EE C227C Convex Optimization and Approximation Lecture 7 - 2/8/2022

Lecture 7: Mirror Descent

Lecturer: Jiantao Jiao *Scribe: Nathan Ju, Geng Zhao*

1 Recap: EGD and Mirror Descent

In this lecture, we give a more thorough treatment of the mirror descent (MD) framework. Let us first define the Bregman divergence between two points:

Definition 1 (Bregman divergence)**.** *The Bregman divergence between x and y with respect to a strictly convex function R is given by*

$$D_R(x,y) = R(x) - R(y) - \langle \nabla R(y), x - y \rangle \quad . \tag{1}$$

Recall (see lecture 5 and 6) the MD algorithm: Our $(t+1)$'th iterate is given by

$$x^{t+1} = \arg\min_{x \in K} D_R(x, x^t) + \eta \langle \nabla f(x^t), x \rangle \quad . \tag{2}$$

Note that different choices of R result in different behavior for the MD algorithm. For example, setting $R(x) = \frac{1}{2}||x||^2$ results in the proximal gradient descent algorithm. As another example, using $R(x) = \sum_{i=1}^n x_i \ln x_i - x_i$ (called the unnormalized negative entropy function) results in the exponential gradient descent (EGD) algorithm. The $(t+1)$'th iterate can be written in a more convenient form; let us define $g^t = \nabla f(x^t)$ and $w^{t+1} = \nabla R^{-1}(\nabla R(x^t) - \eta g^t)$ and rewrite the iterate:

$$\begin{aligned}
x^{t+1} &= \arg\min_{x \in K} D_R(x, x^t) + \eta \langle g^t, x \rangle \\
&= \arg\min_{x \in K} \eta \langle g^t, x \rangle + R(x) - R(x^t) - \langle \nabla R(x^t), x - x^t \rangle \\
&= \arg\min_{x \in K} R(x) - \langle \nabla R(w^{t+1}), x \rangle \\
&= \arg\min_{x \in K} R(x) - R(w^{t+1}) - \langle \nabla R(w^{t+1}), x - w^{t+1} \rangle \\
&= \arg\min_{x \in K} D_R(x, w^{t+1}) \quad .
\end{aligned} \tag{3}$$

In pseudocode,

Algorithm 1 Mirror descent

for $t = 0, ..., T-1$ **do**
  Obtain $g^t \leftarrow \nabla f(x^t)$
  Assign w^{t+1} to be s.t. $\nabla R(w^{t+1}) = \nabla R(x^t) - \eta g^t$
  Assign $x^{t+1} \leftarrow \arg\min_{x \in K} D_R(x, w^{t+1})$
end for
Return $\bar{x} = \frac{1}{T} \sum_{t=0}^{T-1} x^t$

