Optimizing interacting Langevin dynamics using spectral gaps

Anastasia Borovykh [1] **Nikolas Kantas** [2] **Panos Parpas** [1] **Greg Pavliotis** [2]

Abstract

Using independent runs of an optimization algorithm is a standard scheme for finding the minimizer of an objective function. In this paper we seek to improve this practice by allowing the different optimizers to interact. The question that arises is: how can one choose the interaction structure that will result in the fastest convergence while maintaining minimal communication costs? To investigate this issue we formulate it through the optimization of the spectral gap of interacting Langevin dynamics. In the case of a linear interaction, the spectral gap is directly related to the spectrum of the Laplacian matrix that characterizes the interaction. We present early numerical results in both convex and non-convex settings that illustrate the benefit of choosing the right kind of interaction structure.

1. Introduction

The challenge of finding optimal schemes for both sampling and optimization (Ma et al., 2019) is a relevant topic due to the ever-increasing use of machine learning algorithms on data. In recent years the challenge of finding such optimal schemes has been addressed by changing the dynamics through e.g. the addition of nonreversible perturbations (Lelievre et al., 2013) (Duncan et al., 2016), Riemannian manifold optimization (Patterson & Teh, 2013) or preconditioning (AlRachid et al., 2018). Recent attempts have been made to do this using interacting systems (Borovykh et al., 2021). Rather than using independent runs of gradient dynamics, hereby referred to as *particles*, we allow them to interact. The goal in this setting is to choose the interaction in such a way that i) one can speed up convergence to the minimizer while having ii) consensus between all the particles, and iii) maintaining low communication costs. In this paper we aim to shed more light on the challenge of finding the right kind of interaction potential by formulating the problem through the spectral gap in the Langevin dynamics. In previous work the spectral gap of the generator of the dynamics has been shown to control the convergence speed to the invariant measure (Bauerschmidt & Bodineau, 2019) (Bakry et al., 2008) or the mixing speed of Markov chains (Boyd et al., 2004). We first present the relationship between sampling from an invariant measure and minimizing an objective function and show that including interactions does not alter the optimum. Then we characterize convergence to the invariant measure through the spectral gap and connect the spectral gap to the spectrum of the interaction matrix. Finally, our numerical examples show how the right kind of interaction structure can maximize the convergence speed and reduce asymptotic variance in convex and non-convex settings while maintaining low communication costs.

[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>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

2. The setting

Consider an objective function $f : \mathbb{R}^d \to \mathbb{R}$. The goal is to find a minimizer x^*,

$$x^* = \arg\min_{x\in\mathbb{R}^d} f(x).$$

This goal can be reformulated through the problem of sampling from the probability measure

$$\pi(d\mathbf{x}) = \frac{1}{Z}e^{-\frac{\sigma^2}{2}V(\mathbf{x})}d\mathbf{x}, \quad Z = \int e^{-\frac{\sigma^2}{2}V(\mathbf{x})}d\mathbf{x}, \quad (1)$$

where $\mathbf{x} = \left(x^{1^T}, ..., x^{N^T}\right)^T$ is the concatenation vector of all particles and σ^2 acts as a user chosen annealing parameter. An obvious choice is to set $V(\mathbf{x}) = \sum_{i=1}^N f(x^i)$, which is equivalent to sampling *independently* N times from a density proportional to $e^{-\frac{\sigma^2}{2}f}$ and computing the mode. This choice does not involve any interaction between the particles. Instead one could choose,

$$V(\mathbf{x}) = \sum_{i=1}^N f(x^i) + \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N A_{ij}(x^i - x^j)^2, \quad (2)$$

where the interaction potential is chosen to be quadratic with interaction matrix A, where $A_{ij} = 0$ if two particles are *not* connected and do not communicate or interact.

Our first goal is to design dynamics with invariant distribution being π, and in addition such that all particles achieve

consensus at some point in time (even if asymptotically). Consensus strictly means $x^i = x^j$ for all $i, j = 1, ..., N$, but here we adopt the convention that x^i, x^j are very close with high probability. The Langevin dynamics for (2) are given by,

$$dx_t^i = -\nabla f(x_t^i)dt - \sum_{j=1}^{N} A_{ij}(x_t^i - x_t^j) + \sigma dB_t^i,$$

where B_t^i are d-dimensional independent Brownian motions. The initial condition can be set to a value $x_0^i = x^i$ or distribution $x_0^i \sim \pi_0$. The second goal of these dynamics is to converge to stationarity quickly with minimal communication between particles.

