**Optimization Theory and Algorithm II** **September 29, 2022**

# Lecture 7

*Lecturer:Xiangyu Chang* *Scribe: Xiangyu Chang*

*Edited by: Xiangyu Chang*

## 1 Mirror Descent

### 1.1 Projected Gradient Descent

Let us consider a general optimization problem

$$\min_{x} \; f(\mathbf{x}),$$
$$\text{s.t. } \mathbf{x} \in \Omega.$$

**Definition 1** *Suppose that $\Omega \subseteq \mathbb{R}^n$, the indicator function of $\Omega$ is*

$$\delta_\Omega(\mathbf{x}) = \begin{cases} +\infty, & \mathbf{x} \notin \Omega \\ 0, & \mathbf{x} \in \Omega. \end{cases} \tag{1}$$

**Definition 2** *The projection of a point $\mathbf{z}$ onto a set $\Omega$ is defined as*

$$\pi_\Omega(\mathbf{z}) = \arg\min_{\mathbf{x}\in\Omega} \|\mathbf{x} - \mathbf{z}\|_2. \tag{2}$$

**Example 1** *Projection examples:*

- $\Omega = \{\mathbf{x}|\mathbf{x} \succeq 0\}$, *then* $\pi_\Omega(\mathbf{z}) = \max\{\mathbf{z}, 0\}$.
- $\Omega = \{\mathbf{x}|l \preceq \mathbf{x} \preceq u\}$, *then* $\pi_\Omega(\mathbf{z}) = \max(\min\{\mathbf{z}, u\}, l)$.
- $\Omega = B_2 = \{\mathbf{x}|\ \|\mathbf{x}\|_2 \le 1\}$, *then*
$$\pi_\Omega(\mathbf{z}) = \begin{cases} \mathbf{z}, & \|\mathbf{z}\|_2 \le 1, \\ \frac{\mathbf{z}}{\|\mathbf{z}\|_2} & \|\mathbf{z}\|_2 > 1. \end{cases}$$
- $\Omega = \{\mathbf{x}|\mathbf{a}^\top\mathbf{x} = b\}$. **Q:** *What is the $\pi_\Omega(\mathbf{z})$??*

This is equivalent to

$$\min_{\mathbf{x}} \{f(\mathbf{x}) + \delta_\Omega(\mathbf{x})\}. \tag{3}$$

Obviously, $\delta_\Omega$ is convex and non-smooth. Let us compute the proximal operator of $\delta_\Omega$ as follows.

$$prox_{1/\beta\delta_\Omega}(\mathbf{z}^t) = \arg\min_{\mathbf{x}\in(\delta_\Omega)} \left\{\delta_\Omega(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{z}^t\|^2\right\} = \arg\min_{\mathbf{x}\in\Omega} \|\mathbf{x} - \mathbf{z}^t\|^2 := \pi_\Omega(\mathbf{z}^t). \tag{4}$$

Obviously, $\pi_\Omega(\mathbf{z}^t)$ is the projection of $\mathbf{z}_t$ onto $\Omega$.

- $\Omega = \{\mathbf{x}|\mathbf{x} \ge 0\}$, then $\mathbf{x}^{t+1} = prox_{1/\beta\delta_\Omega}(\mathbf{z}^t) = \pi_\Omega(\mathbf{z}^t) = \max\{\mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t), 0\}$.
- $\Omega = \{\mathbf{x}|l \le \mathbf{x} \le u\}$, then $\mathbf{x}^{t+1} = prox_{1/\beta\delta_\Omega}(\mathbf{z}^t) = \pi_\Omega(\mathbf{z}^t) = \max(\min\{\mathbf{x}^t - \frac{1}{\beta}\nabla f(\mathbf{x}^t), u\}, l)$.
- The same with $B_2$ or hyperplane.

These algorithms are called *projected gradient descent*.

### 1.1.1 Bregman Divergence

Another view point of projected gradient descent. Let us consider

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}\in\Omega} \left\{ f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle + \underbrace{\frac{1}{2s_t}\|\mathbf{x} - \mathbf{x}^t\|^2}_{\text{distance term}} \right\}.$$

If $\Omega = \mathbb{R}^n$, then $\mathbf{x}^{t+1} = \mathbf{x}^t - s_t \nabla f(\mathbf{x}^t)$.
If $\Omega \subset \mathbb{R}^n$, then $\mathbf{x}^{t+1} = \pi_\Omega(\mathbf{x}^t - s_t \nabla f(\mathbf{x}^t))$.