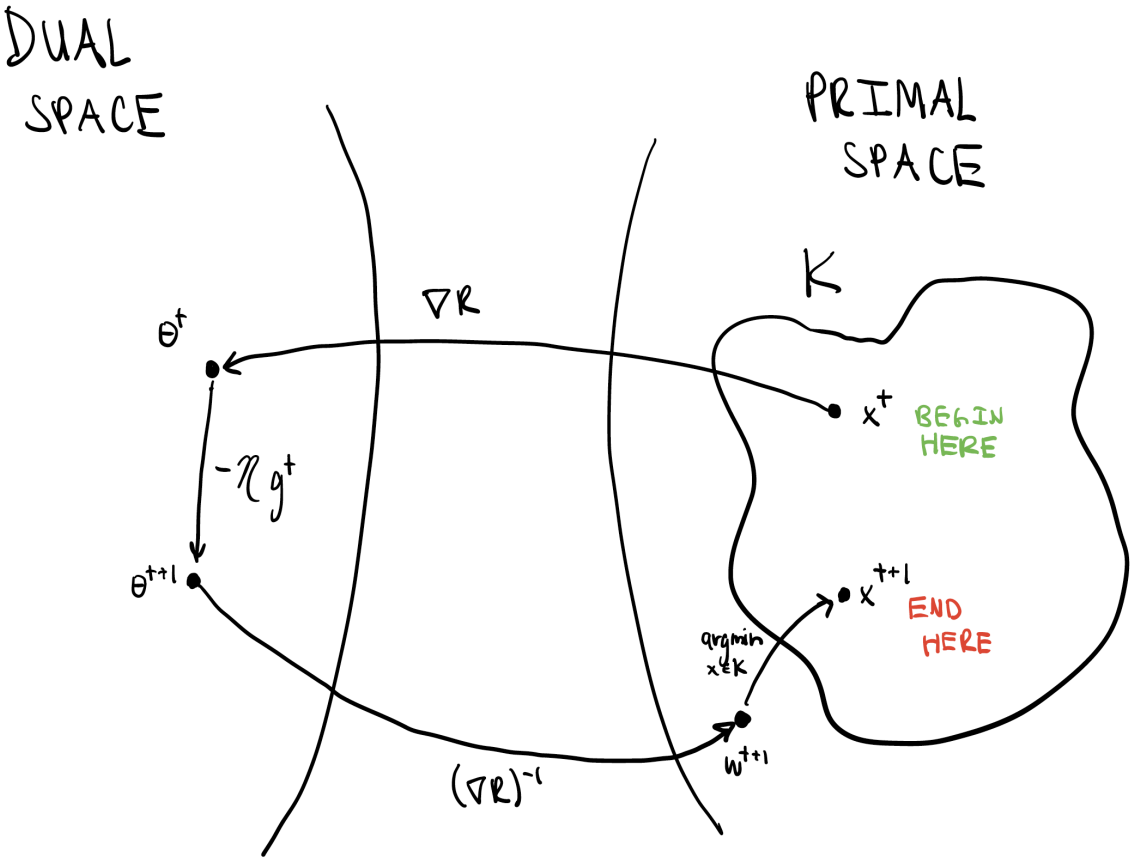
2 Interpretation of "Mirror"

An observation by Nemirovsky and Yudin in 1978 gives a good justification for the "mirror" aspect of mirror descent. As an example, recall that vanilla gradient descent's update rule is given by $x^{t+1} = x^t - \eta \nabla f(x^t)$. The gradient $\nabla f(x^t)$ is actually a linear functional, which lives in the *dual* space of x^t, so adding the two elements is, in a sense, ill-defined. This motivates the mirror viewpoint and the mirror descent framework of the previous section.

In more detail, let X be our primal space (think of this as the space in which x^t lives) and Y the dual space. Using the same notation as before, define

$$\nabla R : X \mapsto Y \tag{4}$$

as the gradient of the Bregman divergence function R, which we assume to be bijective. In the mirror descent framework, ∇R and ∇R^{-1} are the maps between primal and dual space; once a point is mapped to the dual space, the algorithm performs "gradient descent" in this space, then maps the resulting dual point back to the corresponding closest primal point. In this way, the dual space is the "mirror" to the primal space. See the following figure for an illustration.



3 Performance guarantee

Theorem 2. *Consider convex functions $f : K \to \mathbb{R}$ and $R : X \to \mathbb{R}$ with $K \subseteq X \subseteq \mathbb{R}^n$ and assume the followings:*

1. *$\nabla R : X \to \mathbb{R}^n$ is a bijection.*

2. *f has bounded gradient with respect to $\|\cdot\|_*$, i.e., $\|\nabla f(x)\|_* \le G$ for all $x \in K$.*

3. *R is σ-strongly convex with respect to the norm $\|\cdot\|$, i.e.,*

$$D_R(x,y) \ge \frac{\sigma}{2}\|x-y\|^2 \quad \forall\, x,y \in K. \tag{5}$$

Then the MD algorithm guarantees that

$$f(\bar{x}) - f(x^*) \le \frac{1}{\eta T}\left(D + \frac{T\eta^2 G^2}{2\sigma}\right), \tag{6}$$

where $D := D_R(x^*, x^0)$. Further, by the AM-GM inequality, we may choose $\eta = \sqrt{\frac{2\sigma}{TG^2D}}$ to minimize the right hand side, which yields

$$f(\bar{x}) - f(x^*) \le \sqrt{\frac{2DG^2}{\sigma T}}. \tag{7}$$

Remark Here are a few observations and interpretations regarding this result.

