Technical Report on Mirror Descent, Bregman Divergence, and Their ODE Formulations

Walid Krichene, Alexandre Bayen, and Peter LBartlett. Accelerated mirror descent in continuous and discrete time.

Qiyao Wei

November 20, 2020

Outline

1. Background
   - Gradient Descent and Mirror descent
   - Previous Work
2. Contribution of the ODE Paper
   - A Simple Example
   - Main Results
   - Basic Ideas for Proofs/Implementations

Outline

Recap of vanilla gradient descent

- Convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$
- differentiable and L-Lipschitz, i.e. $\|f(x) - f(y)\| \leq L(x - y)$
- $$f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) - f(x^*) \leq \frac{RL}{\sqrt{T}}$$

Recap of vanilla gradient descent

- Convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$
- differentiable and L-Lipschitz, i.e. $\|f(x) - f(y)\| \leq L(x - y)$
- $$f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) - f\left(x^*\right) \leq \frac{RL}{\sqrt{T}}$$

Recap of vanilla gradient descent

- Convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$
- differentiable and L-Lipschitz, i.e. $\|f(x) - f(y)\| \leq L(x - y)$
- $$f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) - f\left(x^*\right) \leq \frac{RL}{\sqrt{T}}$$

Recap of Bregman Divergence

- $$D_{\Phi}(a, b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a - b))$$
- Interactive demo: http://mark.reid.name/blog/meet-the-bregman-divergences.html

Recap of Bregman Divergence

- $$D_\Phi(a, b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a - b))$$
- Interactive demo: http://mark.reid.name/blog/meet-the-bregman-divergences.html

Fundamental assumptions

- f being L-Lipschitz places a constraint on the gradient of f
- Gradient and mirror descent converges in $\mathcal{O}(\frac{1}{t^{1/2}})$

- f is normally constrained on the gradient of its gradient, i.e. ∇f being L-Lipschitz, or

$$\|\nabla f(x) - \nabla f(y)\| \leq L(x - y)$$

- This achieves $\mathcal{O}(\frac{1}{t})$, as we will see in the ODE paper

Fundamental assumptions

- f being L-Lipschitz places a constraint on the gradient of f
- Gradient and mirror descent converges in $\mathcal{O}(\frac{1}{t^{1/2}})$

- f is normally constrained on the gradient of its gradient, i.e. ∇f being L-Lipschitz, or

$$\|\nabla f(x) - \nabla f(y)\| \leq L(x - y)$$

- This achieves $\mathcal{O}(\frac{1}{t})$, as we will see in the ODE paper

Outline

1. Background
   - Gradient Descent and Mirror descent
   - Previous Work
2. Contribution of the ODE Paper
   - A Simple Example
   - Main Results
   - Basic Ideas for Proofs/Implementations

Work Leading up to This Paper

- Su et al.(2014)Su, Boyd, and Candes expressed Nesterov's accelerated method using discretized ODEs
- Allen-Zhu and Orecchia(2014) interprets Nesterov's accelerated method with mirror descent and generalized divergence

Work Leading up to This Paper

- Su et al.(2014)Su, Boyd, and Candes expressed Nesterov's accelerated method using discretized ODEs
- Allen-Zhu and Orecchia(2014) interprets Nesterov's accelerated method with mirror descent and generalized divergence

Outline

Simple ODE and Lyapunov example

- Vanilla gradient descent $x^{(k+1)} = x^{(k)} - s\nabla f\left(x^{(k)}\right)$ with a step size s can be rephrased as $\dot{X}(t) = -\nabla f(X(t))$ with discretization step s
- Taking the time derivative of Lyapunov function $V(X(t)) = \frac{1}{2}\left\|X(t) - x^{\star}\right\|^2$ gives convergence rate $\mathcal{O}(1/t)$, under the Lipschitz restriction on ∇f
- The mirror descent ODE formulation is a generalization, when we replace the Euclidean distance with the Bregman Divergence function

Simple ODE and Lyapunov example

- Vanilla gradient descent $x^{(k+1)} = x^{(k)} - s\nabla f\left(x^{(k)}\right)$ with a step size s can be rephrased as $\dot{X}(t) = -\nabla f(X(t))$ with discretization step s
- Taking the time derivative of Lyapunov function $V(X(t)) = \frac{1}{2}\left\|X(t) - x^{\star}\right\|^2$ gives convergence rate $\mathcal{O}(1/t)$, under the Lipschitz restriction on ∇f
- The mirror descent ODE formulation is a generalization, when we replace the Euclidean distance with the Bregman Divergence function

