Review on Nestorov's Accelerated Gradient Method and Mirror Descent

Jiawei Kyle Zhang

Abstract

In this report, we will discuss two optimization method: Nestorov's accelerated gradient method and mirror descent, from several ways. We will first present the problem settings each algorithm tends to address and then show the main process of each one. For each algorithm, we will show the proof of convergence and also discuss some aspects of the convergence results and developments.

1 Nestorov's Accelerated Gradient Method

1.1 Introduction

The classic gradient descent update step is defined as $x_t = x_{t-1} - \alpha \nabla f(x_{t-1})$, where the α is constant. In class and homework, we discussed some methods that using line search to choose a better α at each step to speed up the convergence. The Nestorov's Accelerated Gradient and the momentum method group it belongs to are other modifications that introduce another variable and update both variables at each time.

Nestorov's Accelerated Gradient is originally proposed by Yurii Nesterov in 1983 and considered to have the fastest convergence rate. However, the intuition behind it is something hard to interpret at that time. As a group of gradient method called momentum are proposed, some scientists gave their insights about it that it can be regarded as a momentum method. We first give the original definition according to Bubeck [2013].

1.2 Definition

Definition 1. *(Nestorov's Accelerated Gradient Method, 1983)*

Suppose f is convex and ∇f is L-Lipshitz. Define ε_{-1} an arbitrary positive number. To obtain ε_t for $t > -1$, repeatedly mutiply ε_{t-1} by a factor $\beta < 1$, i.e. $\varepsilon_t = \beta^i \cdot \varepsilon_{t-1}$ until ε_t satisfies:

$$f(x_t - \varepsilon_t \nabla f(x_t)) \leq f(x_t) - \frac{1}{2}\varepsilon_t \parallel \nabla f(x_t) \parallel^2 \tag{1.1}$$

Starting with $\lambda_0 = 0$, and an arbitrary initial point $y_0 = x_0$, update the following equations

repeatedly to optimize $f(y)$:

$$\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2} \tag{1.2}$$

$$\gamma_t = \frac{1-\lambda_2}{\lambda_{t+1}} \tag{1.3}$$

$$x_{t+1} = y_t - \varepsilon_t \nabla f(y_t) \tag{1.4}$$

$$y_{t+1} = (1-\gamma_t)x_{t+1} + \gamma_t x_t \tag{1.5}$$

.

As shown above, this process iteratively updates two variabels: x_t and y_t, where x_t is the target that obtained from original gradient descent, and y_t is the 'real' target point that updated by the combination of current and previous gradients.

1.3 Convergence Analysis

Nestorov's accelerated gradient method is proven to converge at a rate of $O(1/t^2)$.

Theorem 1. *If f is convex and L-smooth, Nestorov's accelerated gradient method satisfies:*

$$f(x_s) - f(x^*) \leq \frac{2L \parallel x_1 - x^* \parallel^2}{t^2} \tag{1.6}$$

when choose $\varepsilon_{-1} = \frac{1}{L}$.

Proof. Consider following equations for any x and y:

$$\begin{aligned} f(y-\varepsilon\nabla f(y)) - f(x) &= f(y-\varepsilon\nabla f(y)) - f(y) + f(y) - f(x) \\ &\leq -\frac{1}{2}\varepsilon \parallel \nabla f(y) \parallel^2 + \nabla^T f(y)(y-x) \end{aligned} \tag{1.7}$$

Let $y = y_t, x = x_t, \varepsilon = \varepsilon_t$, we have

$$\begin{aligned} f(y_t - \varepsilon_t\nabla f(y_t)) - f(x_t) &= f(x_{t+1}) - f(x_t) \\ &\leq -\frac{1}{2}\varepsilon_t \parallel \nabla f(y_t) \parallel^2 + \nabla^T f(y_t)(y_t - x_t) \end{aligned} \tag{1.8}$$

Rewrite the update equation ((1.4)), we have $\nabla f(y_t) = \frac{1}{\varepsilon_t}(y_t - x_{t+1})$. Thus, we get the following inequality:

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\varepsilon_t} \parallel x_{t+1} - y_t \parallel^2 - \frac{1}{\varepsilon_t}(x_{t+1} - y_t)^T(y_t - x_t) \tag{1.9}$$

Let $y = y_t, x = y^*, \varepsilon = \varepsilon_t$, we similarly have:

$$f(x_{t+1}) - f(y^*) \leq -\frac{1}{2\varepsilon_t} \parallel x_{t+1} - y_t \parallel^2 - \frac{1}{\varepsilon_t}(x_{t+1} - y_t)^T(y_t - y^*) \tag{1.10}$$

Then, from ((1.9))$\times(\lambda_t - 1)$+((1.10)), we get

$$\begin{aligned} \lambda_t f(x_{t+1}) - (\lambda_t - 1)f(x_t) - f(y^*) &= \lambda_t(f(x_{t+1}) - f(y^*)) - (\lambda_t - 1)(f(x_t) - f(y^*)) \\ &\leq -\frac{\lambda_t}{2\varepsilon_t} \parallel x_{t+1} - y_t \parallel^2 - \frac{1}{\varepsilon_t}(x_{t+1} - y_t)^T(\lambda_t y_t - (\lambda_t - 1)x_t - y^*) \end{aligned} \tag{1.11}$$

Notice following property of λ, from ((1.2)) we have

$$\begin{aligned} (2\lambda_t - 1)^2 &= 1 + 4\lambda_{t-1}^2 \\ \lambda_t^2 - \lambda_t &= \lambda_{t-1}^2 \end{aligned} \tag{1.12}$$

Multiply ((1.11)) by λ_t and apply ((1.12)), we get

$$\begin{aligned} & \lambda_t^2(f(x_{t+1}) - f(y^*)) - \lambda_t(\lambda_t - 1)(f(x_t) - f(y^*)) \\ \leq & -\frac{1}{2\varepsilon_t}(\| \lambda_t(x_{t+1} - y_t) \|^2 + 2\lambda_t(x_{t+1} - y_t)^T(\lambda_t y_t - (\lambda_t - 1)x_t - y^*)) \\ = & -\frac{1}{2\varepsilon_t}(\| \lambda_t(x_{t+1} - y_t) + (\lambda_t y_t - (\lambda_t - 1)x_t - y^*) \|^2 \\ & - \| \lambda_t y_t - (\lambda_t - 1)x_t - y^* \|^2) \quad \text{(Complete the square)} \\ = & -\frac{1}{2\varepsilon_t}(\| \lambda_t x_{t+1} - (\lambda_t - 1)x_t - y^* \|^2 - \| \lambda_t y_t - (\lambda_t - 1)x_t - y^* \|^2) \end{aligned} \tag{1.13}$$

Notice γ's definition ((1.3)) and y_t's update equation, we have

$$\begin{aligned} & y_{t+1} = (1 - \gamma_t)x_{t+1} + \gamma_t x_t \\ \Leftrightarrow\ & y_{t+1} = x_{t+1} + \gamma_t(x_t - x_{t+1}) \\ \Leftrightarrow\ & \lambda_{t+1}y_{t+1} = \lambda_{t+1}x_{t+1} + (1 - \lambda_t)(x_t - x_{t+1}) \\ \Leftrightarrow\ & \lambda_{t+1}y_{t+1} - (\lambda_{t+1} - 1)x_{t+1} = \lambda_t x_{t+1} - (\lambda_t - 1)x_t \end{aligned} \tag{1.14}$$

Denote $f(x_t) - f(y^*)$ as δ_t, $\lambda_t x_{t+1} - (\lambda_t - 1)x_t - y^*$ as u_t and apply ((1.14)) to ((1.13)), we have

$$\begin{aligned} \lambda_t^2\delta_{t+1} - \lambda_{t-1}^2\delta_t &\leq \frac{1}{2\varepsilon_t}(\| u_t \|^2 - \| u_{t+1} \|^2) \\ &\leq \frac{1}{2\varepsilon_{best}}(\| u_t \|^2 - \| u_{t+1} \|^2) \end{aligned} \tag{1.15}$$

where ε_{best} is the smallest ε among $\{\varepsilon_0, \varepsilon_1, ..., \varepsilon_t\}$. From the Descent Lemma, we know that

