Accelerated Mirror Gradient Descent over the Infinite Dimensional Space of Probability Measures

Haoxuan Chen
haoxuanc@stanford.edu
Institute for Computational and Mathematical Engineering (ICME)
Stanford University

1 Introduction

In this section, we firstly provide a brief introduction to two popular research topics - accelerated first-order optimization algorithms and optimization over the probability space. On the one hand, gradient-based (or equivalently first-order) optimization algorithms have been used extensively in many scientific fields nowadays, including but not limited to machine learning, computational physics and operations research. Among all first-order methods, one popular research topic is to design and analyze accelerated gradient-based optimization algorithms. Following the phenomenal work Su et al. [2016], extensive studies have been conducted to analyze gradient-based methods based on a framework induced by either differential equations or optimal control Lessard et al. [2016], Zhang et al. [2018], Betancourt et al. [2018], Shi et al. [2019], Wilson et al. [2021], Kovachki and Stuart [2021], Shi et al. [2022], Lu [2022], Shi et al. [2023], Moucer et al. [2023]. Similar analysis Li et al. [2017, 2019], An et al. [2020], Zhu and Ying [2021] have also been applied to stochastic gradient descent and its variants.

On the other hand, many problems in machine learning, statistics and operations research can be formulated as an optimization problem over the space of probability measures, which is of infinite dimension. Therefore, many work have focused on developing efficient and fast algorithms for related problems, such as Wang and Li [2020], Ying [2020, 2021], Kent et al. [2021a,b], Wang and Li [2022], Wang and Yan [2022], Chen et al. [2023]

2 Approach

In this project, we plan to investigate an accelerated version of the mirror descent (MD) algorithm and its stochastic version, which is orginally proposed in Krichene et al. [2015]. Specifically, they have derived a coupled ODE system describing the accelerated MD algorithm when it is applied to some function f:

$$\begin{aligned} \frac{d}{dt}X(t) &= \frac{r}{t}(\nabla\Phi^*(Z(t)) - X(t)) \\ \frac{d}{dt}Z(t) &= -\frac{t}{r}\nabla f(X(t)) \\ X(0) = x_0,\ Z(0) &= z_0,\ \nabla\Phi^*(z_0) = x_0 \end{aligned} \tag{1}$$

where X denotes the primal variable to be optimized, Z denotes the dual variable and Φ is the mirror map. Similarly, an accelerated version of the stochastic mirror descent method has been proposed in Krichene and Bartlett [2017]. We aim to study if there is a way to develop an accelerated version of the algorithm proposed in Ying [2020], i.e., does there exist an analog of accelerated MD algorithm over the space of probability measures? Moreover, is there a way to incorporate the adaptive restarting methodology proposed in O'donoghue and Candes [2015] within this algorithm? Theoretically, we will also try to investigate if there is a partial differential equation (PDE) governing the evolution of

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

the density under the accelerated MD algorithm, just as that of Wang and Li [2022] and Chen et al. [2023].

3 Background Knowledge

Let $\mathcal{P}(\mathbb{R}^d)$ denote the space formed by all the probability measures defined over $\mathbb{R}^d$. Recall that for any functional $F = F(\mu) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, the first variation of F, which is the analog of gradient over infinite dimensional spaces, is defined in a way that the following identity

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(F(\nu + \epsilon h) - F(\nu)) = \int_{\mathbb{R}^d} \frac{\delta F}{\delta \mu}(\nu)(x)h(x)dx \tag{2}$$

holds for any h satisfying $\int_{\mathbb{R}^d} h(x)dx = 0$. Then for any convex functional $\Phi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, we have that the associated Bregman divergence $D_\Phi(\cdot, \cdot) : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is given by

$$D_\Phi(p, q) = \Phi(p) - \Phi(q) - \langle \frac{\delta \Phi}{\delta \mu}(q), p - q \rangle \tag{3}$$