The basic idea of mirror descent is to choose the distance term to fit the problem geometry. So, the mirror descent is

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}\in\Omega} \left\{ f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle + \frac{1}{s_t} D_\phi(\mathbf{x}, \mathbf{x}^t) \right\},$$

where $D_\phi(\mathbf{x}, \mathbf{x}^t)$ is a generalized distance function with respect to $\phi$.

**Definition 3** *The Bregman divergence with respect to a convex function $\phi$ is denoted to be*

$$D_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla\phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{5}$$

**Example 2**

- *Let $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, then $D_\phi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.*
- *Let $\phi(\mathbf{x}) = \sum_i x_i \log x_i, \mathbf{x} \in \mathbb{R}^n_+$, then $D_\phi(\mathbf{x}, \mathbf{y}) = \sum_i (x_i \log x_i / y_i + y_i - x_i)$.*
- *If we further assume that $\mathbf{x}, \mathbf{y} \in \Delta = \{\mathbf{x} | \sum_i x_i = 1, \mathbf{x} \in \mathbb{R}^n_+\}$, that is $\Delta$ is a unit simplex. Then,*

  $$D_\phi(\mathbf{x}, \mathbf{y}) = \sum_i x_i \log x_i / y_i = KL(\mathbf{x}||\mathbf{y}), \tag{6}$$

  *where $KL$ is the KL-divergence or relative entropy.*

Properties of Bregman divergence:

- $D_\phi(\mathbf{x}, \mathbf{y}) \geq 0$. $D_\phi(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} = \mathbf{y}$.
- If $\phi$ is a $\alpha$-strongly convex function, then $D_\phi(\mathbf{x}, \mathbf{y}) \geq \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|^2$.
- $D_\phi(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$, in general not convex in $\mathbf{y}$.
- In general, $D_\phi(\mathbf{x}, \mathbf{y}) \neq D_\phi(\mathbf{y}, \mathbf{x})$.
- $$\nabla_\mathbf{x} D_\phi(\mathbf{x}, \mathbf{y}) = \nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{y}). \tag{7}$$

**Theorem 1** *(Generalized Pythagores Identity)*

$$D_\phi(\mathbf{x}, \mathbf{y}) + D_\phi(\mathbf{z}, \mathbf{x}) - D_\phi(\mathbf{z}, \mathbf{y}) = (\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{y}))^\top(\mathbf{x} - \mathbf{z}). \tag{8}$$

You can compare this with the result:

$$\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{z} - \mathbf{z}\|^2 - \|\mathbf{z} - \mathbf{y}\|^2 = 2(\mathbf{x} - \mathbf{y})^\top(\mathbf{x} - \mathbf{z}).$$

**Theorem 2** *Let $\phi$ be closed, convex and differentiable. Fix any $\mathbf{x}, \mathbf{y} \in (\phi)$, define $\hat{\mathbf{x}} = \nabla\phi(\mathbf{x})$ and $\hat{\mathbf{y}} = \nabla\phi(\mathbf{y})$, then*

$$\nabla\phi^*(\hat{\mathbf{x}}) = \nabla\phi^*(\nabla\phi(\mathbf{x})) = \mathbf{x}, \tag{9}$$
$$D_\phi(\mathbf{x}, \mathbf{y}) = D_{\phi^*}(\hat{\mathbf{y}}, \hat{\mathbf{x}}). \tag{10}$$

Before prove the theorem, let us recall the following lemma:

**Lemma 1** *Suppose that $\phi$ is closed and convex. Then the following are equivalent.*

- $\mathbf{y} \in \partial\phi(\mathbf{x})$,
- $\mathbf{x} \in \partial\phi^*(\mathbf{y})$,
- $\phi(\mathbf{x}) + \phi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$.