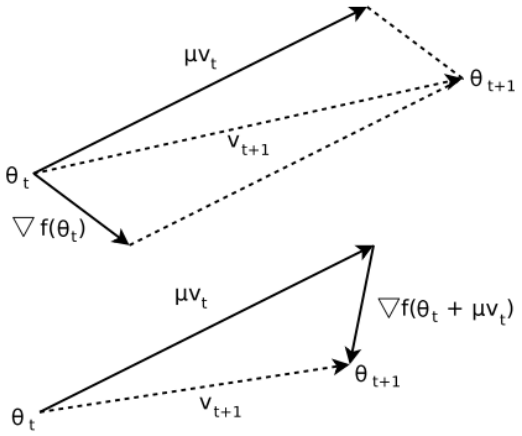
$$f(x_t - \varepsilon_t \nabla f(x_t)) \leq f(x_t) - \varepsilon_t \| \nabla f(x_t) \|^2 + \frac{L}{2}\varepsilon_t^2 \| \nabla f(x_t) \|^2$$

When $\varepsilon \leq \frac{1}{L}$, since $\frac{L}{2}\varepsilon_t^2 \leq \frac{\varepsilon_t}{2}$, the stop condition for ε is always satisfied. So if ε_{-1} is smaller than $\frac{1}{L}$, ε doesn't change and $\varepsilon_{best} = \varepsilon_{-1}$. If ε_{-1} is greater than $\frac{1}{L}$, any future ε_t is at least greater than $\frac{\beta}{L}$. Thus, the inequality ((1.15)) now can be written as

$$\begin{aligned} \lambda_t^2\delta_{t+1} - \lambda_{t-1}^2\delta_t &\leq \frac{1}{2\varepsilon_{best}}(\| u_{t-1} \|^2 - \| u_t \|^2) \\ &\leq \frac{1}{2min\left\{\varepsilon_{-1}, \frac{\beta}{L}\right\}}(\| u_{t-1} \|^2 - \| u_t \|^2) \end{aligned}$$

Sum over from $t = 1$ to $t = t - 1$, we have

$$\begin{aligned} \lambda_{t-1}^2\delta_t - \lambda_0^2\delta_1 &\leq \frac{1}{2min\left\{\varepsilon_{-1}, \frac{\beta}{L}\right\}}(\| u_1 \|^2 - \| u_t \|^2) \\ \Longrightarrow \lambda_{t-1}^2\delta_t &\leq \frac{1}{2min\left\{\varepsilon_{-1}, \frac{\beta}{L}\right\}} \| u_1 \|^2 \quad (\lambda_0 = 0) \end{aligned} \tag{1.16}$$



(a) **(top)** Momentum **(bottom)** Nesterov's accelerated gradient

Figure 1

Notice $\lambda_t \geq \frac{1}{2}(t+1)$ and $x^* = y^*$, we get the following result:

$$f(x_t) - f(x^*) \leq \frac{2 \parallel x_1 - x^* \parallel^2}{min\left\{\varepsilon_{-1}, \frac{\beta}{L}\right\} t^2} \tag{1.17}$$

If we know f is L-smooth, we just choose $\varepsilon_{-1} = \frac{1}{L}$, and the Theorem 1 is proven true. Equation ((1.17)) shows that the Nestorov's Accelrated Gradient Method converges at a rate of $O(\frac{1}{t^2})$. □

1.4 Development

In later years, momentum gradient methods are proposed. We first introduce the original momentum gradient descent method.