where the dot product $\langle \cdot, \cdot \rangle$ above satisfies $\langle f, g \rangle = \int_{\mathbb{R}^d} f(x)g(x)dx$ for any $f, g \in L^2(\mathbb{R}^d)$. Below we consider a special case of the mirror descent algorithm, which chooses the KL divergence as the distance metric in and proximal formulation and is depicted in Ying [2020]. In order to obtain the PDE dynamics governing the evolution of the probability measures under this specific mirror descent algorithm, we begin by expressing the KL divergence between any two probability measures as a Bregman divergence. In fact, consider the entropy function $H : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ defined as follows

$$H(\mu) = \int_{\mathbb{R}^d} \mu(x) \log(\mu(x))dx \tag{4}$$

Computing the first variation of the entropy functional H yields $\frac{\delta H}{\delta \mu} = \log(\mu) + 1$. Substituting this identity into (3) yields that the associated Bregman divergence is given by

$$\begin{aligned} D_H(p, q) &= H(p) - H(q) - \int_{\mathbb{R}^d} (p - q)(\log q + 1)dx \\ &= \int_{\mathbb{R}^d} (p \log p)dx - \int_{\mathbb{R}^d} (p \log q)dx \\ &= \int_{\mathbb{R}^d} p \log(\frac{p}{q})dx = KL(p \| q) \end{aligned} \tag{5}$$

which coincides with the KL divergence between p and q, as desired. With the properties listed above, we proceed to present the continuous PDE dynamics in the following section.

4 Continuous PDE Dynamics

Recall that for mirror descent defined over the finite dimensional space $\mathbb{R}^d$, the ODE dynamics coupling the primal variable $X \in \mathbb{R}^d$ and the dual variable $Z \in \mathbb{R}^d$ is given by

$$\begin{aligned} X(t) &= \nabla \phi^*(Z(t)) \\ \frac{d}{dt} Z(t) &= -\nabla f(X(t)) \\ X(0) &= x_0,\ Z(0) = z_0,\ \nabla \phi^*(z_0) = x_0 \end{aligned} \tag{6}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is the convex objective function, $\phi : \mathbb{R}^d \to \mathbb{R}$ is the convex mirror map defining the Bregman divergence defined in the mirror descent and $\phi^*(y) = \sup_{x \in \mathbb{R}^d}(x^T y - \phi(x))$ is the convex conjugate of ϕ. From Danskin's Theorem, we may further deduce that

$$\nabla \phi^*(y) = \arg\max_{x \in \mathbb{R}^d}(x^T y - \phi(x)) \tag{7}$$

Therefore, under the setting of applying mirror descent associated with the KL divergence over the space of probability measures, we may replace the mirror map ϕ with the entropy function H

computed above. Then we have that the gradient of the conjugate map H^* can be defined via a similar form as (7) above

$$\nabla H^*(q) = \arg\max_{p\in\mathcal{P}(\mathbb{R}^d)}(\langle p, q\rangle - H(p)) \tag{8}$$

By taking the first variation with respect to p in the expression above, we may solve for the optimal p^* as follows

$$q - (\log p^* + C) = 0 \Rightarrow \nabla H^*(q) = p^* \propto e^q \tag{9}$$

i.e., $p(x) = (\int e^{q(x)}dx)^{-1}e^{q(x)}$ for any $x \in \mathbb{R}^d$. Therefore, by using ρ and μ to represent the primal and dual variables respectively, we can rewrite (6) to deduce that the PDE dynamics governing the mirror descent over probability spaces is given by

$$\begin{aligned} \rho(x,t) &= \left(\int_{\mathbb{R}^d} e^{\mu(x)}dx\right)^{-1} e^{\mu(x)} \\ \frac{\partial}{\partial t}\mu(x,t) &= -\frac{\delta E}{\delta \rho}(\rho(x,t)) \\ \rho(x,0) &= \rho_0(x),\ \mu(x,0) = \mu_0(x),\ \rho_0 \propto e^{\mu_0} \end{aligned} \tag{10}$$

