process variance due to the (appearing as oscillations) and result in convergence closer to the optimum.

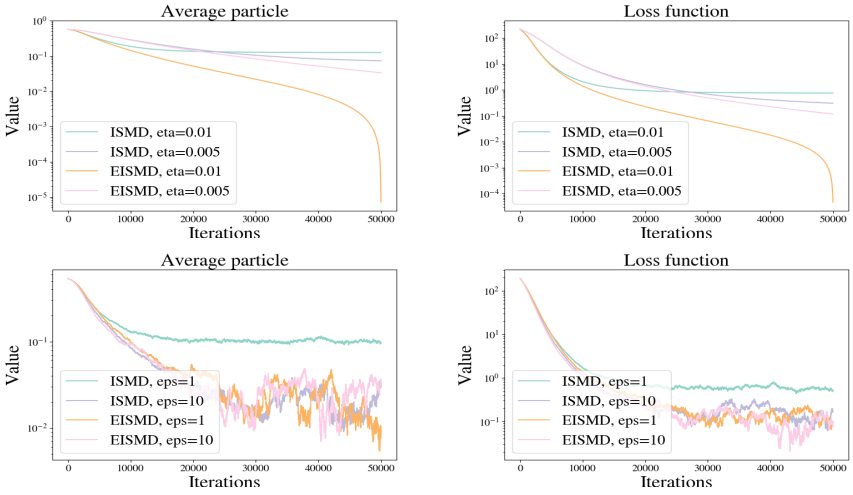


Fig. 6: ISMD and EISMD for Problem (B) and graph (i). Bottom row details as in Figure 5. Top row panels same but with $\sigma = 0$, $\epsilon = 1$ and $\eta = 0.01, 0.005$.

5.4. The choice of interaction graph. Here we set the communication structure to be an Erdos-Rényi communication graph (ii). This is a random graph where each edge is chosen with a certain probability p. The communication between the nodes is thus determined by this connectivity probability with on average each node being connected to $p \times N$ other nodes. The weight matrix A is set to be a doubly stochastic version of this communication graph and additionally we use [7] to optimize

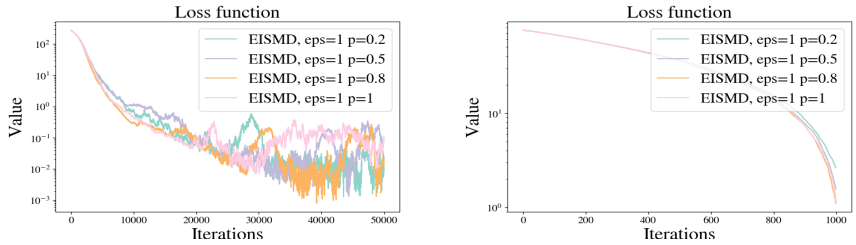


Fig. 7: Loss functions of EISMD for problem (B) and graph (ii). Left panel contains full results and right magnifies the initial epochs. We set $\epsilon = 1$, $\sigma = 0.1$, $\eta = 0.01$ and vary the number of nodes each particle communicates with to 2, 5, 8 or 10.

the coefficients to get the fastest mixing in (3.3) whilst maintaining minimum communications. Figure 7 shows results for problem (B) and different graph connectivities. The benefits of EISMD become even more clear from these results; using a graph with an optimal interaction structure while maintaining minimal interaction costs in combination with EISMD can result in very fast and computationally efficient convergence. Even with a very low communication cost per round, specifically each node communicating with approximately two other nodes per round, it results in a convergence speed comparable to full communication seen earlier.

5.5. Accelerated dual convergence. The right choice of mirror map (preconditioner) in the unconstrained setting can also result in accelerated convergence of the $\boldsymbol{\lambda}$ variable. We know from the discussion in Section 4.3 the form of a good preconditioner. For the linear case, it is given by $\nabla^2\Psi(\boldsymbol{\lambda}) = -\mathcal{L}_\beta(QQ^T)^{-1}\mathcal{L}_\beta$ with Q being block diagonal composition of Q_i-s. We use the regularized Laplacian here, specifically $\mathcal{L}_\beta = \mathcal{L} + 0.01 \cdot \mathbf{1}_N\mathbf{1}_N^T \otimes I_d$. We show the benefits of this preconditioner in EPISMD (shown in (3.4)) numerically in Figure 8 for a barbell graph with two clusters (iii). Clearly, preconditioning allows to converge much *faster and closer* to the optimum. We observe that preconditioning enables the algorithm to converge much faster and closer to the optimum. In the case of the barbell graph (iii) where the graph Laplacian is not well-conditioned, the ability to speed up the convergence using the preconditioning of the dual variable can lead to very substantial benefits in many real-world clustered systems.