Definition 2. *(Momentum Gradient Descent)*

Different from classical gradient descent, the momentum gradient descent defines a momentum velocity as:

$$v_t = \mu_{t-1} v_{t-1} - \epsilon_t \nabla f(x_t) \tag{1.18}$$

and the x_t is updated by the current ∇f and previous momentum v_{t-1}:

$$x_t = x_{t-1} + v_t \tag{1.19}$$

At each gradient descent step, we both iteratively update the (1.18) and the (1.19). It is called momentum because the x_t is updated not only by the ∇f but also the previous 'accumulated' v_{t-1}s. This effect is pretty like the momentum in kinematics, where the position is updated by the velocity (momentum) and the velocity (momentum) is updated by the force.

Some later work showed that the Nestorov's Accelerated Gradient could also be regarded as a particular momentum gradient (e.g. Ruder [2016]).

The most famous development is from Ilya Sutskever, who is now the OpenAI cheif scientist and became much more famous recently. In his PhD paper, he gave his insight of Nestorov's Accelerated Gradient.

Definition 3. *(Sutskever Momentum, Sutskever [2013])*

Define new variables like follows:

$$\begin{aligned} v_t &:= x_{t+1} - x_t \\ \mu_t &:= -\gamma_t \end{aligned}$$

Thus, the equation (1.5) *can be rewritten as* $y_{t+1} = x_{t+1} + \mu_t v_t$. *The Nestorov's Accelerated Gradient update equations* (1.4) *can be then rewritten as*

$$x_{t+1} = x_t + \mu_{t-1} v_{t-1} - \epsilon_t \nabla f(x_t + \mu_{t-1} v_{t-1})$$

Rephrasing the equation and combining with (1.20)*, the Sutskever momentum method is*

$$v_t = \mu_{t-1} v_{t-1} - \epsilon_t \nabla f(x_t + \mu_{t-1} v_{t-1}) \tag{1.20}$$

$$x_t = x_{t-1} + v_t \tag{1.21}$$

The update equation form of Sutskever momentum really looks like the original momentum gradient method. So nowadays it is widely accepted as a form of momentum.

Sutskever's insight here is, the key difference between Nestorov's accelerated gradient and momentum gradient method is that Nestorov's accelerated gradient does not move along the current derivative direction, but move towards the derivative of approximate 'next' point, which makes it change more responsive. As illustrated in fig. 1. More momentum forms are discussed in this website Ville [2016].

2 Mirror Descent

2.1 Intuition

When developing Projected Gradient Descent, we have the following update step:

$$x_{n+1} = [x_n - \alpha \nabla f(x_n)]^+$$

where $[\delta]^+$ is the solution to $\min_{y \in S} \frac{1}{2} ||\delta - y||^2$. Substitute y with x_{n+1}, we can get following equation:

$$\begin{aligned} x_{n+1} &= \min_{x \in S} ||x - (x_n - \alpha \nabla f(x_n))||^2 \\ &= \min_{x \in S} ||x - x_n + \alpha \nabla f(x_n)||^2 \\ &= \min_{x \in S} \{||x - x_n||^2 + 2\alpha \nabla^T f(x_n)(x - x_n) + M_n\} \\ &= \min_{x \in S} \{\nabla^T f(x_n)(x - x_n) + \frac{1}{2\alpha} ||x - x_n||^2\} \end{aligned} \tag{2.1}$$

where M_n is independent of x. Observing the equation, the second term is the Euclidean distance between x and x_n. A natural thinking is what if we use other distances to better the convergence when dealing with set S with special geometry. (The intuition part is mainly from Tlienart [2021], and following contents are mainly from CMU [2020])

2.2 Proximal Point View

Definition 4. *(Bregman Divergence)*

The Bregman Divergence from x to y w.r.t. a strictly convex function h is defined as

$$D_h(y||x) = h(y) - h(x) - \nabla^T h(x)(y - x) \tag{2.2}$$

By choosing different types of $h(x)$, we can get different Bregman divergence. Two typical examples are:

1. When $h(x) = \frac{1}{2}||x||^2$, the corresponding Bregman divergence is

$$D_h(y||x) = \frac{1}{2}||y - x||^2$$

2. When $h(x) = \sum x_i \ln x_i - x_i$, the corresponding Bregman divergence is

$$D_h(y||x) = \sum (y_i \ln \frac{y_i}{x_i} - y_i + x_i)$$

When $\sum x_i = 1$ and $\sum y_i = 1$, we further get $D_h(x||y) = \sum y_i \ln \frac{y_i}{x_i}$, which is called KL-divergence.