We can rewrite these dynamics in vectorized form,

$$d\mathbf{x}_t = -\nabla f(\mathbf{x}_t)dt - \mathcal{L}\mathbf{x}_t dt + \sigma d\mathbf{B}_t, \tag{3}$$

where f is the concatenation of gradients and the graph Laplacian is given by $\mathcal{L} = (\text{Diag})(A\mathbf{1}_N) - A) \otimes I_d$ with smallest eigenvalue zero. The self-adjoint generator of (3) is,

$$\mathcal{A}\phi = -\sum_{j=1}^{dN} \nabla_j V(\mathbf{x}) \frac{\partial}{\partial x_j}\phi + \frac{1}{2}\sigma^2 \sum_{i,j=1}^{dN} \frac{\partial^2}{\partial x_i \partial x_j}\phi.$$

The spectral gap λ for a stochastic process with generator $\mathcal{A}$ is defined as the smallest non-zero eigenvalue of the generator,

$$\lambda = \sup\big\{\text{Re}(z) : z \in \sigma(\mathcal{A}), z \neq 0\big\},$$

where $\sigma(\mathcal{A})$ is the spectrum of the generator $\mathcal{A}$ (Pavliotis, 2014). The spectral gap determines the convergence rate to stationarity. To appreciate how relevant this is for optimisation consider the risk bound,

$$\begin{aligned}|\mathbb{E}[V(\mathbf{x}_t)] - V(\mathbf{x}^*)| \leq |\mathbb{E}[V(\mathbf{x}_t)] - \mathbb{E}_\pi[V(\mathbf{x})]| \\ + |\mathbb{E}_\pi[V(\mathbf{x})] - V(\mathbf{x}^*)|.\end{aligned}$$

The first term is affected by λ through (4) and the second term depends on σ^2 and how π is centered around $\mathbf{x}^*$. After some number of iterations, $\mathbf{x}_t$ can be thought as approximate samples from π and the mode can be used as an estimate of $\mathbf{x}^*$. This requires that N concatenated copies of the optimizer of $f(x)$ are equal to the one from $V(x)$ in (2). This can be achieved for A being doubly stochastic.

Lemma 2.1 (Interaction does not alter the optimum)**.** *Let $x^* := \arg\min_x f(x)$, and let $V(\mathbf{x})$ be as defined in (2). Let A be doubly stochastic. Then $\mathbf{x}^*$ is a minimizer of $V(\mathbf{x})$.*

Proof. See Lemma 21 in (Borovykh et al., 2021). □

3. Convergence results through spectral gaps

We present two performance measures based on the spectral gaps that control the convergence speed to the invariant measure (1). The relevant quantities are: i) $\lambda_{\max} - \lambda_{\min}$, the largest and smallest eigenvalues of A and ii) λ_2, the second-smallest eigenvalue of A.

3.1. A spectral gap result

Let π_t be the probability law of the process $\mathbf{x}_t$ at time t. A standard convergence result is

$$||\pi_t - \pi|| \leq Ce^{-\lambda t}||\pi_0 - \pi||, \tag{4}$$

which is valid various choices of the norm $||\cdot||$, e.g. $L^2(\mathbb{R}^{dN}, \pi^{-1}dx)$ with $C = 1$ for reversible, self-adjoint dynamics (Lelievre et al., 2013) or the Wasserstein distance (Bakry et al., 2013); see (Pavliotis, 2014) for an overview.

Note that the spectral gap of the generator can be computed explicitly for a quadratic potential function (Metafune et al., 2002) (Lelievre et al., 2013) (Ottobre et al., 2012). Consider a linear system with,

$$f(x) = ||Mx - b||_2^2, \tag{5}$$

where M is some matrix and let $\hat{M} = M \otimes I_N$. The convergence rate can be expressed through the eigenvalues of the interaction matrix.

Proposition 1. *Consider the dynamics in (3) with f as in (5). The spectrum of $\mathcal{A}$ is given by,*

$$\sigma(\mathcal{A}) = \Big\{ \sum_{j=1}^{r} -n_j\lambda_j, n_j \subset \mathbb{N} \Big\},$$

where $\{\lambda_j\}_{j=1}^r$ are the r distinct eigenvalues of the matrix B defined as $B := \hat{M} + \mathcal{L}$.

