To interact or not? The convergence properties of interacting stochastic mirror descent

Anastasia Borovykh [1] **Nikolas Kantas** [2] **Panos Parpas** [1] **Grigorios A. Pavliotis** [2]

Abstract

An open problem in optimization with imperfect information is the computation of an exact minimizer. Possible solutions include using a decreasing step size, an increasing batch size or variance reduction techniques. In this work we take a different approach and run replicas of mirror descent which we allow to interact. In particular, we study the convergence of interacting stochastic mirror descent and show that interaction can decrease the variance. Therefore, interaction is an alternative to decreasing the learning rate or to increasing the batch size. We show that in a convex, ill-conditioned problem interacting mirror descent results in faster and closer convergence to the optimum, and in a nonconvex setting interactions can help to escape from saddle points. In the distributed case we also discuss a variant of our algorithm that uses second-order information in order to encourage convergence to a solution that achieves both consensus and optimality.

1. Introduction

Optimization models that arise in artificial intelligence and statistical learning applications often include noisy estimates of the function and its gradient. In such a situation it is known that the optimization algorithm will converge to a neighborhood of the minimizer, the size of which is proportional to the noise variance and the stepsize used. Several methods have been proposed for controlling the noise in stochastic gradient optimization. A standard approach is to use a vanishing step size e.g. (Mertikopoulos & Staudigl, 2018), in which case the noise is decreased at the expense of slower convergence. Alternatively, decreasing the noise variance by increasing the batch size over time has been proposed e.g. (Byrd et al., 2012) but this requires an increase in computational cost. In addition to these approaches various variance reduction methods have been proposed e.g. (Johnson & Zhang, 2013), (Defazio et al., 2014), (Gorbunov et al., 2019), (Csiba & Richtárik, 2018).

Another option is to run independent replicas of the algorithm and average the results. We will refer to each of these runs as a particle. The question we address in this paper is whether it is beneficial to allow these particles *to interact* with each other. We study this question using the general framework of Stochastic Mirror Descent (SMD) (Nemirovsky & Yudin, 1983), an efficient method used to solve both constrained and unconstrained problems. We derive convergence rates and make explicit the tradeoff between communication and variance reduction in both a centralized and a distributed setting. We show that the variance can be reduced by using more particles and/or a sufficiently high interaction strength. The interacting algorithm converges faster and closer to the optimizer in ill-conditioned settings due to the noise reduction property. Furthermore, we show that interaction helps escape saddle points in nonconvex settings. Lastly, in a distributed setting we show that increasing the interaction strength allows for the algorithm to converge to consensus and we show how second-order information can be used to achieve convergence to consensus and optimality. Interestingly, the use of second-order information does not require noticeably more storage or CPU time compared to first-order methods.

The work most closely related to ours is that of (Raginsky & Bouvrie, 2012) in which the authors show that interaction leads to variance reduction. Our work also has parallels with the vast distributed optimization literature from which we list some indicative recent references: (Duchi et al., 2011), (Lin et al., 2016), (Shahrampour & Jadbabaie, 2017), (Koloskova et al., 2019), (De & Goldstein, 2016).

2. The setup

In this paper we will consider generic convex and nonconvex optimization problems of the form

$$\min_{x\in\mathcal{X}}\{f(x)\},$$

with $\mathcal{X} \subset \mathbb{R}^d$ be a closed convex constraint set. In the theoretical analysis we are interested in computing $x^* =$

[1] Department of Computing, Imperial College London, UK [2] Department of Mathematics, Imperial College London, UK. Correspondence to: Anastasia Borovykh <a.borovykh@imperial.ac.uk>.

Workshop 'Beyond first-order methods in ML systems', 37^{th} International Conference on Machine Learning, Vienna, Austria, 2020.