(a) The strong convexity of R can be formulated alternatively (by writing out D_R explicitly) as $R(x) - R(y) - \langle \nabla R(x), x-y\rangle \ge \frac{\sigma}{2}\|x-y\|^2$ for all $x, y \in K$, which implies (by interchanging x and y) that

$$\langle \nabla R(x) - \nabla R(y), x-y\rangle \ge \sigma\|x-y\|^2 \quad \forall\, x,y \in K, \tag{8}$$

an important property for strongly convex functions in optimization theory.

(b) The norm $\|\cdot\|$ is arbitrary and is an algorithmic choice. E.g., in EGD, $\|\cdot\|$ is the ℓ^1 norm and $\|\cdot\|_*$ is the ℓ^∞ norm.

(c) The bound (7) depends primarily on the constants D and G; we can always scale R accordingly to make $\sigma = 1$ (which will affect D). The constant G can be interpreted as the Lipschitz constant of function f, i.e., $G = \sup_{x,y\in K, x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|}$. This allows us to incorporate the subgradient method into this framework (so that g^t can be a subgradient of f at x^t).

Sketch of Proof The proof is nearly identical to the analysis of EGD in the previous lecture. It follows the same initial steps using convexity of f, the law of cosines (which is valid for any Bregman divergence), and the generalized Pythagorean theorem (again valid for any Bregman divergence). The only difference is in the next step where we use the strong convexity assumption, in place of Pinsker's inequality, along with Hőlder's inequality to derive

$$D_R(x^t, w^{t+1}) - D_R(x^{t+1}, w^{t+1}) \le \|g^t\|_*\|x^t - x^{t+1}\| - \frac{\sigma}{2}\|x^t - x^{t+1}\|^2.$$

The rest of the proof is entirely the same and requires no further properties. □

4 Comparing EGD with subgradient method

At this point, it is important to examine the motivation for the EGD algorithm. Now that we are able to cast both EGD and the subgradient method into the MD framework, we will provide a comparison between the two.

We follow the discussion in the online lecture notes [1, Pg 619-624].

The subgradient method can be treated as mirror descent in domain $X \subseteq B_2(r) = \{x \in \mathbb{R}^n : \|x\|_2 \le r\}$ with $R_{\text{subgrad}}(x) = \frac{1}{2}\|x\|_2^2$, so

$$D_{R_{\text{subgrad}}}(x,y) = \frac{1}{2}\|x-y\|_2^2 \le O\big(\max_{x,y\in X}\|x-y\|_2^2\big) \tag{9}$$

for any $x, y \in X$.

Similarly, the EGD method operates in domain $X \subseteq \Delta_n = \{x \in \mathbb{R}^n : x \ge 0, \|x\|_1 = 1\}$, i.e., the standard simplex[1]. In Lecture 5 and 6, we used the generalized negative entropy function $H(x)$ as the distance generating function and claimed that $D_H(x, \frac{1}{n}\mathbf{1}) \le O(\ln n)$ for any $x \in \Delta_n$. This time, however, we use a slightly different distance generating function

$$R_{\text{EGD}}(x) = (1+\delta)\sum_{i=1}^n \left(x_i + \frac{\delta}{n}\right)\log\left(x_i + \frac{\delta}{n}\right), \tag{10}$$