where $E = E(\rho(x,t)) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is the energy functional over the space of probability measures that we aim to optimize. Similarly, by replacing the mirror map Φ in (1) above, which is the dynamics governing the accelerated mirror descent, we obtain the following PDE system that corresponds to the continuous formulation of the accelerated mirror descent algorithm (associated with the KL divergence) over the space of probability measures:

$$\begin{aligned} \frac{\partial}{\partial t}\rho(x,t) &= \frac{r}{t}\left(\left(\int_{\mathbb{R}^d} e^{\mu(x)}dx\right)^{-1} e^{\mu(x)} - \rho(x,t)\right) \\ \frac{\partial}{\partial t}\mu(x,t) &= -\frac{t}{r}\frac{\delta E}{\delta \rho}(\rho(x,t)) \\ \rho(x,0) &= \rho_0(x),\ \mu(x,0) = \mu_0(x),\ \rho_0 \propto e^{\mu_0} \end{aligned} \tag{11}$$

where r is some fixed constant and picked to be $r = 3$ in the Nesterov acceleration method. Currently we are still trying to prove that the dynamics (11) does converge faster (i.e, have a better convergence rate) compared to the dynamics (10), which justifies the efficacy of accelerated mirror descent algorithm over the space of probability measures.

5 Numerical Experiments

In addition to theoretical analysis, we also consider performing a few numerical experiments to justify the acceleration phenonmenon, Below we list a few examples that we plan to use for testing our algorithms. Similar to the settings adpoted in Ying [2020], Chen et al. [2023], below is a general form of the loss function that we aim to test on:

$$F(\rho) = D(\rho\|\nu) + \int_{\mathbb{R}^d}\rho(x)V(x)dx + \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\rho(x)W(x,y)\rho(y)dxdy \tag{12}$$

where $D(\cdot\|\cdot)$ is some fixed metric between any two probability distributions, $V(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is some potential function and $W(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is some symmetric positive kernel corresponding to the kinetic energy. We remark that the setting above has various applications in many settings, ranging from the Keller-Segel model in mathematical biology, granular flows in kinetic theory and training of two-layer neural networks in machine learning. Currently we are still wrapping up the numerical experiments in all the settings above to justify that the discretized algorithm corresponding to the continuous system (11) is indeed more effective than that of the continuous system (10).

6 Future Work

In this subsection, we also include a few possible topics for future work that we discover in our study. **On the one hand, for algorithmic studies, below is a list of questions that remain unanswered, which we think will be interesting to investigate:**

(Q1) Is there a way to incorporate the adaptive restarting methodology proposed in O'donoghue and Candes [2015] within this algorithm? Within the specific setting of training two layer neural networks under the mean-field formulation, an algorithm based on the Fisher-Rao gradient flow has been proposed in Rotskoff et al. [2019]

On the other hand, regarding theoretical studies, we also provide a list of potential questions as follows:

(Q2) For accelerated MD and stochastic MD algorithms, previous work have only derived the corresponding continuous time dynamics Krichene et al. [2015], Krichene and Bartlett [2017], Xu et al. [2018b,a]. It seems that none of them has conducted numerical analysis to quantify the distance between the discrete algorithm and the continuous dynamics. In contrast, such quantification has been performed for SGD, ASGD and Nesterov accelerated algorithm via the modified equation approach by researchers from the applied math community. See Li et al. [2017, 2019], Kovachki and Stuart [2021], An et al. [2020], Zhu and Ying [2021]. Also, can we analyze the ODE (1) based on control theory, just as what has been done in Lessard et al. [2016], Zhao et al. [2021], Moucer et al. [2023]? In general, since the coupling between vanilla Gradient Descent (GD) and MD has been studied in Allen-Zhu and Orecchia [2014], it seems to me that what has been analyzed for accelerated GD is also doable for accelerated MD.

(Q3) How can the accelerated MD algorithm be interpreted under the Optimal Transport (OT) framework Mishchenko [2019], Aubin-Frankowski et al. [2022], Eckstein [2023]? Some recent work Deb et al. [2023], Karimi et al. [2024] has been trying to link the MD algorithm with the Sinkhorn algorithm via a PDE based approach. How does the accelerated MD algorithm look like under their framework? Also, how does the MD algorithm relate to the primal-dual method designed for the Wasserstein gradient flow Carrillo et al. [2022]?