$\arg\min_{x\in\mathcal{X}}\{f(x)\}$ under the assumptions of smoothness and (μ-strong) convexity for f. Let $\Phi : \mathcal{X} \to \mathbb{R}^d$ be a μ-strongly convex function w.r.t. a norm $||\cdot||$ and assume it is continuously differentiable. We will refer to this as a *mirror map* used to convert the constrained optimization problem to an unconstrained one; see for more details e.g. Assumption 9.3 in (Beck, 2017). Let Φ^* be the conjugate of Φ. The *Bregman divergence* is defined as $D_\Phi(x,y) = \Phi(x) - \Phi(y) - \nabla\Phi^T(y)(x-y)$. The Bregman divergence is meant to quantify how far a point x is from y and Φ can be thought of as a distance generating function that adapts to the geometry or structure of $\mathcal{X}$. Let $D_{\Phi,\mathcal{X}} := \sup_{x\in\mathcal{X},x'\in\mathcal{X}^o}\sqrt{2D_\Phi(x,x')}$. Let κ be the strong convexity constant of $\mathcal{V}$, defined as $\nabla\mathcal{V} = \nabla f \circ \nabla\Phi^*$.

We are interested in investigating the performance and properties of the interacting stochastic mirror descent (ISMD) algorithm for estimating the minimizer x^*, based on the Itô stochastic differential equation,

$$dz_t^i = -\nabla f(x_t^i)dt + \theta\sum_{j=1}^N A_{ij}(z_t^j - z_t^i)dt + \sigma dB_t^i,$$
$$x_t^i = \nabla\Phi^*(z_t^i) := \arg\min_{x\in\mathcal{X}} D_\Phi(x, z_t^i),$$

where each particle $i = 1,...,N$ is driven by an independent Brownian motion B_t^i and Φ is the mirror map. The interesting feature here is that particles interact through $A = \{A_{ij}\}_{i,j=1}^N$, which is an $N\times N$ doubly-stochastic matrix representing the interaction weights. This interaction will attract particles towards each other. The matrix A represents an interaction graph which imposes communication constraints on the agents: each particle i can communicate directly only with its immediate neighbors, i.e. $j \in \{1,\dots,N\}$ for whom $A_{ij} \neq 0$. In the absence of A (i.e. when $A = 0$) the dynamics would correspond to parallel independent replicas of SMD. The parameter θ represents the *interaction strength*. As we will show, it plays a crucial role in mitigating the variance of the algorithm and in obtaining consensus. Without the constraints and with no noise the setting is that of gradient descent (GD); without constraints and with noise it is stochastic gradient descent (SGD). We define the graph Laplacian as $L := \text{Diag}(A\mathbf{1}_N) - A$, and let $\mathcal{L} := L \otimes I_d$, where $\otimes$ is the Kronecker product. We assume throughout that the network graph corresponding to the graph Laplacian L is connected, which in turn implies (Mesbahi & Egerstedt, 2010) that eigenvalues are nonnegative, $\lambda_0 = 0 < \underline{\lambda} \leq \lambda_2 \leq ... \leq \lambda_{dN}$.

3. Analyzing the convergence properties

We will decompose each particle position as a sum of the particle average and the fluctuation term, $z_t^i := \bar{z}_t^N + \tilde{z}_t^i$, where we let $\bar{z}_t^N := \frac{1}{N}\sum_{i=1}^N z_t^i$ and $\tilde{z}_t^i := z_t^i - \bar{z}_t^N$. Observe, $d\bar{z}_t^N = -\frac{1}{N}\sum_{i=1}^N \nabla f(x_t^i)dt + \frac{\sigma}{N}\sum_{i=1}^N dB_t^i$, using the double stochasticity of A.

3.1. A general bound

The following bound to optimality holds.