Now, we replace the 2-Norm term in (2.1) with Bregman divergence, i.e., using Bregman divergence to approximate the distance to the current point x_n, which comes to the proximal point view.

Definition 5. *(Proximal Point View on Mirror Descent)*

For a unconstrained optimization problem, replace the 2-Norm term in (2.1) *with Bregman divergence, we have following gradient descent method:*

$$\begin{aligned} & x_{t+1} = \min_x \{\nabla^T f(x_t)(x - x_t) + \frac{1}{\alpha} D_h(x||x_t)\} \\ \Leftrightarrow \quad & 0 = \alpha \nabla^T f(x_t) + \nabla h(x_{t+1}) - \nabla h(x_t) \\ \Leftrightarrow \quad & x_{t+1} = \nabla h^{-1}(\nabla h(x_t) - \alpha \nabla^T f(x_t)) \end{aligned} \tag{2.3}$$

When dealing with constrained optimization problem, we can just make

$$\begin{aligned} x'_{t+1} &= (\nabla h)^{-1}(\nabla h(x_t) - \alpha \nabla^T f(x_t)) \\ x_{t+1} &= \min_{x \in S} D_h(x||x'_{t+1}) \end{aligned} \tag{2.4}$$

Here I give a simple proof on why 2.4 is valid.

Proof.

$$\begin{aligned} x_{t+1} &= \min_{x \in S} D_h(x||x'_{t+1}) \\ &= \min_{x \in S} \{h(x) - h(x'_{t+1}) - \nabla^T h(x_{t+1})(x - x'_{t+1})\} \\ &= \min_{x \in S} \{h(x) - h(x'_{t+1}) - (\nabla h(x_t) - \alpha \nabla f(x_t))^T (x - x'_{t+1})\} \qquad (2.5) \\ &= \min_{x \in S} \{h(x) - (\nabla h(x_t) - \alpha \nabla f(x_t))^T x\} \quad \text{(ignore some terms independent of x)} \\ &= \min_{x \in S} \{h(x) - h(x_t) - \nabla h(x_t)(x - x_t) + \alpha \nabla f(x_t)^T (x - x_t)\} \\ &= \min_{x \in S} \{\nabla^T f(x_t)(x - x_t) + \frac{1}{\alpha} D_h(x||x_t)\} \end{aligned}$$

which is exactly what mirror descent trying to optimize. The (2.5) line is because the $\nabla h(x)$ and $(\nabla h)^{-1}(x)$ cancel out. □

2.3 Mirror Map View

In this section, we discuss the Mirror Map view, which is the name Mirror Descent comes from.

Definition 6. *(Mirror Map View on Mirror Descent)*

To minimize a convex function f over a convex set S, we first choose a differentiable strictly convex function $h : \mathbb{R}^n \longrightarrow \mathbb{R}$, which gives us the mirror map $\nabla h : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Starting with x_0, the mirror descent algorithm is following:

1. *Map x_t to dual space: $\theta_t \longleftarrow \nabla h(x_t)$.*
2. *Do gradient descent in dual space: $\theta_{t+1} \longleftarrow \theta_t - \alpha \nabla f(x_t)$.*
3. *Map back to the prime space: $x_{t+1}^{'} \longleftarrow (\nabla h)^{-1}(\theta_{t+1})$.*
4. *If $x_{t+1}^{'}$ out of set S, map it back using Bregman divergence: $x_{t+1} = \min_{x \in S} D_h(x||x_{t+1}^{'})$.*

The Mirror Map view provides another interpretation of the update equation 2.4 proposed in Proximal Point view. It decomposes the whole update equation into several steps: using mirror map to map between the **dual space** (where we move along the derivative direction) and **primal space** (where the optimization object exists).