Simple ODE and Lyapunov example

- Vanilla gradient descent $x^{(k+1)} = x^{(k)} - s\nabla f\left(x^{(k)}\right)$ with a step size s can be rephrased as $\dot{X}(t) = -\nabla f(X(t))$ with discretization step s
- Taking the time derivative of Lyapunov function $V(X(t)) = \frac{1}{2}\|X(t) - x^{\star}\|^2$ gives convergence rate $\mathcal{O}(1/t)$, under the Lipschitz restriction on ∇f
- The mirror descent ODE formulation is a generalization, when we replace the Euclidean distance with the Bregman Divergence function

Simple ODE and Lyapunov example—continued

We use $\mathbb{E}$ to denote the primal space $\mathbb{R}^n$, and $\mathbb{E}^\star$ to denote the dual space. In this case, we replace the original Lyapunov $V(X(t)) = \frac{1}{2}\|X(t) - x^\star\|^2$ by a function on the dual space $V(Z(t)) = D_{\psi^*}(Z(t), z^\star)$, where $Z(t) \in \mathbb{E}^\star$ corresponds to $X(t) \in \mathbb{E}$, and ψ^* is a convex function defined on $\mathbb{E}^\star$ such that $\nabla\psi^* : \mathbb{E}^\star \mapsto \mathbb{E}$. Here the Bregman Divergence is defined as $D_{\psi^*}(Z(t), z^\star) = \psi^*(Z(t)) - (\psi^*(z^\star) + \nabla\psi^*(z^\star) \cdot (Z(t) - z^\star))$.

$$\begin{cases} X = \nabla\psi^*(Z) \\ \dot{Z} = -\nabla f(X) \\ X(0) = x_0, Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases}$$

Outline

ODE Formulation of Nesterov's Accelerated Method in Bregman Divergence

Combines Su et al.(2014)Su, Boyd, and Candes and Allen-Zhu and Orecchia(2014)

The desired lyapunov function is $V(X(t), Z(t), t) = \frac{t^2}{r}(f(X(t)) - f^\star) + rD_{\psi^*}(Z(t), z^\star)$. This gives us the proposed ODE system

$$\begin{cases} \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X) \\ \dot{Z} = -\frac{t}{r}\nabla f(X) \\ X(0) = x_0, Z(0) = z_0, \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases}$$

ODE Formulation of Nesterov's Accelerated Method in Bregman Divergence—continued

Combines Su et al.(2014)Su, Boyd, and Candes and Allen-Zhu and Orecchia(2014)

Theorem

Suppose that f has Lipschitz gradient, and that ψ^* is a smooth distance generating function. Let $(X(t), Z(t))$ be the unique solution to the accelerated mirror descent ODE with $r \geq 2$ Then for all $t > 0, f(X(t)) - f^\star \leq \frac{r^2 D_{\psi^*}(z_0, z^\star)}{t^2}$

Outline

Proof Idea of Solution and Convergence

- Obtain a smoothed form of the ODE by replacing t with $\max(t, \delta)$
- Prove convergence of smoothed ODE to original ODE with Arzela-Ascoli Theorem, then Cauchy-Lipschitz Theorem guarantees existent and unique solution
- By construction of the Lyapunov function

Proof Idea of Solution and Convergence

- Obtain a smoothed form of the ODE by replacing t with $\max(t, \delta)$
- Prove convergence of smoothed ODE to original ODE with Arzela-Ascoli Theorem, then Cauchy-Lipschitz Theorem guarantees existent and unique solution
- By construction of the Lyapunov function

Proof Idea of Solution and Convergence

- Obtain a smoothed form of the ODE by replacing t with $\max(t, \delta)$
- Prove convergence of smoothed ODE to original ODE with Arzela-Ascoli Theorem, then Cauchy-Lipschitz Theorem guarantees existent and unique solution
- By construction of the Lyapunov function

Summary

- We looked at the convergence rate of gradient and mirror descent under different assumptions
- We looked at the ODE formulation of Nesterov's accelerated method in Bregman divergence and its convergence properties

For Further Reading I

- Pan Xu, Tianhao Wang, and Quanquan Gu. [*Accelerated stochasticmirror descent: From continuous-time dynamics to discrete-time algorithms*]. In International Conference on Artificial Intelligence and Statistics, pages 1087 to 1096, 2018.
- Yujia Jin and Aaron Sidford. [*Efficiently solving mdps with stochastic mirror descent*]. arXiv preprint arXiv:2008.12776, 2020.