Proposition 1 (Convergence of ISMD[1]). *Assume that the function $f : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex function with respect to Φ. Then it holds,*

$$\int_0^T e^{\mu(t-T)}\mathbb{E}[(f(x_t^i) - f(x^*))]dt \leq \frac{1}{2}e^{-\mu T}D_{\Phi,\mathcal{X}}^2$$
$$+ \frac{\sigma^2}{2N\mu}||\Delta\Phi^*||_\infty + \int_0^T e^{\mu(t-T)}\frac{L}{\mu}\mathbb{E}\left[||\tilde{z}_t^i||_*\right]dt$$
$$+ \int_0^T e^{\mu(t-T)}\frac{2L+\mu^2}{\mu N}\sum_{i=1}^N\mathbb{E}\left[||\tilde{z}_t^i||_*\right]dt,$$

where L is the Lipschitz constant of f and $||\cdot||_*$ is the dual norm.

The deviation from the minimum is upper-bounded by four terms. The first two terms are the standard optimization errors, where we observe that the noise variance is reduced by a factor of N. The third and fourth terms are penalties incurred due to each of the particles having different values. These two terms measure the deviation of each individual particle from the particle average. There is thus a tradeoff between the interaction and the variance. If the interaction term is bounded and not increasing with N, the more particles, the smaller the distance to the optimum, as is witnessed by the term $\frac{\sigma^2}{2N}||\nabla\Phi^*||_\infty$.

3.2. Bounding the fluctuation

To complete the analysis on the distance to optimality of ISMD we show that the fluctuation term $\tilde{z}_t^i$ can be bounded. Since we are working in the dual space here, we will use the strong convexity constant κ of $\mathcal{V}$.

Proposition 2 (Bounding the fluctuation term[1]). *For a μ-strongly convex f, it holds,*

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N||\tilde{z}_t^i||_*^2\right] \leq \frac{K}{N}e^{-\theta(\kappa+\underline{\lambda})t}\sum_{i=1}^N||\tilde{z}_0^i||_2^2$$
$$+ \frac{dK}{\theta(\kappa+\underline{\lambda})}\sigma^2\frac{(N-1)}{N}\left(1 - e^{-\theta(\kappa+\underline{\lambda})t}\right).$$

From the above result we remark the following:

- in the setting with noise, for a strongly convex objective, approximate consensus can be achieved at sufficiently long times t and for a sufficiently large $\theta(\kappa+\underline{\lambda})$;

[1] For the proof we refer to (Borovykh et al., 2020).

- in the strongly convex case the fluctuation term converges even if there is no interaction if κ is sufficiently negative; for the convex case we have $\kappa = 0$ and consensus can be achieved only if the interaction strength θ is sufficiently high;
- for the no-noise setting, i.e. if $\sigma = 0$, the fluctuation term converges to zero as $t \to \infty$.

3.3. Noise reduction

ISMD improves the convergence since in the term $\frac{1}{N}\sigma||\Delta\Phi(z_t^i)||_\infty$ noise is reduced by a factor $1/N$. We remark however that this will only be achieved if the term $||\tilde{z}_t^i||_*$ is bounded and non-increasing with N. This fluctuation term is controlled by the strong convexity of the objective or by imposing a sufficiently high interaction strength. In other words, when $\theta(\kappa + \underline{\lambda})$ is sufficiently high the fluctuation term is sufficiently small. As long as the decrease in the value of $\frac{1}{N}\sigma||\Delta\Phi(z_t^i)||_\infty$ is larger than the increase in the value of $||\tilde{z}_t^i||_*$, interaction with N particles achieves a *closer* convergence to the optimum. In other words, *the number of particles can be seen as an alternative to a decreasing learning rate or vanishing noise variance, where the latter is achieved by e.g. increasing the batch size.*

