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

Qiyao Wei

September 2020

1 Vanilla Gradient Descent and its Convergence Rate

([Zhang(2019)])

Throughout this report we consider a convex function $f : \mathbb{R}^n \mapsto \mathbb{R}$, and assume that it is differentiable and L-Lipschitz, i.e. $\|\nabla f(x)\| \leq L$. It is quite natural to use gradient descent under such formulation.

If we let $x^* = \arg\min f(x)$, and start with $\|x_0 - x^*\| \leq d$, then we would choose gradient step size $\eta_t = \frac{d}{L\sqrt{(t)}}$, and do $x_{t+1} = x_t - \eta_t * \nabla f(x_t)$. The convergence rate for vanilla gradient descent is as follows.

Theorem 1.1 (Convergence of gradient descent) Let x_0 be such that $\|x_0 - x^*\| \leq d$. The gradient descent algorithm for T iterations starting at $(x_0, f(x_0))$ satisfies (proof omitted, but interested readers can always find this simple proof in the original lecture notes ([Zhang(2019)]))

$$f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) - f\left(x^*\right) \leq \frac{RL}{\sqrt{T}}$$

2 Bregman Divergence

([Zhang(2019)])

The most natural way I found to go from the familiar squared Euclidian Distance (SED and it also has many other names like L2 norm) towards Bregman Divergences is to simply note that the Bregman Divergence is a natural extension from SED, capturing all the Lp norms and much more. We refer to this source (http://mark.reid.name/blog/meet-the-bregman-divergences.html) for an interactive demo.

The Bregman Divergence is defined as

$$D_\Phi(a,b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a-b))$$

As an example, for $a \in \mathbb{R}^d$, the distance defined as $\Phi(a) = \frac{1}{2}\|a\|^2$ between a and b is the same as SED

$$\begin{aligned} D_\Phi(a,b) &= \frac{1}{2}\|a\|^2 - \left(\frac{1}{2}\|b\|^2 + b \cdot (a-b)\right) \\ &= \frac{1}{2}\|a-b\|^2 \end{aligned}$$

Since the introduction part of the ODE paper touches on mirror descent in a more detailed way, I am omitting the last part of the lecture notes in ([Zhang(2019)]). Moving on to the ODE paper will give us more than enough understanding of mirror descent.

3 Important Key Assumptions

The analysis in ([Zhang(2019)]) tells us that vanilla gradient descent, as well as mirror descent under the same assumptions of ”f being L-Lipschitz”, converges in $\mathcal{O}(\frac{1}{t^{1/2}})$. This is only under the simplest assumption. The vast majority of literature in this field, including the paper we will be seeing next, requires a more strict assumption, namely that ”f being L-smooth”. We will explicitly define this assumption later in this report, but it is vital to draw comparison here emphasizing the difference between ”f being L-Lipschitz”, which restricts the gradient of f, and ”f being L-smooth”, which restricts the gradient of the gradient of f. ”f being L-smooth” is also often referred to as ”∇ f being Lipschitz”, hence the confusion.

4 Vanilla Gradient Descent and Mirror Descent Under Lyapunov Arguments

([Krichene et al.(2015)Krichene, Bayen, and Bartlett])

The fact that many machine learning and optimization problems can be characterized using ODE equations is well-known. As a simple 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.

In order to prove the convergence rate of vanilla gradient descent, we consider Lyapunov arguments. For example, for the simple ODE $\dot{X}(t) = -\nabla f(X(t))$, we define a Lyapunov function $V(X(t)) = \frac{1}{2}\left\|X(t) - x^\star\right\|^2$. Then

$$\frac{d}{dt}V(X(t)) = \left\langle \dot{X}(t), X(t) - x^\star \right\rangle = \left\langle -\nabla f(X(t)), X(t) - x^\star \right\rangle \leq -\left(f(X(t)) - f^\star\right)$$

Where the final inequality is due to the convexity of f. The convergence we want to prove, expressed as

$$\begin{aligned} f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^\star &\leq \frac{1}{t}\int_0^t f(X(\tau))d\tau - f^\star \\ &\leq \frac{V\left(x_0\right) - V(X(t))}{t} \\ &\leq \frac{V\left(x_0\right)}{t} \\ &= \mathcal{O}(\frac{1}{t}) \end{aligned}$$