**Proof 1** *Proof of the above theorem. By Lemma 1, we have that*

$$\phi^*(\hat{\mathbf{x}}) = \langle \hat{\mathbf{x}}, \mathbf{x} \rangle - \phi(\mathbf{x}), \tag{11}$$
$$\phi^*(\hat{\mathbf{y}}) = \langle \hat{\mathbf{y}}, \mathbf{y} \rangle - \phi(\mathbf{y}). \tag{12}$$

*Therefore, $\nabla\phi^*(\hat{\mathbf{x}}) = \mathbf{x}$ and $\nabla\phi^*(\hat{\mathbf{y}}) = \mathbf{y}$. Compute that*

$$\begin{aligned} D_{\phi^*}(\hat{\mathbf{y}}, \hat{\mathbf{x}}) &= \phi^*(\hat{\mathbf{y}}) - \phi^*(\hat{\mathbf{x}}) - \langle \nabla\phi^*(\hat{\mathbf{x}}), \hat{\mathbf{y}} - \hat{\mathbf{x}} \rangle & (13)\\ &= \langle \hat{\mathbf{y}}, \mathbf{y} \rangle - \phi(\mathbf{y}) - \langle \hat{\mathbf{x}}, \mathbf{x} \rangle + \phi(\mathbf{x}) - \langle \mathbf{x}, \hat{\mathbf{y}} - \hat{\mathbf{x}} \rangle & (14)\\ &= D_\phi(\mathbf{x}, \mathbf{y}). & (15) \end{aligned}$$

### 1.2 Bregman Projection

**Definition 4** *The projection of $\mathbf{y}$ on to $\Omega$ under the Bregman divergence is denoted as*

$$\pi_\Omega^\phi(\mathbf{y}) = \arg\min_{\mathbf{x}\in\Omega} D_\phi(\mathbf{x}, \mathbf{y}). \tag{16}$$

*Obviously, the minimizer exists due to the convexity of $D_\phi(\mathbf{x}, \mathbf{y})$ in $\mathbf{x}$.*

**Theorem 3** *(Optimality Condition) Suppose that $\phi$ is differentiable, then for any $\mathbf{y} \in \mathbb{R}^n$, let $\pi_\Omega^\phi(\mathbf{y}) = \arg\min_{\mathbf{x}\in\Omega} D_\phi(\mathbf{x}, \mathbf{y})$, then*

$$(\nabla\phi(\pi_\Omega^\phi(\mathbf{y})) - \nabla\phi(\mathbf{y}))^\top(\pi_\Omega^\phi(\mathbf{y}) - \mathbf{z}) \leq 0, \tag{17}$$

*where for any $\mathbf{z} \in \Omega$.*

**Theorem 4**

$$D_\phi(\mathbf{z}, \mathbf{y}) \geq D_\phi(\mathbf{z}, \pi_\Omega^\phi(\mathbf{y})) + D_\phi(\pi_\Omega^\phi(\mathbf{y}), \mathbf{y}). \tag{18}$$

It can be proved by Theorem 1.

### 1.3 Bregman Projected Gradient Descent == Mirror Descent

Recall that PGD

$$\begin{aligned} \mathbf{x}^{t+1} &= \pi_\Omega\Big(\arg\min_{\mathbf{x}\in\mathbb{R}^n}\Big\{f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle + \frac{1}{2s_t}\|\mathbf{x} - \mathbf{x}^t\|^2\Big\}\Big) & (19)\\ &= \pi_\Omega(\mathbf{x}^t - s_t\nabla f(\mathbf{x}^t)). & (20) \end{aligned}$$

It comes from PGD's inspiration, the Bregman Projected Gradient Descent is

$$\mathbf{x}^{t+1} = \pi_{\Omega}^{\phi}\Big(\arg\min_{\mathbf{x}\in\mathbb{R}^n}\left\{f(\mathbf{x}^t) + \langle\nabla f(\mathbf{x}^t), \mathbf{x}-\mathbf{x}^t\rangle + \frac{1}{s_t}D_\phi(\mathbf{x},\mathbf{x}^t)\right\}\Big) \tag{21}$$

$$= \pi_{\Omega}^{\phi}((\nabla\phi)^{-1}(\nabla\phi(\mathbf{x}^t) - s_t\nabla f(\mathbf{x}^t))). \tag{22}$$

The reason is that we first to solve the unconstrained optimization

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{f(\mathbf{x}^t) + \langle\nabla f(\mathbf{x}^t), \mathbf{x}-\mathbf{x}^t\rangle + \frac{1}{s_t}D_\phi(\mathbf{x},\mathbf{x}^t)\right\}$$

to obtain the optimal value $\mathbf{y}^{t+1}$ satisfies

$$\nabla\phi(\mathbf{y}^{t+1}) = \nabla\phi(\mathbf{x}^t) - s_t\nabla f(\mathbf{x}^t).$$