3.4. Extension to a distributed setting

Assume now that the particles are optimizing $f(x) = \sum_{i=1}^N f_i(x)$. We have at our disposal a cluster with N processors to solve the optimization objective with each particle having access to f_i (and thus ∇f_i). The proof of Proposition 2 exploits the fact that each particle has access to the *same* objective function f. When each particle is optimizing a different f_i, the terms related to the gradient ∇f_i no longer vanish. Assuming the condition for consensus holds, $\theta(\kappa - \underline{\lambda}) < 0$, we have,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N ||\tilde{z}_t^i||_*^2\right] \leq \frac{K}{N}e^{-\theta(\kappa+\underline{\lambda})t}\sum_{i=1}^N ||\tilde{z}_0^i||_2^2 + \frac{K}{\theta(\kappa+\underline{\lambda})} + \frac{dK}{\theta(\kappa+\underline{\lambda})}\sigma^2\frac{(N-1)}{N}.$$

In this case, even in the no-noise case, consensus can no longer be achieved without imposing a sufficiently high interaction strength θ. Therefore, the variance reduction effect will be visible only if we impose a sufficiently strong interaction between the particles.

However, using a large interaction strength, even in the strongly convex case, may change the solution computed by the algorithm. This issue with decentralized first-order methods is well-known, see (Shi et al., 2015; Yuan et al., 2016). One possible way to address this issue is to include second-order information. For the deterministic case we propose the following dynamics, with ∇^2 the Hessian,

$$dx_t^i = \sum_{j=1}^N A_{ij}(x_t^j - x_t^i) - v_i(t) \tag{1}$$

$$dv_t^i = \sum_{j=1}^N A_{ij}(v_t^j - v_t^i) + \nabla^2 f_i(x_t^i)dx_t^i.$$

The scheme above can be viewed as the continuous time equivalent of the algorithm proposed in (Qu & Li, 2017). Its detailed convergence analysis and extension to the non-Euclidean and stochastic setting will appear elsewhere. We note that because the equation for v includes a Hessian vector product the cost of computing this term is of the same order as computing the gradient.

4. Numerical experiments

In this section we illustrate the performance of ISMD in a few test problems. We use the Euler discretization of the continuous SMD dynamics $z_{t+1}^i = z_t^i - \eta\epsilon\nabla f(x_t^i) + \epsilon\theta\sum_{j=1}^N A_{ij}(z_t^j - z_t^i) + \sigma\sqrt{\epsilon}\mathcal{N}(0,1)$, which we run for T_N time steps.

4.1. An ill-conditioned optimization problem

Consider the problem $\min_{x\in\mathcal{X}} ||Wx - b||_2^2$, where $\mathcal{X} = \Delta_n$, the unit simplex, $W \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$. Unless otherwise mentioned, we set $\epsilon = 0.1$ and $T_N = 2000$. We generate W randomly with a chosen condition number $\kappa = 1000$. We let $b_i \sim \mathcal{N}(0,1)$. We furthermore set $A_{ij} = \frac{1}{N}$. We consider the linear system optimized with ISMD with 1, 10 and 100 particles. We set $m = 100$ and $d = 100$, $\sigma = 0.05$, $\theta = 10$ and use a *fixed* learning rate $\eta = 0.0001$. From Figure 1 we observe that in a setting with a high condition number – the ill-conditioned setting – the convergence speed using interaction can be significantly faster than when considering just a single particle. At the same time the distance to the optimum is significantly improved.

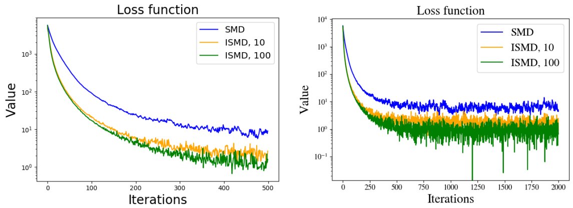

Figure 1. A comparison between the initial convergence of SMD and ISMD for the linear system with condition number 1000. We observe a speedup in convergence using interacting particles.