with a small $\delta = 10^{-16}$. Heuristically, this is nearly identical to the negative entropy function, and we justify this choice as follows: First, writing $\bar{x}_i = x_i + \delta/n$ and $\bar{R}_{\text{EGD}}(x) = \sum_{i=1}^n \bar{x}_i \log \bar{x}_i$, we use Cauchy-Schwarz inequality to derive

$$\begin{aligned}\langle h, \nabla^2 R_{\text{EGD}}(x) h\rangle = (1+\delta)\langle h, \nabla^2 \bar{R}_{\text{EGD}}(x) h\rangle &= (1+\delta)\sum_{i=1}^n \frac{h_i^2}{\bar{x}_i} \\ &= \left(\sum_{i=1}^n \bar{x}_i\right)\left(\sum_{i=1}^n \frac{h_i^2}{\bar{x}_i}\right) \ge \left(\sum_{i=1}^n |h_i|\right)^2 = \|h\|_1^2 \end{aligned} \tag{11}$$

for all $x \in \Delta_n$ and all $h \in \mathbb{R}^n$, implying strong convexity (with respect to ℓ^1 norm). Next, for any $x, y \in \Delta_n$, again writing $\bar{x}_i = x_i + \delta/n$ and $\bar{y}_i = y_i + \delta/n$, we have

$$\begin{aligned} D_{R_{\text{EGD}}}(x,y) &= R_{\text{EGD}}(x) - R_{\text{EGD}}(y) - \langle \nabla R_{\text{EGD}}(y), x - y\rangle \\ &= (1+\delta)\left(\sum_{i=1}^n \bar{x}_i \log \bar{x}_i - \sum_{i=1}^n \bar{y}_i \log \bar{y}_i - \sum_{i=1}^n (1+\log \bar{y}_i)(\bar{x}_i - \bar{y}_i)\right) \\ &= (1+\delta)\left(\sum_{i=1}^n \bar{x}_i \log \frac{\bar{x}_i}{\bar{y}_i} + \sum_{i=1}^n y_i - \sum_{i=1}^n x_i\right) \\ &\le (1+\delta)\left(\sum_{i=1}^n \bar{x}_i \log \frac{n+\delta}{\delta} + 1\right) = O(\ln n), \end{aligned} \tag{12}$$

where we simply use the trivial bound $\delta/n \le \bar{x}_i, \bar{y}_i \le 1 + \delta/n$. Notice that this guarantee holds for all pairs $x, y \in \Delta_n$, an improvement over the guarantee we had on $D_H(x,y)$ (which only holds when $y = \frac{1}{n}\mathbf{1}$). Further, a more careful treatment of the last step in (12) reveals that, whenever $\|x-y\|_1 = \sum_{i=1}^n |\bar{x}_i - \bar{y}_i| \le a$ for some $a \in (0,1)$, we have

$$D_{R_{\text{EGD}}}(x,y) = (1+\delta)\left(\sum_{i=1}^n \bar{x}_i \log \frac{\bar{x}_i}{\bar{y}_i} + \bar{y}_i - \bar{x}_i\right) \le \left(a + \frac{\delta}{n}\right)\log\frac{a+\delta/n}{\delta/n} - a, \tag{13}$$

achieved when $x_1 = a$, $y_1 = 0$, and $x_i = y_i$ for $i > 1$. This can be seen by first fixing $|x_i - y_i|$ for all $i \in [n]$ and maximizing over y, and then maximizing over x fixing y and $\|x-y\|_1$ using convexity of the negative entropy function. By Taylor expanding (13) to second order, we conclude that

$$D_{R_{\text{EGD}}}(x,y) \le O(\log n)\max_{x,y\in X}\|x-y\|_1^2. \tag{14}$$

[1]In [1], the simplex setup has domain $X = \{x \in \mathbb{R}^n : x \ge 0, \|x\|_1 \le 1\}$. It is easy to see that this is equivalent to the EGD setup for Δ_{n+1} in our class.