Therefore,

$$\mathbf{x}^{t+1} = \pi_{\Omega}^{\phi}(\mathbf{y}^{t+1}) = \pi_{\Omega}^{\phi}((\nabla\phi)^{-1}(\nabla\phi(\mathbf{x}^t) - s_t\nabla f(\mathbf{x}^t))),$$

where $(\nabla\phi)^{-1}$ is the inverse function of $\nabla\phi$. Moreover, if we suppose that $\phi$ is strongly convex, then by Theorem 2, we have

$$\mathbf{x}^{t+1} = \pi_{\Omega}^{\phi}(\mathbf{y}^{t+1}) = \pi_{\Omega}^{\phi}(\nabla\phi^*(\nabla\phi(\mathbf{x}^t) - s_t\nabla f(\mathbf{x}^t))),$$

due to $(\nabla\phi)^{-1} = \nabla\phi^*$.

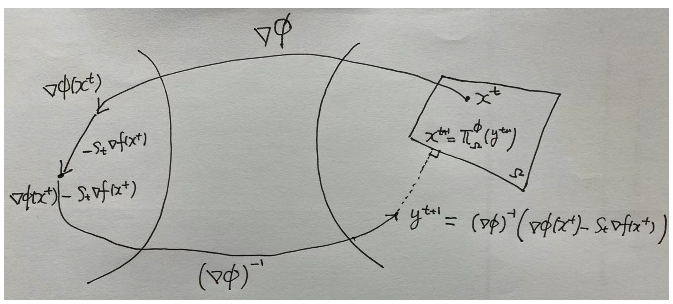


Figure 1: Primal space and Mirror space

**Example 3** 
- *Let* $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, *then* $D_\phi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$. *We have the Projected gradient descent algorithm.*
- *Let* $\phi(\mathbf{x}) = \sum_i x_i\log x_i$, *and* $\mathbf{x},\mathbf{y}\in\Omega = \{\mathbf{x}|\sum_i x_i = 1, \mathbf{x}\in\mathbb{R}_+^n\}$, *that is* $\Omega$ *is a unit simplex. Then, let us consider*

$$\pi_{\Omega}^{\phi}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\Omega} D_\phi(\mathbf{x},\mathbf{y}) \tag{23}$$

$$= \arg\min_{\mathbf{x}\in\Omega}\{\sum_i x_i\log x_i/y_i\}. \tag{24}$$

*Write down the Largrange function as* $L(\mathbf{x},\lambda) = \sum_i x_i\log x_i/y_i + \lambda(\sum_i \mathbf{x}_i - 1)$. *Take* $\frac{\partial L}{\partial x_i} = 0$, *then get* $x_i = y_i\exp(-\lambda-1)$. *According to* $\sum_i x_i = 1$, *then* $\exp(-\lambda-1) = \frac{1}{\sum_i y_i}$. *So,* $x_i = \frac{y_i}{\sum_j y_j}$, *that is*

$$\pi_{\Omega}^{\phi}(\mathbf{y}) = \mathbf{x}^* = \frac{\mathbf{y}}{\|\mathbf{y}\|_1}.$$

*Let us compute* $\mathbf{y}^{t+1}$ *according to the unconstrained optimization, then*

$$\nabla\phi(\mathbf{y}^{t+1}) = \nabla\phi(\mathbf{x}^t) - s_t\nabla f(\mathbf{x}^t),$$

*implies*

$$1 + \log y_i = 1 + \log x_i - s_t[\nabla f(\mathbf{x}^t)]_i.$$

*So,*

$$y_i^{t+1} = x_i^t \exp\{-s_t[\nabla f(\mathbf{x}^t)]_i\},$$

*then*

$$x_i^{t+1} = \frac{y_i^{t+1}}{\sum_j y_j^{t+1}} = \frac{x_i^t \exp\{-s_t[\nabla f(\mathbf{x}^t)]_i\}}{\sum_j x_j^t \exp\{-s_t[\nabla f(\mathbf{x}^t)]_j\}}.$$

### 1.3.1 Convergence Analysis of Mirror Descent

**Theorem 5** *Assume that $f$ is convex and L-Lipschz, $\phi$ is $\alpha$-strongly convex, and $\{\mathbf{x}^t\}_{t=0}^{\infty}$ is from the Mirror descent algorithm, then*

$$f^{best} - f^* \le \frac{R + \frac{L^2}{2\alpha}\sum_{t=0}^{T-1} s_t^2}{\sum_{t=0}^{T-1} s_t}, \tag{25}$$