5.6. Federated learning in a nonconvex example. In this example we study the performance of ISMD and EISMD for problem (C) and graph (i). In Figure 9 we compute the gradient over a random batch of data points. Each particle computes the gradient in each iteration over a batch of size 10, which results in implicit noise in the algorithm. We observe that the exact method results converge significantly faster even in this nonconvex setting.

6. Discussion. Our work successfully presented the benefits of mirror maps and how to use them while still achieving consensus in a distributed setting. For future work several interesting extensions can be considered. If one discretizes the continuous time dynamics, e.g. as in constant step size Euler-Maruyama schemes, then a bias will be present that depends on the discretization interval. In this sense exactness is lost, but could be recovered again using decreasing step sizes or constant ones combined with ideas from sampling methods [14]. Addressing this in detail and comparing with

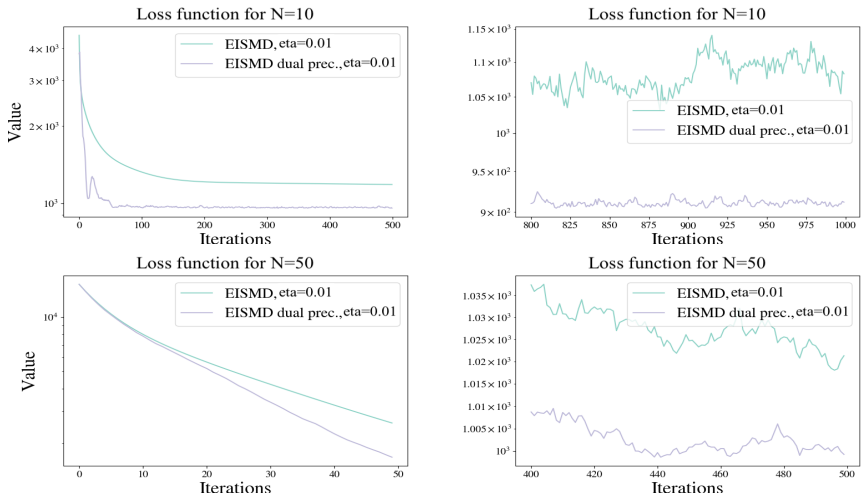


Fig. 8: Loss functions comparing EPISMD and EISMD for Problem A and graph (iii). We use $\sigma = 0.01$ and top row panels use a barbell graph with two clusters and 10 nodes and bottom panels two clusters and 50 nodes.

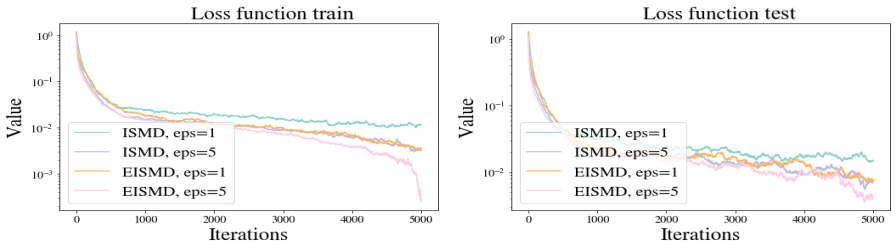


Fig. 9: ISMD and EISMD for problem (C) and graph (i). Left shows training loss against epoch and right shows test loss.

the discrete time methods in Table 1 is left for further work. The analysis shown here for EISMD and EPISMD can be extended to a setting where the solution of (1.1) does not lie in the constraint space $\mathcal{X}$ using steps similar to [26]. Furthermore, we only numerically explored the nonconvex setting in EISMD using a small neural network with a negative entropy mirror map; future work can include a theoretical analysis of the nonconvex case as well as exploring further the benefits of different mirror maps using EPISMD in the distributed training of neural networks. Last, the algorithm could be modified to one applying the Laplacian onto the mirrored variables; preliminary numerical results showed that in high-dimensional and/or nonconvex settings this could be of benefit; a more detailed converge analysis of this phenomenon may be of interest which would require the definition of a different Lyapunov function.

Acknowledgements. This project was funded by JPMorgan Chase & Co under J.P. Morgan A.I. Research Awards in 2019 and 2021. G.A.P. was partially supported by the EPSRC through grant number EP/P031587/1.