Now we compare the convergence of EGD and subgradient method for general convex optimization problems. Note that in order to apply these methods, we first need to transform the domain X of the problem into the standard simplex or the unit ball by scaling and translating. This can be done for any compact feasible domain X, and the scaling only affects the Lipschitz constant of f by a factor of the diameter of X (in ℓ^1 for EGD and in ℓ^2 for subgradient method).

Applying the efficiency estimate (7) in Theorem 2 and the bound (14), we see that the guaranteed upper bound for EGD is

$$E_{\text{EGD}} := \frac{\sqrt{O(\ln n)} \cdot \max_{x,y \in X} \|x - y\|_1 \max_{x \in X} \|f'(x)\|_\infty}{\sqrt{T}}, \tag{15}$$

and by (9) the corresponding bound for subgradient method is

$$E_{\text{subgrad}} := \frac{\sqrt{O(1)} \cdot \max_{x,y \in X} \|x - y\|_2 \max_{x \in X} \|f'(x)\|_2}{\sqrt{T}}. \tag{16}$$

The ratio of these two quantities is

$$\frac{E_{\text{EGD}}}{E_{\text{subgrad}}} = O(\sqrt{\ln n}) \cdot \underbrace{\frac{\max_{x,y \in X} \|x - y\|_1}{\max_{x,y \in X} \|x - y\|_2}}_{A} \cdot \underbrace{\frac{\max_{x \in X} \|f'(x)\|_\infty}{\max_{x \in X} \|f'(x)\|_2}}_{B}. \tag{17}$$

A smaller ratio $\frac{E_{\text{EGD}}}{E_{\text{subgrad}}}$ indicates that, as far as theoretical guarantees are concerned, EGD outperforms subgradient method, and vice versa.

Notice that $\|u\|_p \geq \|u\|_q$ whenever $1 \leq p \leq q \leq \infty$. Hence, the term $A \geq 1$ always (i.e., against EGD) and can range between 1 and $\sqrt{n}$ depending on the geometry of the domain X; similarly, $B \leq 1$ always (i.e., in favor of EGD) and can range from 1 to as small as $\frac{1}{\sqrt{n}}$ depending on the geometry of f. The factor of $O(\sqrt{\ln n})$ is against EGD but in practice just a moderate absolute constant. Overall, the relative performance of EGD and subgradient method really depends on the geometry of X and f.

To make the comparison concrete, let us examine a few extreme examples. We first compare the cases when X is an (ℓ^2) ball versus when X is the standard simplex.

- When X is a ball $B_2(r) \subseteq \mathbb{R}^n$, the diameter in ℓ^1 is $\sqrt{n}$ larger than in ℓ^2, i.e., $A = \sqrt{n}$. Since $B \geq \frac{1}{\sqrt{n}}$, the ratio $\frac{E_{\text{EGD}}}{E_{\text{subgrad}}} \geq 1$, meaning that the classical subgradient method outperforms EGD.

- When X is the standard simplex $\Delta_n \subseteq \mathbb{R}^n$, its diameters in ℓ^1 and ℓ^2 are both of constant order, i.e., $A = O(1)$. Since $B \leq 1$ and $O(\sqrt{\ln n})$ is in practice a moderate absolute constant, the ratio $\frac{E_{\text{EGD}}}{E_{\text{subgrad}}} \leq O(1)$, meaning that EGD outperforms the classical subgradient method.

Next, we examine the dependency on the geometry of f:

- When all first order partial derivatives of f (in X) are of the same order, i.e., f is nearly equally sensitive to each variable, f' has n roughly equal coordinates and hence

$$B = O\left(\frac{\|\mathbf{1}\|_\infty}{\|\mathbf{1}\|_2}\right) = O\left(\frac{1}{\sqrt{n}}\right). \tag{18}$$

- When just $O(1)$ first order partial derivatives of f (in X) are of the same order while the remaining are negligible, i.e., f is only sensitive to a constant number of variables, we have

$$B = O\left(\frac{\|\mathbf{e}_1\|_\infty}{\|\mathbf{e}_1\|_2}\right) = O(1). \tag{19}$$

5 Comparison with Newton's method

We can now compare Newton's method with mirror descent. In Newton's method, the update step is given by

$$x^{t+1} = \arg\min_{x \in K} \frac{1}{2}(x - x^t)^\top H(x^t)(x - x^t) + \eta \langle g^t, x \rangle \tag{20}$$

where $H(x^t)$ is an approximation to the Hessian of the function f around x^t. We will see in the future that Newton's method performs well near the point of optimality, but in general is not guaranteed to converge.