Proof. The result is a straight-forward adaptation of Theorem 3.1 in (Metafune et al., 2002) or Proposition 10 in (Lelievre et al., 2013). □

From the definition of the spectral gap, it follows that λ for the stochastic process (3), $\lambda = \lambda_2^B$ where λ_2^B is the second-smallest eigenvalue of B. The work of (Boyd et al., 2004) similarly optimizes the second-largest eigenvalue modulus of the transition matrix to maximize the mixing rate of Markov chains.

3.2. A Log-Sobolev convergence result

The problem that we study in this work is related to the study of Gibbs measures for unbounded spin systems (Bauerschmidt & Bodineau, 2019). Our second convergence result is a bound on the relative entropy for π and a test function φ :

$\mathbb{R}^d \to \mathbb{R}$ defined as $\text{Ent}_\pi(\varphi) = \pi(\varphi \log \varphi) - \pi(\varphi) \log \pi(\varphi)$ with $\pi(\varphi) := \int \varphi(x)\pi(dx)$. The bound on the relative entropy in turn can be used to bound the Wasserstein distance in (4) and hence the total variation (Bakry et al., 2013). We present the result for the mean-field case; it holds more generally with $\langle A \rangle = \lambda_{\max} - \lambda_{\min}$ as long as $\langle A \rangle < N$.

Proposition 2. *The Gibbs measure π of the N-particle system with* mean-field *interaction $A_{ij} = \frac{1}{N-1}$ for $i \neq j$ as in (3) satisfies a Log-Sobolev inequality (LSI) with $\langle A \rangle = 1$,*

$$\text{Ent}_\pi(f^2) \leq 2\left(1 + \frac{2N\langle A \rangle}{N - \langle A \rangle}\right)\pi(|\nabla f|^2).$$

Proof. Observe that we can rewrite $V(\mathbf{x})$ as,

$$\sum_{i=1}^{N}\left(f(x^i) + \frac{1}{2}\left(\sum_k A_{ki}\right)x_i^2\right) - \frac{1}{2}\sum_{j\neq i}^{N} A_{ij}x^i x^j.$$

The Gibbs measure (1) of the N-particle system becomes

$$\pi(d\mathbf{x}^N) = \frac{1}{Z_N} e^{-\frac{\sigma^2}{2}\sum_{i\neq j} A_{ij}x_i x_j} \prod_{j=1}^{N} \mu(dx^i),$$

where $\mu(dx) = \frac{1}{Z} e^{-\sigma^2\left(f(x) + \frac{(N-1)}{2N}x^2\right)}dx$ and we used the double stochasticity of A. The result follows by application of Theorem 1 from (Bauerschmidt & Bodineau, 2019). □

This implies that (4) holds with $\lambda = \left(1 + \frac{2N\langle A \rangle}{N - \langle A \rangle}\right)^{-1}$.

4. Numerical examples

We demonstrate numerically i) the benefit of interactions in speeding up convergence to optimality and consensus, ii) the role of the spectrum of the interaction matrix in controlling the convergence rate, and iii) the effect of the interaction matrix on the communication costs. We simulate the SDEs using Euler discretization with learning rate 0.01, $\Delta t = 0.01$, noise $\sigma = 0.01$, and in the linear system M has condition number 100.

4.1. The effect of interaction in a convex system

To show the importance of interactions in a linear system (5) we compare an independent system to one with a mean-field interaction given by $A_{ij} = \frac{1}{N-1}$ for $i \neq j$ and $A_{ii} = 0$. Figure 1 shows the convergence speed and the histogram of the 500 final losses $V(\mathbf{x}_t)$.

4.2. The role of the convergence metrics

Consider again the linear system (5); we now analyze the influence of the spectral gap through the convergence of two different Barbell interaction graphs. In Figure 2 we observe an initial faster convergence of the system with the higher second-smallest eigenvalue.