References

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Jing An, Jianfeng Lu, and Lexing Ying. Stochastic modified equations for the asynchronous stochastic gradient descent. *Information and Inference: A Journal of the IMA*, 9(4):851–873, 2020.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

José A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows. *Foundations of Computational Mathematics*, pages 1–55, 2022.

Shi Chen, Qin Li, Oliver Tse, and Stephen J Wright. Accelerating optimization over the space of probability measures. *arXiv preprint arXiv:2310.04006*, 2023.

Nabarun Deb, Young-Heon Kim, Soumik Pal, and Geoffrey Schiebinger. Wasserstein mirror gradient flow as the limit of the sinkhorn algorithm. *arXiv preprint arXiv:2307.16421*, 2023.

Stephan Eckstein. Hilbert's projective metric for functions of bounded growth and exponential convergence of sinkhorn's algorithm. *arXiv preprint arXiv:2311.04041*, 2023.

Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. Sinkhorn flow as mirror flow: A continuous-time framework for generalizing the sinkhorn algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 4186–4194. PMLR, 2024.

Carson Kent, Jose Blanchet, and Peter Glynn. Frank-wolfe methods in probability space. *arXiv preprint arXiv:2105.05352*, 2021a.

Carson Kent, Jiajin Li, Jose Blanchet, and Peter W Glynn. Modified frank wolfe in probability space. *Advances in Neural Information Processing Systems*, 34:14448–14462, 2021b.

Nikola B Kovachki and Andrew M Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021.

Walid Krichene and Peter L Bartlett. Acceleration and averaging in stochastic mirror descent dynamics. *arXiv preprint arXiv:1707.06219*, 2017.

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28, 2015.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

Haihao Lu. An o (sr)-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Mathematical Programming*, 194(1):1061–1112, 2022.

Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.

Céline Moucer, Adrien Taylor, and Francis Bach. A systematic approach to lyapunov analyses of continuous-time models in convex optimization. *SIAM Journal on Optimization*, 33(3):1558–1586, 2023.

Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15:715–732, 2015.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *International conference on machine learning*, pages 5508–5517. PMLR, 2019.

Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *Advances in Neural Information Processing Systems*, 32, 2019.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2022.

Bin Shi, Weijie Su, and Michael I Jordan. On learning rates and schrödinger operators. *Journal of Machine Learning Research*, 24(379):1–53, 2023.

Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Li Wang and Ming Yan. Hessian informed mirror descent. *Journal of Scientific Computing*, 92(3):90, 2022.

Yifei Wang and Wuchen Li. Information newton's flow: second-order optimization method in probability space. *arXiv preprint arXiv:2001.04341*, 2020.

Yifei Wang and Wuchen Li. Accelerated information gradient flow. *Journal of Scientific Computing*, 90:1–47, 2022.

Ashia C Wilson, Ben Recht, and Michael I Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

Pan Xu, Tianhao Wang, and Quanquan Gu. Accelerated stochastic mirror descent: From continuous-time dynamics to discrete-time algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1087–1096. PMLR, 2018a.

Pan Xu, Tianhao Wang, and Quanquan Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning*, pages 5492–5501. PMLR, 2018b.

Lexing Ying. Mirror descent algorithms for minimizing interacting free energy. *Journal of Scientific Computing*, 84(3):51, 2020.

Lexing Ying. Natural gradient for combined loss using wavelets. *Journal of Scientific Computing*, 86(2):26, 2021.

Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. *Advances in neural information processing systems*, 31, 2018.

Shipu Zhao, Laurent Lessard, and Madeleine Udell. An automatic system to detect equivalence between iterative algorithms. *arXiv preprint arXiv:2105.04684*, 2021.

Yuhua Zhu and Lexing Ying. A sharp convergence rate for a model equation of the asynchronous stochastic gradient descent. *Communications in Mathematical Sciences*, 19(3):851–863, 2021.