2.4 Convergence Analysis

Before the convergence analysis, the dual norm has to be introduced.

Definition 7. *(Dual Norm)*

In class, we have already discussed the definition of general norms. Let $||\cdot||$ is a norm, the corresponding dual norm $||\cdot||_*$ is defined as:

$$||y||_* := \sup\{x^T y : ||x|| < 1\}$$

Theorem 2. *Suppose $||\cdot||$ is a norm, its dual norm then is $||\cdot||_*$. h is m-strongly convex. For all x_t, $||\nabla f(x_t)|| \leq G$. The Mirror descent satisfies:*

$$\sum_{t=1}^{n} f(x_t) \leq \sum_{t=1}^{n} f(x^*) + \frac{D_h(x^*||x_1)}{\alpha} + \frac{\alpha \sum_{t=1}^{n} ||\nabla f(x_t)||_*^2}{2m} \tag{2.6}$$

Before the proof, we introduce the generalized Cauchy-Schwarz inequality for any norms as a lemma.

Lemma 1. *(Generalized Cauchy-Schwarz) For any x and y, $x^T y \leq ||x|| \cdot ||y||_*$*

Proof. When $||x|| = 0$, it is obvious. For $||x|| \neq 0$, $||\frac{x}{||x||}|| = \frac{1}{||x||}||x|| = 1$. From the definition of dual norm, we have $||y||_* \geq (\frac{x}{||x||})^T y$, which finishes the proof. □

Now we start the proof for the Theorem 2.

Proof. Denote $\Phi_t = \frac{D_h(x^*||x_t)}{\alpha}$. We first examine the distance between Φ_{t+1} and Φ_t.

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \frac{1}{\alpha}(D_h(x^*||x_{t+1}) - D_h(x^*||x_t)) \\
&= \frac{1}{\alpha}(h(x^*) - h(x_{t+1}) - \nabla^T h(x_{t+1})(x^* - x_{t+1}) \\
&\qquad -h(x^*) + h(x_t) + \nabla^T h(x_t)(x^* - x_t)) \\
&= \frac{1}{\alpha}(h(x_t) - h(x_{t+1}) - (\nabla h(x_t) - \alpha\nabla f(x_t))^T(x^* - x_{t+1}) \\
&\qquad +\nabla h^T(x_t)(x^* - x_t)) \\
&= \frac{1}{\alpha}(h(x_t) - h(x_{t+1}) - \nabla^T h(x_t)(x_t - x_{t+1}) + \alpha(\nabla^T(x_t)(x^* - x_{t+1}))) \\
&\leq \frac{1}{\alpha}(-\frac{m}{2}||x_{t+1} - x_t||^2 + \alpha\nabla^T f(x_t)(x^* - x_{t+1})) \qquad (2.7)
\end{aligned}$$

The last inequality (2.7) comes from the definition of strong convexity for any norm: $f(x)$ is m-strong convex, if and only if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) + m||x-y||^2$. This definition is different from the form that the professor gave in class ($\nabla f \succeq mI$ w.r.t. any norms), but I have not come up with a way to show the equivalence of them.

Putting (2.7) together with $f(x_t) - f(x^*)$, we have

$$\begin{aligned}
& f(x_t) - f(x^*) + (\Phi_{t+1} - \Phi t) \\
\leq\ & f(x_t) - f(x^*) + \nabla^T f(x_t)(x^* - x_{t+1}) - \frac{m}{2\alpha}||x_{t+1} - x_t||^2 \\
\leq\ & f(x_t) - f(x^*) + \nabla^T f(x_t)(x^* - x_t) - \frac{m}{2\alpha}||x_{t+1} - x_t||^2 + \nabla^T f(x_t)(x^t - x_{t+1}) \\
\leq\ & -\frac{m}{2\alpha}||x_{t+1} - x_t||^2 + \nabla^T f(x_t)(x^t - x_{t+1}) \quad \text{(because of the convexity of } f) \\
\leq\ & -\frac{m}{2\alpha}||x_{t+1} - x_t||^2 + ||x_t - x_{t+1}|| \cdot ||\nabla f(x_t)||_* \quad \text{(from the lemma 1)} \\
\leq\ & -\frac{m}{2\alpha}||x_{t+1} - x_t||^2 + \frac{1}{2}(\frac{\alpha}{m}||\nabla f(x_t)||_*^2 + \frac{m}{\alpha}||x_t - x_{t+1}||^2) \text{(by the AM-GM inequality)} \\
\leq\ & \frac{\alpha}{2m}||\nabla f(x_t)||_*^2 \qquad (2.8)
\end{aligned}$$