REFERENCES

[1] Y. Arjevani and O. Shamir, *Communication complexity of distributed convex learning and optimization*, arXiv preprint arXiv:1506.01900, (2015).

[2] M. Badiei Khuzani and N. Li, *Stochastic Primal-Dual Method on Riemannian Manifolds with Bounded Sectional Curvature*, arXiv e-prints, (2017), pp. arXiv–1703.

[3] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*, John Wiley & Sons, 2013.

[4] A. Beck, *First-order methods in optimization*, vol. 25, SIAM, 2017.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*, 2015.

[6] A. Borovykh, P. Parpas, N. Kantas, and G. Pavliotis, *On stochastic mirror descent with interacting particles: convergence properties and variance reduction*, Physica D.Nonlinear Phenomena 418, (2021).

[7] S. Boyd, P. Diaconis, and L. Xiao, *Fastest mixing Markov Chain on a graph*, SIAM review, 46 (2004), pp. 667–689.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), p. 1–122.

[9] S. Bubeck, *Convex optimization: Algorithms and complexity*, arXiv preprint arXiv:1405.4980, (2014).

[10] F. Bullo, *Lectures on Network Systems*, Kindle Direct Publishing, 1.5 ed., 2021.

[11] S. Chen, A. Garcia, M. Hong, and S. Shahrampour, *Decentralized Riemannian Gradient Descent on the Stiefel Manifold*, arXiv preprint arXiv:2102.07091, (2021).

[12] R. D'Orazio, N. Loizou, I. Laradji, and I. Mitliagkas, *Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize*, arXiv preprint arXiv:2110.15412, (2021).

[13] J. C. Duchi, A. Agarwal, and M. J. Wainwright, *Dual averaging for distributed optimization: Convergence analysis and network scaling*, IEEE Transactions on Automatic control, 57 (2011), pp. 592–606.

[14] A. Durmus and E. Moulines, *Sampling from strongly log-concave distributions with the unadjusted Langevin algorithm*, arXiv preprint arXiv:1605.01559, 5 (2016).

[15] G. Franca, D. Robinson, and R. Vidal, *ADMM and accelerated ADMM as continuous dynamical systems*, in International Conference on Machine Learning, PMLR, 2018, pp. 1559–1567.

[16] B. Gharesifard and J. Cortés, *Distributed continuous-time convex optimization on weight-balanced digraphs*, IEEE Transactions on Automatic Control, 59 (2013), pp. 781–786.

[17] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach, and L. Massoulie, *Statistically preconditioned accelerated gradient method for distributed optimization*, in International Conference on Machine Learning, PMLR, 2020, pp. 4203–4227.

[18] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge university press, 1994.

[19] D. Jakovetić, J. Xavier, and J. M. Moura, *Fast distributed gradient methods*, IEEE Transactions on Automatic Control, 59 (2014), pp. 1131–1146.

[20] C. Li, C. Chen, D. Carlson, and L. Carin, *Preconditioned stochastic gradient Langevin dynamics for deep neural networks*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.

[21] S. Liang, L. Wang, and G. Yin, *Exponential convergence of distributed primal–dual convex optimization algorithm without strong convexity*, Automatica, 105 (2019), pp. 298–306.

[22] P. Lin, W. Ren, and J. A. Farrell, *Distributed continuous-time optimization: nonuniform gradient gains, finite-time convergence, and convex constraint set*, IEEE Transactions on Automatic Control, 62 (2016), pp. 2239–2253.

[23] S. Liu, Z. Qiu, and L. Xie, *Continuous-time distributed convex optimization with set constraints*, IFAC Proceedings Volumes, 47 (2014), pp. 9762–9767.

[24] H. Lu, R. M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM Journal on Optimization, 28 (2018), pp. 333–354.

[25] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, *Communication-efficient learning of deep networks from decentralized data*, in Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

[26] P. Mertikopoulos and M. Staudigl, *On the convergence of gradient-like flows with noisy gradient input*, SIAM Journal on Optimization, 28 (2018), pp. 163–197.