Where the second inequality is because integrating

$$\frac{d}{dt}V(X(t)) \leq -\left(f(X(t)) - f^\star\right) V(X(t)) - V\left(x_0\right) \leq tf^\star - \int_0^t f(X(\tau))d\tau$$

We now apply the logic above to the mirror descent formulation. For simplicity of notation, 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 $V(X(t)) = \frac{1}{2}\left\|X(t) - x^\star\right\|^2$ by a function on the dual space $V(Z(t)) = D_{\psi^*}\left(Z(t), z^\star\right)$, 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 $\nabla\psi^* : \mathbb{E}^\star \mapsto \mathbb{E}$. Just to reiterate, the Bregman Divergence is defined as $D_{\psi^*}\left(Z(t), z^\star\right) = \psi^*(Z(t)) - (\psi^*(z^\star) + \nabla\psi^*(z^\star) \cdot (Z(t) - z^\star))$.

$$\begin{aligned} \frac{d}{dt}V(Z(t)) = \frac{d}{dt}D_{\psi^*}\left(Z(t), z^\star\right) &= \frac{d}{dt}\left(\psi^*(Z(t)) - \psi^*\left(z^\star\right) - \left\langle\nabla\psi^*\left(z^\star\right), Z(t) - z^\star\right\rangle\right) \\ &= \left\langle \nabla\psi^*(Z(t)) - \nabla\psi^*\left(z^\star\right), \dot{Z}(t)\right\rangle = \left\langle X(t) - x^\star, \dot{Z}(t)\right\rangle \end{aligned}$$

where the derivatives wrt t for $z^\star$ is of course zero. Therefore, if the dual variable Z obeys the dynamics $\dot{Z} = -\nabla f(X)$, then

$$\frac{d}{dt} V(Z(t)) = -\langle \nabla f(X(t)), X(t) - x^\star \rangle \leq -(f(X(t)) - f^\star)$$

and by the same argument as before, $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^\star$ converges to 0 at a $\mathcal{O}(1/t)$ rate. . We summarize the mirror descent system with

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

Just as a quick example, if we take $\psi^*(Z) = \frac{1}{2}\|z\|^2$, then $\nabla\psi^*(Z)$ is the identity, X and Z coincide, and we retrieve vanilla gradient descent.

5 Now the ODE Paper

Aside from the background in the previous section, the ODE paper also mentions Nesterov's accelerated method, which provably converges in $\mathcal{O}(\frac{1}{t^2})$. In terms of Lyapunov arguments, the hard work has already been done for us in expressing it as a second-order ODE ([Su et al.(2014)Su, Boyd, and Candes]), so by choosing the proper Lyapunov function $\mathcal{E}(t) = \frac{t^2}{r}(f(X) - f^\star) + \frac{r}{2}\left\|X + \frac{t}{r}\dot{X} - x^\star\right\|^2$, one can prove convergence in $\mathcal{O}(\frac{1}{t^2})$. However, this Lyapunov function is only defined for the Euclidean norm.

We can now appreciate the gap this paper is trying to bridge. ([Su et al.(2014)Su, Boyd, and Candes]) only talks about the ODE formulation of Nesterov's accelerated method in the Euclidean space. Therefore, this paper takes that one step further, by extending the ODE formulation of Nesterov's accelerated method to the general space of Bregman Divergences with mirror descent. Notably, the only thing that changed between the two papers is the continuous time formulation, since the discretization technique in them is identical.

We define the assumptions here. This paper assumes $\psi^\star$ is L-smooth. For function f to be L-smooth wrt a reference norm $\|\cdot\|_*$, we must have $D_f(z,y) \leq \frac{L}{2}\|z-y\|_*^2$. Referred to in papers like ([Allen-Zhu and Orecchia(2014)]), L-smooth is also defined for a function f as $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$. It is worth noting that this is equivalent to our inverse mapping function $\nabla\psi^\star$ being L-Lipschitz.

The desired lyapunov function is $V(X(t), Z(t), t) = \frac{t^2}{r}(f(X(t)) - f^\star) + rD_{\psi^*}(Z(t), z^\star)$. With the same computation as before, we would have the proposed ODE system

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

In terms of Euclidean norm, taking $\psi^*(z) = \frac{1}{2}\|z\|^2$, we have $\nabla\psi^*(z) = z$, thus $Z = X + \frac{t}{r}\dot{X}$, and the ODE system is equivalent to $\frac{d}{dt}\left(X + \frac{t}{r}\dot{X}\right) = -\frac{t}{r}\nabla f(X)$, which is equivalent to the ODE (2) studied in ([Su et al.(2014)Su, Boyd, and Candes]), which we recover as a special case.