After getting the inequality (2.8), we are ready to telescope from $t = 1$ to $t = n$:

$$\begin{aligned}
\sum_{t=1}^{n} f(x_t) - \sum_{t=1}^{n} f(x^*) &\leq \Phi_1 - \Phi_n + \frac{\alpha}{2m}\sum_{t=1}^{n}||\nabla f(x_t)||_*^2 \\
&\leq \Phi_1 + \frac{\alpha}{2m}\sum_{t=1}^{n}||\nabla f(x_t)||_*^2 \qquad (2.9)
\end{aligned}$$

Now, we nearly proved the theorem. In fact, there is still a missing part need to be proved to finish the proof on constraint optimization cases: $D_h(x_{t+1}||x^*) \leq D_h(x'_{t+1}||x^*)$. This can be seen as the generalized "Projection not expansion" property. However, I have not come up with it or found enough resources.

To further show the convergence, we can use the two techniques we discussed in class (i.e., updating the average of x_i or defining the x_{best}) to get a convergence result. Say we use the first method:

$$
\begin{aligned}
f(\frac{1}{n}\sum_{t=1}^{n} x_t) - f(x^*) &\leq \frac{D_h(x^*||x_1)}{\alpha n} + \frac{\alpha}{2mn}\sum_{t=1}^{n} ||\nabla f(x_t)||_*^2 \\
&\leq \frac{D_h(x^*||x_1)}{\alpha n} + \frac{\alpha G^2}{2m} \qquad (2.10)
\end{aligned}
$$

To ensure the error go to zero as the n goes to infinite, we can choose α satisfying:

$$
\begin{aligned}
&-\frac{1}{\alpha^2}\frac{D_h(x^*||x_1)}{n} + \frac{G^2}{2m} = 0 \\
&\Longrightarrow \alpha = \sqrt{\frac{2mD_h(x^*||x_1)}{nG^2}}
\end{aligned}
$$

Write back α to (2.10):

$$
RHS = 2\sqrt{\frac{G^2 D_h(x^*||x_1)}{2mn}} \qquad (2.11)
$$

which goes to 0. □

From the result, we can know the mirror descent normally converges at the rate of $O(\frac{1}{\sqrt{t}})$. But the $D_h(x^*||x_1)$ term provides us a way to improve the performance: we can choose better D_h to reduce the distance between x_* and x_1 under the Bregman divergence measurement.

3 Acknowledgements

It's been a great journey taking this class. I surprisingly found myself gradually able to understand many algorithms while reading papers and learning other subjects like machine learning in this semester. Sincerest thanks for Professor Srikant and our TAs.

References

Sebastien Bubeck. Orf523: Nesterov's accelerated gradient descent, 2013. URL https://web.archive.org/web/20210302210908/https://blogs.princeton.edu/imabandit/2013/04/01/acceleratedgradientdescent/. Accessed on December 8, 2023.

CMU. 17: Mirror descent, 2020. URL http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15850-f20/www/notes/lec19.pdf. Accessed on December 8, 2023.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Ilya Sutskever. *Training recurrent neural networks*. University of Toronto Toronto, ON, Canada, 2013.

Tlienart. Mirror descent, 2021. URL https://tlienart.github.io/posts/2018/10/27-mirror-descent-algorithm/. Accessed on December 8, 2023.

Jlmel Ville. Nesterov accelerated gradient and momentum, 2016. URL https://jlmelville.github.io/mize/nesterov.html. Accessed on December 8, 2023.