[27] A. Nedić, S. Lee, and M. Raginsky, *Decentralized online optimization with global objectives and local communication*, in 2015 American Control Conference (ACC), IEEE, 2015,

pp. 4497–4503.
[28] A. Nedic, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization, 27 (2017), pp. 2597–2633.
[29] S. Pu and A. Nedić, *Distributed stochastic gradient tracking methods*, Mathematical Programming, 187 (2021), pp. 409–457.
[30] G. Qu and N. Li, *Harnessing smoothness to accelerate distributed optimization*, IEEE Transactions on Control of Network Systems, 5 (2017), pp. 1245–1260.
[31] M. Raginsky and J. Bouvrie, *Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence*, in 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), IEEE, 2012, pp. 6793–6800.
[32] S. S. Ram, A. Nedić, and V. V. Veeravalli, *Distributed stochastic subgradient projection algorithms for convex optimization*, Journal of optimization theory and applications, 147 (2010), pp. 516–545.
[33] S. Shahrampour and A. Jadbabaie, *Distributed online optimization in dynamic environments using mirror descent*, IEEE Transactions on Automatic Control, 63 (2017), pp. 714–725.
[34] O. Shamir, N. Srebro, and T. Zhang, *Communication-efficient distributed optimization using an approximate newton-type method*, in International conference on machine learning, PMLR, 2014, pp. 1000–1008.
[35] G. Shi, A. Proutiere, and K. H. Johansson, *Network synchronization with convexity*, SIAM Journal on Control and Optimization, 53 (2015), pp. 3562–3583.
[36] W. Shi, Q. Ling, G. Wu, and W. Yin, *EXTRA: An exact first-order algorithm for decentralized consensus optimization*, SIAM Journal on Optimization, 25 (2015), pp. 944–966.
[37] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for machine learning*, Mit Press, 2012.
[38] Y. Sun and S. Shahrampour, *Distributed mirror descent with integral feedback: Asymptotic convergence analysis of continuous-time dynamics*, arXiv preprint arXiv:2009.06747, (2020).
[39] H. Wang and A. Banerjee, *Bregman alternating direction method of multipliers*, Advances in Neural Information Processing Systems, (2014).
[40] J. Wang and N. Elia, *A control perspective for centralized and distributed convex optimization*, in 2011 50th IEEE conference on decision and control and European control conference, IEEE, 2011, pp. 3800–3805.
[41] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, proceedings of the National Academy of Sciences, 113 (2016), pp. E7351–E7358.
[42] K. Yuan, Q. Ling, and W. Yin, *On the convergence of decentralized gradient descent*, SIAM Journal on Optimization, 26 (2016), pp. 1835–1854.
[43] X. Zeng, P. Yi, and Y. Hong, *Distributed continuous-time algorithm for constrained convex optimizations via nonsmooth analysis approach*, IEEE Transactions on Automatic Control, 62 (2016), pp. 5227–5233.

Appendix A. Auxiliary results.

A.1. The benefits of the mirror map. To show the ability of mirror descent to adapt to a particular geometry we present the classical mirror descent proof which can be interpreted as having a central server; this server broadcasts $x_t \in \mathcal{X}$ to all nodes which compute $\nabla f_i(x_t)$ and send this back to the server. In continuous-time the algorithm at the server is given by

$$dz_t = -\sum_{i=1}^{N} \nabla f_i(x_t)dt + \sigma dB_t.$$

The aim is to converge to the optimal point (x^*, z^*) given as

$$x^* = \arg\min_{x\in\mathcal{X}} f(x), \quad z^* = \nabla\Phi(x^*).$$

We will make use of the inequality,

$$\int_0^t e^{-\alpha(t-s)} K ds \leq \frac{K}{\alpha}, \tag{A.1}$$

and $||\Delta\Phi^*(z_t)||_\infty \leq \infty$ (by Assumption 3).

LEMMA A.1 (Convergence of centralized mirror descent). *Let $f(x) = \sum_{i=1}^N f_i(x)$ be μ_f-strongly convex with respect to Φ. Then,*

$$\mathbb{E}\left[D_\Phi(x^*, x_t)\right] \leq e^{-\mu_f t} D_\Phi(x^*, x_0) + \frac{\sigma^2}{2\mu_f}||\Delta\Phi^*||_\infty.$$