4.2. Interaction strength in nonconvex optimization

In this section we show that using interaction can result in a better convergence of ISMD and show it has the beneficial property of escaping saddle points while attaining consensus. We will consider the well-known Müller-Brown (MB) potential which is a standard toy example for nonconvex optimization and molecular dynamics simulations (Wales et al., 2003). The MB potential is the sum of four Gaussians in $\mathbb{R}^2$ and is given by $f(x, y) = \sum_{i=1}^{4} A_i \exp(a_i(x-\bar{x}_i)^2+b_i(x-\bar{x}_i)(y-\bar{y}_i)+c_i(y-\bar{y}_i)^2)$. The MB potential has several saddle points and local minima. We consider $A_{ij} = \frac{1}{N}$ for all i, j, and initialize the particles to a small area around a saddle point (local minimum with index 1) $x_0 = [0.212, 0.2930, 0.624]^T$; and set $\epsilon = 1$. The results for interacting SGD are shown in the top plot of Figure 2. Regular GD with independent particles starting near a saddle point are not able to escape this point. Imposing a sufficiently high interaction strength adds instability to the system so that the saddle point is no longer a stable point, and the particles escape it towards a local minimum. An alternative solution for escaping saddle points is to add noise to the system, see the bottom plot of Figure 2. In this case particles can diverge into different local minima. Imposing a high interaction strength is needed to achieve consensus.

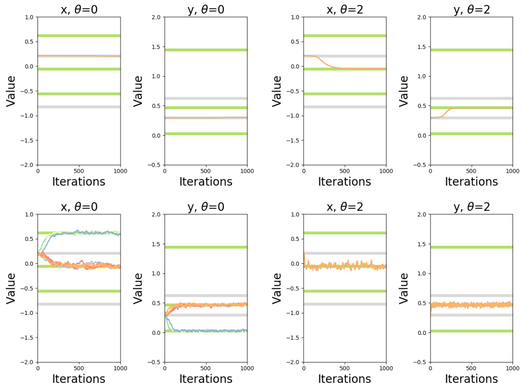

Figure 2. Optimizing the Müller-Brown potential with $A = [-200, -100, -170, 15]$, $a = [-1, -1, -6.5, 0.7],$, $b = [0, 0, 11, 0.6]$, $c = [-10, -10, -6.5, 0.7]$, $\bar{x} = [1, 0, -0.5, -1]$, $\bar{y} = [0, 0.5, 1.5, 1]$ starting from a saddle point with 10 particles using interacting SGD with a low and high interaction strength. Interactions help escape saddle points in interacting GD with learning rate $\eta = 5e-6$ (T) and the interaction strength imposes consensus in interacting SGD with $\sigma = 0.005$ (B).

4.3. Interaction in a distributed setting

In our final experiments we report preliminary numerical experiments using an Euler discretization of the second order dynamics in (1). Due to space limitations we only report results for the deterministic case in the Euclidean setting (i.e $\phi(x) = \|x\|_2^2$). We used the same setting described in Section 4.1, except that $\mathcal{X} = \mathbb{R}^n$ and the objective function function was changed to $\min_{x\in\mathbb{R}^n} \sum_{i=1}^{N} \|W_i x - b_i\|_2^2$. The results in Figure 3 clearly show that including second order information helps the algorithm converge to a consensus solution that is also optimal for the original model. For this problem IMD does not reach consensus and the mean of all the particles is still far away from the solution. As explained in Section 3.4, increasing the interaction strength can help with consensus but may not necessarily converge to the right point. Finally we note that the algorithm only includes Hessian vector products and can therefore be implemented at about the same cost of two gradient evaluations using standard automatic differentiation techniques.

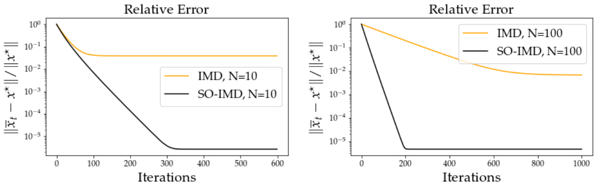

Figure 3. A comparison between the relative error between IMD SO-IMD for the linear system with condition number 1000. We observe that the interacting particle system converges to the exact solution only in the case where second order information is included. As expected adding more particles does not improve the convergence of IMD.