*where $R = \sup_{\mathbf{x}\in\Omega} D_\phi(\mathbf{x}, \mathbf{x}^0)$ and $f^{best} = \min_{0\le t\le T} f(\mathbf{x}^t)$. Moreover, take $s_t = \frac{\sqrt{2\alpha R}}{L\sqrt{T}}$, then*

$$f^{best} - f^* \le L\sqrt{\frac{2R}{\alpha T}}. \tag{26}$$

**Proof 2** *By the convexity of $f$, for $t \ge 0$ and any $\mathbf{x} \in \Omega$, we have*

$$\begin{aligned}
f(\mathbf{x}^t) - f(\mathbf{x}) &\le \langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}\rangle & (27)\\
&= \frac{1}{s_t}\langle \nabla\phi(\mathbf{x}^t) - \nabla\phi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}\rangle & (28)\\
&= \frac{1}{s_t}\left[D_\phi(\mathbf{x}^t, \mathbf{y}^{t+1}) + D_\phi(\mathbf{x}, \mathbf{x}^t) - D_\phi(\mathbf{x}, \mathbf{y}^{t+1})\right] & (29)\\
&\le \frac{1}{s_t}\left[D_\phi(\mathbf{x}^t, \mathbf{y}^{t+1}) + D_\phi(\mathbf{x}, \mathbf{x}^t) - D_\phi(\mathbf{x}, \mathbf{x}^{t+1}) - D_\phi(\mathbf{x}^{t+1}, \mathbf{y}^{t+1})\right] & (30)
\end{aligned}$$

*where the first equation comes from the optimal condition, i.e., $\nabla\phi(\mathbf{y}^{t+1}) - \nabla\phi(\mathbf{x}^t) + \frac{1}{s_t}\nabla f(\mathbf{x}^t)$=0, the and the second inequality is induced by the general Pythagores identity 1, and the last inequality uses Theorem 4.*

*Applying the telescopic sum technique in the term $D_\phi(\mathbf{x}, \mathbf{x}^t) - D_\phi(\mathbf{x}, \mathbf{x}^{t+1})$ from $t = 0$ to $t = T - 1$, we can bound it with $D_\phi(\mathbf{x}, \mathbf{x}^0)$. For the remaining,*

$$\begin{aligned}
D_\phi(\mathbf{x}^t, \mathbf{y}^{t+1}) - D_\phi(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) &= \phi(\mathbf{x}^t) - \phi(\mathbf{x}^{t+1}) - \langle \nabla\phi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}^{t+1}\rangle & (31)\\
&\le \langle \nabla\phi(\mathbf{x}^t) - \nabla\phi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}^{t+1}\rangle - \frac{\alpha}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 & (32)\\
&= s_t\langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^{t+1}\rangle - \frac{\alpha}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 & (33)\\
&\le s_t L\|\mathbf{x}^t - \mathbf{x}^{t+1}\| - \frac{\alpha}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 & (34)\\
&\le \frac{(s_t L)^2}{2\alpha} & (35)
\end{aligned}$$

*where the first inequality uses the $\alpha$-strongly convex property and the last inequality uses $az - bz^2 \le \frac{a^2}{4b}$ for $\forall z \in \mathbb{R}$.*

*Hence, one has*

$$s_t\left(f(\mathbf{x}^t) - f(\mathbf{x}^*)\right) \le D_\phi(\mathbf{x}, \mathbf{x}^t) - D_\phi(\mathbf{x}, \mathbf{x}^{t+1}) + \frac{(s_t L)^2}{2\alpha} \tag{36}$$

*Summing it over from $t = 0$ to $t = T - 1$ and letting $x := x^*$, we proved,*

$$\sum_{t=0}^{T-1} s_t \left( f(\mathbf{x}^t) - f(\mathbf{x}^*) \right) \leq R + \frac{L^2}{2\alpha} \sum_{t=0}^{T-1} s_t^2. \tag{37}$$

*Plugging in $f^{best} \leq f(\mathbf{x}_t)$ for $0 \leq t \leq T$,*

$$f^{best} - f^* \leq \frac{R + \frac{L^2}{2\alpha} \sum_{t=0}^{T-1} s_t^2}{\sum_{t=0}^{T-1} s_t}, \tag{38}$$

*which complete the proof. If $s_t = \frac{\sqrt{2\alpha R}}{L\sqrt{T}}$ is a constant, it's trivial to prove that $f^{best} - f^*$ has a sub-liner convergence rate.*

# References