Proof. Let $V_t = D_{\Phi^*}(z_t, z^*)$. We have through Itô's lemma,

$$\begin{aligned} dV_t &= -\sum_{i=1}^N (x^* - x_t)^T \nabla f_i(x_t) dt + \frac{1}{2}\sigma^2 \mathrm{tr}(\Delta\Phi^*(z_t)) dt + \sigma(x_t - x^*)^T dB_t \\ &\leq (-\mu_f D_\Phi(x^*, x_t) - f(x_t) + f(x^*) + \frac{\sigma^2}{2}||\Delta\Phi^*||_\infty) dt + \sigma(x_t - x^*)^T dB_t \\ &\leq (-\mu_f D_{\Phi^*}(z_t, z^*) + \frac{\sigma^2}{2}||\Delta\Phi^*||_\infty) dt + \sigma(x_t - x^*)^T dB_t, \end{aligned}$$

where in the first inequality we have used the μ_f-strong convexity w.r.t. Φ and in the second inequality that by the properties of the mirror map $D_\Phi(x^*, x_t) = D_{\Phi^*}(z_t, z^*)$ and that $f(x^*) - f(x_t) \leq 0$. Using (A.1), taking expectations and applying Grönwall inequality the result follows. ∎

Comparing the results of mirror (preconditioned) gradient descent to the Euclidean setting where $\Phi_E(x) = \frac{1}{2}||x||_2^2$, the effectiveness of the algorithm hinges on how large μ_f is compared to μ_E where μ_E is such that f is μ_E-strongly convex with respect to $\frac{1}{2}||x||_2^2$. An optimal mirror map is then the one that is given by $\Phi(x) = f(x)$, in which case it holds that $\mu_f = 1$. Discretizing the continuous-time algorithm would then result in achieving convergence with a larger discretization step size. The works of [17, 34, 1] study the discrete-time convergence if the mirror map is chosen to be one of the objective functions, i.e. $\Phi(x) = f_i(x)$ for some i and show that convergence can be obtained in one time step with the right (estimate of the) preconditioner. One can seek to choose the mirror map Φ in such a way that $D_\Phi(x_0, x^*)$ is smaller than $D_{\Phi_E}(x_0, x^*)$ and get faster convergence, and in such a way that $||\Delta\Phi^*||_\infty$ is smaller than $||\Delta\Phi_E^*||_\infty$ for closer convergence.

A.2. Approximate convergence of distributed mirror descent. Recall the vectorized ISMD dynamics of (3.1) are

$$d\mathbf{z}_t = (-\eta\nabla f(\mathbf{x}_t) - \epsilon\mathcal{L}\mathbf{z}_t)\, dt + \sigma d\mathbf{B}_t, \quad \mathbf{x}_t = \nabla\Phi^*(\mathbf{z}_t). \tag{A.2}$$

If $x^\star$ as per Lemma 4.1 does not exist, even in the deterministic case exact consensus and optimality at convergence can no longer be achieved. We present a result which shows that exponential convergence holds only up to a certain neighborhood of $(\mathbf{x}^\dagger, \mathbf{z}^\dagger)$ both minimizing $f(\mathbf{x}) + \frac{1}{2}\mathbf{z}^T\mathcal{L}\mathbf{z}$, the size of this neighborhood depending on the noise and the distance between the $f(x^\dagger)$ and $f(x^\star)$.

PROPOSITION A.2 (Approximate convergence of (A.2)). *Let Assumptions 1-3 hold and assume that f is μ_f-strongly convex w.r.t. Φ. Let $V_t = \frac{1}{N}\sum_{i=1}^N D_{\Phi^*}(z_t^i, z^\dagger)$, where z_t^i obeys the dynamics of (3.1) (or (A.2)). Then we have:*

$$\mathbb{E}\left[V_t\right] \leq e^{-\eta\mu_f t}\frac{1}{N}\sum_{i=1}^N D_{\Phi^*}(z_0^i, z^\dagger) + \frac{\sigma^2}{2\eta\mu_f}||\Delta\Phi^*||_\infty + \frac{1}{\mu_f}\left(f(\mathbf{x}^\dagger) - f(\mathbf{x}^\circ)\right),$$

where $x^\circ = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$.

Proof. Denote $\mathbf{x}^\dagger = \arg\inf\{f(\mathbf{x}) + \frac{1}{2}\nabla\Phi(\mathbf{x})^T\mathcal{L}\nabla\Phi(\mathbf{x})\}$, $\mathbf{z}^\dagger = \nabla\Phi(\mathbf{x}^\dagger)$. Then

$$dV_t = -\frac{1}{N}\sum_{i=1}^N (x_t^i - x^\dagger)^T\eta\nabla f_i(x_t^i)dt + \epsilon\frac{1}{N}\sum_{i=1}^N (x_t^i - x^\dagger)^T\sum_{j=1}^N A_{ij}(z_t^j - z_t^i)dt$$

$$+\frac{1}{2}\sigma^2\frac{1}{N}\sum_{i=1}^N \mathrm{tr}(\Delta\Phi^*(z_t^i))dt + \frac{1}{N}\sum_{i=1}^N (x_t^i - x^\dagger)^T dB_t^i. \tag{A.3}$$