5. Conclusions and further work

Our analysis showed that by controlling the interaction the variance of stochastic gradients could be reduced. Extending such a result to a distributed setting could help deal with problems of instability in distributed optimization settings (De & Goldstein, 2016). In such a setting the communication costs play an important role. In future work the topology that maximizes the speed of convergence of the distributed optimization algorithm while keeping the communication cost low will be analyzed (Kar et al., 2008), (Tsianos et al., 2012), (Nedić et al., 2018). In this way efficient distributed algorithms can be developed that still benefit from variance reduction. Furthermore, the potential of including second-order information was shown here numerically and will be adressed with a theoretical analysis in further work.

In an alternative direction, we plan to use tools from the analysis of SDEs and in particular the rate of convergence to their invariant distribution. Studying the convergence of the $Law(z_t)$ is a cornerstone in the analysis of sampling schemes and can also provide valuable insights into the optimization problems; see (Raginsky et al., 2017; Shi et al., 2020; Hsieh et al., 2018) for recent works in this direction. In the numerical examples we showed the benefits of interaction in the nonconvex case. We will address the theoretical convergence guarantees in the nonconvex setting in future work, extending the analysis in e.g. (Raginsky et al., 2017; Hsieh et al., 2018) to the interacting setting.

Acknowledgements

This project was funded by JPMorgan Chase & Co under a J.P. Morgan A.I. Research Award 2019. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material does not constitute a solicitation or offer in any jurisdiction. G.P. was partially supported by the EPSRC through grant numbers EP/P031587/1.

References

Beck, A. *First-order methods in optimization*, volume 25. SIAM, 2017.

Borovykh, A., Parpas, P., Kantas, N., and Pavliotis, G. On stochastic mirror descent with interacting particles: convergence properties and variance reduction. *Submitted to Physica D.*, 2020.

Byrd, R. H., Chin, G. M., Nocedal, J., and Wu, Y. Sample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1):127–155, 2012.

Csiba, D. and Richtárik, P. Importance sampling for minibatches. *The Journal of Machine Learning Research*, 19 (1):962–982, 2018.

De, S. and Goldstein, T. Efficient distributed sgd with variance reduction. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 111–120. IEEE, 2016.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pp. 1646–1654, 2014.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

Gorbunov, E., Hanzely, F., and Richtárik, P. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 2878–2887, 2018.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Kar, S., Aldosari, S., and Moura, J. M. Topology for distributed inference on graphs. *IEEE Transactions on Signal Processing*, 56(6):2609–2613, 2008.

Koloskova, A., Stich, S. U., and Jaggi, M. Decentralized stochastic optimization and gossip algorithms with compressed communication. *arXiv preprint arXiv:1902.00340*, 2019.

Lin, P., Ren, W., and Farrell, J. A. Distributed continuous-time optimization: nonuniform gradient gains, finite-time convergence, and convex constraint set. *IEEE Transactions on Automatic Control*, 62(5):2239–2253, 2016.

Mertikopoulos, P. and Staudigl, M. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

Mesbahi, M. and Egerstedt, M. *Graph theoretic methods in multiagent networks*. Princeton University Press, 2010.

Nedić, A., Olshevsky, A., and Rabbat, M. G. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106 (5):953–976, 2018.

Nemirovsky, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Qu, G. and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Raginsky, M. and Bouvrie, J. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 6793–6800. IEEE, 2012.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Shahrampour, S. and Jadbabaie, A. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, 2017.

Shi, B., Su, W. J., and Jordan, M. I. On learning rates and schrödinger operators. *arXiv preprint arXiv:2004.06977*, 2020.

Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Tsianos, K., Lawlor, S., and Rabbat, M. G. Communication/computation tradeoffs in consensus-based distributed optimization. In *Advances in neural information processing systems*, pp. 1943–1951, 2012.

Wales, D. et al. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

Yuan, K., Ling, Q., and Yin, W. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.