Now let us analyze the update step in mirror descent. Recall that it is given by

$$x^{t+1} = \arg\min_{x \in K} D_R(x, x^t) + \eta \langle \nabla f(x^t), x \rangle \quad . \tag{21}$$

If we expand $D_R(\cdot, x^t)$ locally around x^t, then it is approximately equivalent to $\frac{1}{2}(x - x^t)^\top \nabla^2 R(x^t)(x - x^t)$. So the difference between mirror descent and Newton's method is that in the former, we are taking the Hessian of the strongly convex function R instead of the function f. As opposed to Newton's method, mirror descent does enjoy global convergence guarantee.

We will offer more discussion in the future when we study second order methods.

6 Discussion: Why entropy function?

Recall from Lecture 5 that for exponential gradient descent on the standard simplex we used distance generating function $R(x) = H(x) = \sum_{i=1}^n (x_i \log x_i - x_i)$, i.e., the generalized negative entropy function. In general, we want the distance generating function $R(x)$ to minimize the parameter $D := D_R(x^*, x^0)$ in Theorem 2 to achieve better performance guarantee for mirror descent, subject to the strong convexity constraint under the given norm $\|\cdot\|$. In the simplex case, i.e., when $X = \Delta_n$ and $\|\cdot\| = \|\cdot\|_1$, it can be shown that the entropy function achieves a value of $D \le \max_{x \in X} D_H(x, x^0) = O(\log n)$ when we choose $x^0 = \frac{1}{n}\mathbf{1} = \min_{x' \in X} H(x')$, the H-center of X. In this setup, this upper bound on D cannot be reduced by more than an absolute constant factor with other choices of $R(x)$. Similarly, when $X = B_2(r)$ is the Euclidean ball and $\|\cdot\| = \|\cdot\|_2$, the choice of $R(x) = \frac{1}{2}\|x\|^2$ leads to $D \le O(1)r^2$, again optimal up to an absolute constant.

However, it is important to note that the entropy function is not the only reasonable choice: similar convergence guarantees can be obtained for other distance generating functions, e.g., $\omega(x) = \sum_{i=1}^n x_i^{p(n)}$ or $\omega(x) = \|x\|_{p(n)}^2$ with $p(n) = 1 + O(1/\ln n)$. These alternatives can be even better than the standard setup in some sense. Namely, $\omega(x)$ is continuously differentiable on the entire $X = \Delta_n$, and moreover $D_\omega(x, y) \le O(\log n)$ for all $x, y \in X$ *not just when* $y = \frac{1}{n}\mathbf{1}$ (cf. $D_R(x, y) \le O(r^2)$ for all $x, y \in B_2(r)$ with $R(x) = \frac{1}{2}\|x\|^2$, and the distance generating function R_{EGD} we used above also has a similar guarantee in the simplex setup). These properties are valuable in certain situations, e.g., when updating prox-centers in the bundle methods, which we will cover in the future. The only drawback is that their proximal maps are harder to compute (recall the closed form solution for the standard EGD from Lecture 5), often solved using the bisection method on simple univariate equations. In practice, however, this computational overhead is usually insignificant. See Section 5.2 of [2] for more discussions on different proximal setups.

References

[1] A. Nemirovski, "Lectures on modern convex optimization," https://www2.isye.gatech.edu/~nemirovs/LMCOTR2022Spring.pdf, 2022, accessed: 2022–02-14.

[2] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.