Using the μ_f-strong convexity of f as in the proof of Lemma A.1 we obtain,

$$\begin{aligned}\nabla f(\mathbf{x}_t)^T(\mathbf{x}^\dagger - \mathbf{x}_t) \leq& f(\mathbf{x}^\dagger) - f(\mathbf{x}_t) - \mu_f D_\phi(\mathbf{x}^\dagger, \mathbf{x}_t)\\ \leq& f(\mathbf{x}^\dagger) - f(\mathbf{x}^\circ) + f(\mathbf{x}^\circ) - f(\mathbf{x}_t) - \mu_f D_\phi(\mathbf{x}^\dagger, \mathbf{x}_t)\\ \leq& f(\mathbf{x}^\dagger) - f(\mathbf{x}^\circ) - \mu_f D_\phi(\mathbf{x}^\dagger, \mathbf{x}_t)\end{aligned}$$

Substituting in (A.3) and using (2.7), (4.2) and taking expectations gives

$$\frac{d\mathbb{E}[V_t]}{dt} \leq -\eta\mu_f\mathbb{E}[V_t] + \eta\left(f(\mathbf{x}^\dagger) - f(\mathbf{x}^\circ)\right) + \frac{1}{2}\sigma^2||\Delta\Phi^*||_\infty.$$

Standard Grönwall arguments gives the result. ∎

Note that when (4.1) $\mathbf{x}^\circ$, $\mathbf{x}^\star$ and $\mathbf{x}^\star$ coincide [35, Lemma 7], so $C_f = 0$.

Appendix B. Additional information for the numerical results.

B.1. Dynamics with hyperparameters and their discrezation. In (A.2) and (3.1) we included η, ϵ to allow tuning of the relative effect of the gradient and the interaction. The expression of Proposition A.2 can be sharpened to include more precise contributions from $\epsilon\mathcal{L}\mathbf{z}_t$ using similar Grönwall arguments to bound $\mathbb{E}[||z_t - z^\dagger||^2]$ instead of simply using (4.2). This is omitted here for brevity and we note ϵ does have an effect in the numerical results. For the discretization, let Δt be a constant discretization interval and $B_k^{i,j} \sim \mathcal{N}(0, \sigma^2\Delta t)$ form i.i.d. sequences for $j = 1, 2$ and $i = 1, \ldots, N$. Then the Euler discretization of the dynamics in (A.2) are given by

$$\begin{aligned}z^i_{(k+1)\Delta t} &= z^i_{k\Delta t} - \eta\nabla f_i(x^i_{k\Delta t})\Delta t + \epsilon\textstyle\sum_{j=1}^N A_{ij}(z^j_{k\Delta t} - z^i_{k\Delta t})\Delta t + B_k^{i,1},\\ x^i_{(k+1)\Delta t} &= \nabla\Phi^*(z^i_{(k+1)\Delta t}).\end{aligned}$$

Similarly for the EISMD dynamics of (3.3) we can include a learning rate η and interaction strength ϵ

$$d\mathbf{z}_t = -\eta\nabla f(\mathbf{x}_t)dt - \epsilon\mathcal{L}\mathbf{z}_t dt - \mathcal{L}\boldsymbol{\lambda}_t dt + \sigma d\mathbf{B}_t, \quad d\boldsymbol{\lambda}_t = \mathcal{L}\mathbf{x}_t dt. \tag{B.1}$$

Similarly, the discretized dynamics of (B.1) are given by

$$\begin{aligned}z^i_{(k+1)\Delta t} =& z^i_{k\Delta t} - \Delta t\eta\nabla f_i(x^i_{k\Delta t}) + \Delta t\epsilon\textstyle\sum_{j=1}^N A_{ij}(z^j_{k\Delta t} - z^i_{k\Delta t})\\ &+ \Delta t\textstyle\sum_{j=1}^N A_{ij}(\lambda^j_{k\Delta t} - \lambda^i_{k\Delta t}) + B_k^{i,2},\\ \lambda^i_{(k+1)\Delta t} =& \lambda^i_{k\Delta t} - \textstyle\sum_{j=1}^N A_{ij}(x^j_{k\Delta t} - x^i_{k\Delta t})\Delta t\end{aligned}$$

The rest of the cases in Section 3 follow similarly.