It is now straightforward to establish the convergence rate of the solution.
Theorem 2. Suppose that f has Lipschitz gradient, and that ψ^* is a smooth distance generating function. Let $(X(t), Z(t))$ be the solution to the accelerated mirror descent ODE (5) with $r \geq 2$ Then for all $t > 0, f(X(t)) - f^\star \leq \frac{r^2 D_{\psi^*}(z_0, z^\star)}{t^2}$

Proof. By construction of the ODE, $V(X(t), Z(t), t) = \frac{t^2}{r}\left(f(X(t)) - f^\star\right) + rD_{\psi^*}\left(Z(t), z^\star\right)$ is a Lyapunov function. It follows that for all $t > 0$, $\frac{t^2}{r}\left(f(X(t)) - f^\star\right) \leq V(X(t), Z(t), t) \leq V\left(x_0, z_0, 0\right) = rD_{\psi^*}\left(z_0, z^\star\right)$

We then discretize the mirror descent system for more general applications. We proceed with the following "mixed forward-backward Euler scheme", by taking a step size of $\sqrt{s}$, letting $t_k = k\sqrt{s}$, and defining $x^{(k)} = X\left(t_k\right) = X(k\sqrt{s})$ (expanding on the discussion of discretization, one can refer to [Xu et al.(2018)Xu, Wang, and Gu] for 3 different discretization techniques, applicable to deterministic and stochastic mirror descent):

Algorithm 1 Accelerated mirror descent with distance generating function ψ^*, regularizer R, step size s, and parameter $r \geq 3$

```
Initialize $\tilde{x}^{(0)} = x_0, \tilde{z}^{(0)} = x_0,\left(\text{ or } z^{(0)} \in (\nabla\psi)^{-1}\left(x_0\right)\right)$
for $k \in \mathbb{N}$ do
    $x^{(k+1)} = \lambda_k \tilde{z}^{(k)} + \left(1 - \lambda_k\right) \tilde{x}^{(k)}$, with $\lambda_k = \frac{r}{r+k}$
    $\tilde{z}^{(k+1)} = \arg\min_{\tilde{z} \in \mathcal{X}} \frac{ks}{r}\left\langle \nabla f\left(x^{(k+1)}\right), \tilde{z}\right\rangle + D_\psi\left(\tilde{z}, \tilde{z}^{(k)}\right)$
(If $\psi$ is non-differentiable, $z^{(k+1)} = z^{(k)} - \frac{kr}{s}\nabla f\left(x^{(k+1)}\right)$ and $\tilde{z}^{(k+1)} = \nabla\psi^*\left(z^{(k+1)}\right)$.)
    $\tilde{x}^{(k+1)} = \arg\min_{\tilde{x} \in \mathcal{X}} \gamma s\left\langle \nabla f\left(x^{(k+1)}\right), \tilde{x}\right\rangle + R\left(\tilde{x}, x^{(k+1)}\right)$
```

Theorem 3. The discrete-time accelerated mirror descent Algorithm 1 with parameter $r \geq 3$ and step sizes $\gamma \geq L_R L_{\psi^*}, s \leq \frac{\ell_R}{2L_f\gamma}$, guarantees that for all $k > 0$

$$f\left(\tilde{x}^{(k))}\right) - f^\star \leq \frac{r}{sk^2}\tilde{E}^{(1)} \leq \frac{r^2 D_{\psi^*}\left(z_0, z^\star\right)}{sk^2} + \frac{f\left(x_0\right) - f^\star}{k^2}$$

The paper also provides an example using KL-divergence. However, it breezes through the procedure with hasty references.

As a conclusion, we have given a gentle introduction to gradient descent and mirror descent in its Lyapunov formulation. Further investigation into the nature of Lyapunov arguments and Nesterov's accelerated method is required in order to better understand the details of this paper.

References

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

[Krichene et al.(2015)Krichene, Bayen, and Bartlett] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015.

[Su et al.(2014)Su, Boyd, and Candes] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in neural information processing systems*, pages 2510–2518, 2014.

[Xu et al.(2018)Xu, Wang, and Gu] 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, 2018.

[Zhang(2019)] Fred Zhang. Mirror descent and online learning. 2019.