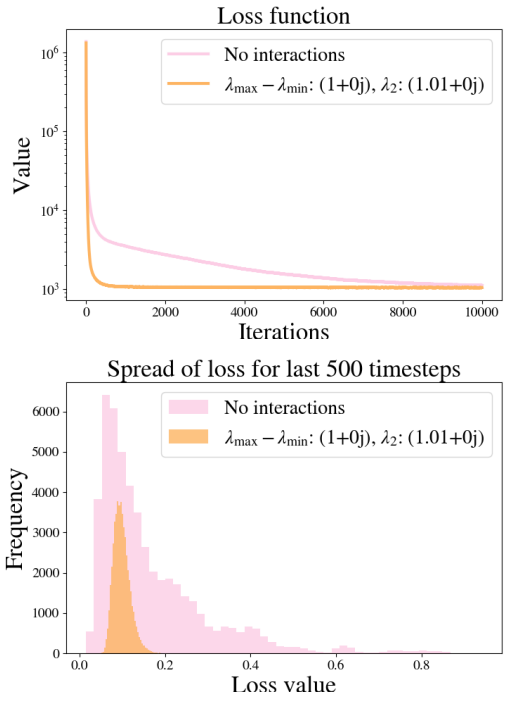

Figure 1. A linear system: convergence speed (T) and histogram of the last 500 losses (B) for 100 interacting (orange) and non-interacting particles (pink).

4.3. Communication costs in a non-convex setting

Lastly, we show the benefits of optimizing the interaction graph in terms of both achieving consensus faster as well as using less communication. We consider the interaction *parametrized* as follows,

$$A_{ij} = \phi(x^i - x^j) = \begin{cases} a, \text{ for } ||x^i - x^j||_2 \leq \frac{1}{\sqrt{2}}, \\ b, \text{ for } \frac{1}{\sqrt{2}} < ||x^i - x^j||_2 \leq 1, \\ 0, \text{ for } ||x^i - x^j||_2 > 1, \end{cases}$$

where x^i and x^j are the particles. Note that the matrix is thus *time-varying*. The parameters a and b determine the *strength* of the interaction: a high b/a will result in a stronger attraction between particles that are further away from each other. Such heterophilious dynamics have been shown to lead to faster convergence (Motsch & Tadmor, 2014). We consider the Muller-Brown potential as the optimization objective, where each $x^i = [\mathrm{x}, \mathrm{y}]$,

$$f(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{4} A_i e^{a_i(\mathrm{x}-\bar{\mathrm{x}}_i)^2 + b_i(\mathrm{x}-\bar{\mathrm{x}}_i)(\mathrm{y}-\bar{\mathrm{y}}_i) + c_i(\mathrm{y}-\bar{\mathrm{y}}_i)}.$$

The problem of finding the optimal interaction structure here amounts to optimizing b/a. In Figure 3 we plot the

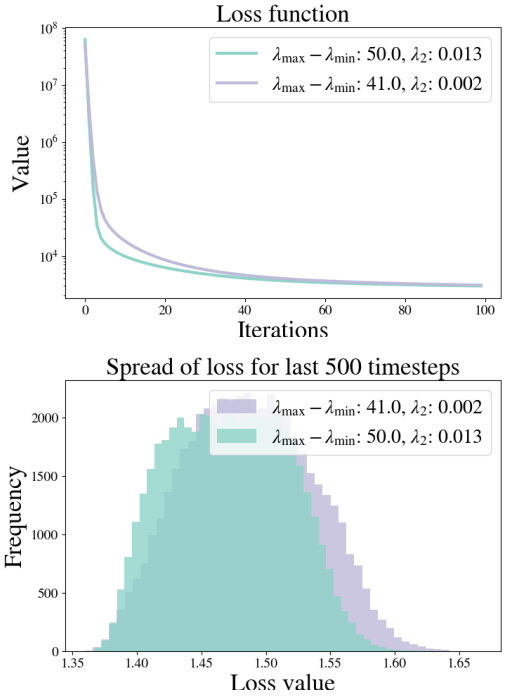

Figure 2. Convergence speed (T) and histogram of the last 500 losses (B) for 100 interacting agents with a Barbell graph with two clusters of 48 (green) and 40 (purple) particles each with a connection length of 2 (green) and 20 (purple).

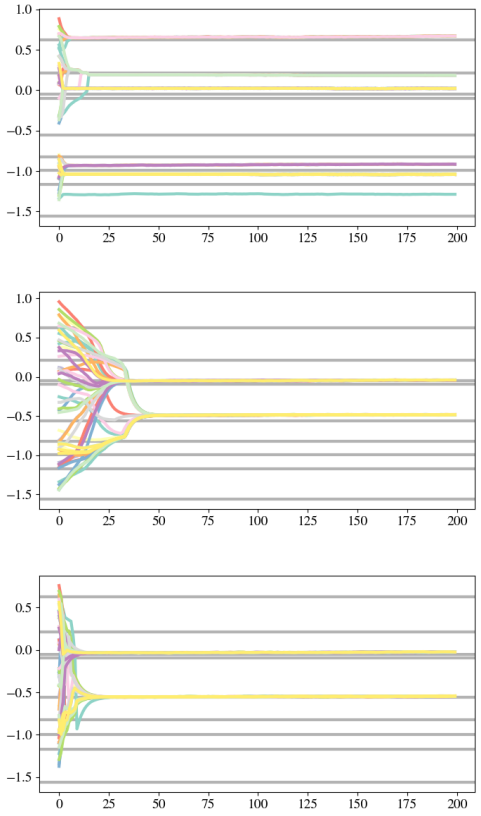

Figure 3. Non-convex Muller-Brown potential: evolution of the different particles (colored lines) to various minima and saddle points (grey lines) (T) $a = 10$ and $b = 1$, (C) $a = 1$ and $b = 1$ and (B) $a = 1$ and $b = 10$. More consensus is achieved faster for higher b/a.

x-coordinate evolution over time and Figure 4 shows the communication cost (i.e. how many agents communicate their values) for different values of b/a. A higher b/a results in faster convergence to *consensus* with *lower* communication cost.

5. Conclusion

We showed how the spectral gap of the generator of the optimization dynamics can be used to control convergence speed and consensus (or spread of the particles). The expression for the spectral gap of the generator can be explicitly derived in the setting of a linear interaction potential and objective function where it is defined through the spectrum of the interaction matrix. By allowing for interactions between the particles one can *control* the convergence of the dynamics as well as the asymptotic variance. In future work we will derive explicit convergence bounds for the distance to the optimum and show how to *optimize* the interaction matrix for fast convergence while *not exceeding a pre-defined communication budget*.

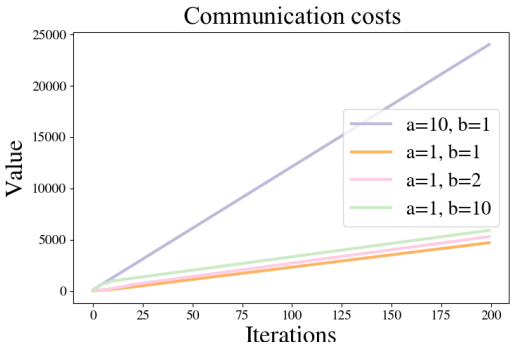

Figure 4. Non-convex Muller-Brown potential: the cumulative communication costs i.e. number of times $A_{ij} \neq 0$ for different values of b/a. The right kind of communication results in fast convergence and low communication cost.

References

AlRachid, H., Mones, L., and Ortner, C. Some remarks on preconditioning molecular dynamics. *The SMAI journal of computational mathematics*, 4:57–80, 2018.

Bakry, D., Cattiaux, P., and Guillin, A. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *Journal of Functional Analysis*, 254(3): 727–759, 2008.

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Bauerschmidt, R. and Bodineau, T. A very simple proof of the LSI for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

Borovykh, A., Kantas, N., Parpas, P., and Pavliotis, G. A. On stochastic mirror descent with interacting particles: convergence properties and variance reduction. *Physica D: Nonlinear Phenomena*, 418:132844, 2021.

Boyd, S., Diaconis, P., and Xiao, L. Fastest mixing Markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

Duncan, A. B., Lelievre, T., and Pavliotis, G. Variance reduction using nonreversible langevin samplers. *Journal of Statistical Physics*, 163(3):457–491, 2016.

Lelievre, T., Nier, F., and Pavliotis, G. A. Optimal nonreversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2): 237–274, 2013.

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42): 20881–20885, 2019.

Metafune, G., Pallara, D., and Priola, E. Spectrum of Ornstein-Uhlenbeck operators in Lp spaces with respect to invariant measures. *Journal of Functional Analysis*, 196(1):40–60, 2002.

Motsch, S. and Tadmor, E. Heterophilious dynamics enhances consensus. *SIAM review*, 56(4):577–621, 2014.

Ottobre, M., Pavliotis, G., and Pravda-Starov, K. Exponential return to equilibrium for hypoelliptic quadratic systems. *Journal of Functional Analysis*, 262(9):4000–4039, 2012.

Patterson, S. and Teh, Y. W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *NIPS*, pp. 3102–3110, 2013.

Pavliotis, G. A. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.