EE C227C Convex Optimization and Approximation Lecture 9 - 02/15/2022

Lecture 9: Lazy Mirror Descent

Lecturer: Jiantao Jiao *Scribe: Anna Deza, Kunhe Yang*

In this lecture, we continue our discussion on variants of mirror descent. We will first introduce the notion of *Lazy Mirror Descent*, then move on to a new result in the exponential gradient descent literature that replaces the update rule $w_i^{t+1} = w_i^t e^{-\eta g_i^t}$ with its first-order Taylor series expansion approximation $w_i^{t+1} = w_i^t(1-\eta g_i^t)$. Surprisingly, this alternative update rule has a better performance guarantee, and this is not because of analytical artifacts, but because the algorithm itself is intrinsically stronger. We will explain the reason behind this phenomenon in this lecture.

1 Lazy Mirror Descent

Motivation: Exponential Gradient Descent To motivate the idea of Lazy Mirror Descent algorithm, we consider a special example of Mirror Descent: Exponential Gradient Descent (EGD). In this example, we choose $R(x) = \sum_{i=1}^n x_i \log x_i$ to be the negative entropy function. Following the MD update rule, at each time, we first do an exponential decrease $w_i^{t+1} = w_i^t e^{-\eta g_i^t}$, then normalize it by $x_i^{t+1} = w_i^{t+1}/\|w^{t+1}\|_1$. However, a closer look at the trajectory suggests that the decrease and normalization actually commutes. This implies that, conceptually, we can keep decreasing w without any normalization, then only when outputting the final probability distribution, we normalize it once.

$$w_i^{t+1} = w_i^0 e^{-\eta \sum_{i=1}^t g_i^t} \qquad \Rightarrow \qquad p_i^{t+1} = w_i^{t+1}/\|w^{t+1}\|_1 \tag{1}$$

This observation suggests that, instead of going back and forth between the primal space and dual space, we can accumulate all the descent vectors in the dual space, and only when being asked to output x^t in the primal space, we project the current dual point back to the primal using the same methodology as mirror descent. This motivates the idea of lazy mirror descent.

1.1 Lazy Mirror Descent Algorithm

In the Lazy Mirror Descent Algorithm, we start from a local minimum x^0 of $R(\cdot)$ in the primal space. For instance, in the EGD example we have discussed above, x^0 corresponds to the uniform distribution, which minimizes the negative entropy function. Applying the mirror map ∇R, we can see that the initial point in the dual space is exactly $\theta^0 = \nabla R(x^0) = 0$. Accumulating all the descent vectors, the algorithm computes

$$\theta^t = \sum_{s=0}^{t-1}(-\eta g^s). \tag{2}$$

If asked for a point in the primal space at time t, the algorithm will perform an inverse mirror map and then project it back to the convex domain:

$$w^t = (\nabla R)^{-1}(\theta^t), \qquad x^t = \arg\min_{x\in\mathcal{K}} D_R(x, w^t). \tag{3}$$

We can immediately see that the x^t given by (2) and (3) can be equivalently computed by

$$\begin{aligned} x^t &= \arg\min_{x\in\mathcal{K}} D_R(x, w^t) = \arg\min_{x\in\mathcal{K}} \big(R(x) - \langle \nabla R(w^t), x\rangle\big) \\ &= \arg\min_{x\in\mathcal{K}} \left(\left\langle \eta \sum_{s=0}^{t-1} g^s, x \right\rangle + R(x) \right). \end{aligned}$$

Above is the lazy mirror descent algorithm. See Figure 1 for an illustration.

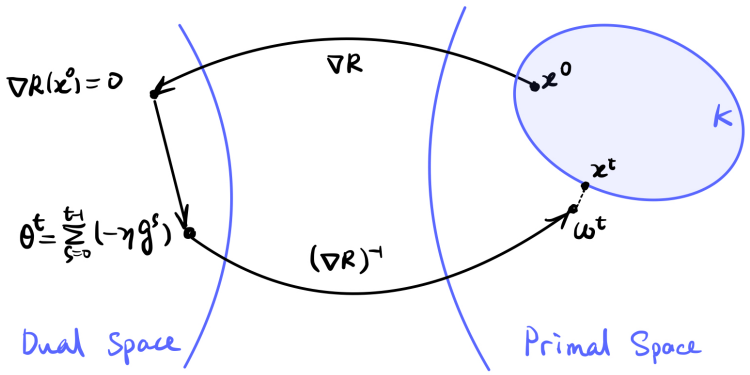

Figure 1: Illustration of lazy mirror descent

MD vs Lazy MD Although Lazy MD follows exactly the same trajectory as MD in the EGD example, in general the behavior of MD and Lazy MD can be different. This is because lazy MD reduces computation by dropping the step of going back and forth between the primal and the dual, and simply averaging the gradients in the dual. To be more specific, instead of having

$$\nabla R(w^{t+1}) \leftarrow \nabla R(x^t) - \eta g^t$$

in lazy MD, we can simply compute

$$\nabla R(w^{t+1}) \leftarrow \nabla R(w^t) - \eta g^t. \tag{4}$$

How to obtain $\mathbf{g^t}$ On the one hand, as in the EGD example, $\{g^t\}_{t=0}^{T-1}$ are not necessarily interpreted as the gradients of a function f. Instead, they could be any sequence of bounded vectors given by an adversary, where our goal is minimizing the notion of regret, which is defined as $\sum_{t=0}^{T-1}\langle g^t, x^t - x\rangle$ for any given $x \in \mathcal{K}$.

On the other hand, if we do want to minimize a convex function f, then g^t needs to be a sub-gradient of f at x^t, which requires us to map back to the primal space at every time (see Algorithm 1 for details). Even in this scenario, we no longer need to repeatedly compute the map from the primal to the dual. This can be advantageous in some situations, such as distributed settings, or the settings where inverse mirror map $(\nabla R)^{-1}$ is much easier to compute than the mirror map ∇R itself. In the next subsection, we will see the performance guarantee of using lazy MD to optimize a strongly convex function f.

Algorithm 1 Lazy Mirror Descent

Initialize $\nabla R(w^0) = \nabla R(x^0) = 0$.

for $t = 0, 1, \cdots, T-1$ **do**

Obtain $g^t \leftarrow \nabla f(x^t)$

Let w^{t+1} be such that $\nabla R(w^{t+1}) \leftarrow \nabla R(w^t) - \eta g^t$

$x^{t+1} \leftarrow \arg\min_{x\in\mathcal{K}} D_R(x, w^{t+1})$

end for

1.2 Performance Guarantee

Theorem 1. *Let $R : \mathcal{X} \to \mathbb{R}$ be a σ-strongly convex function defined on a convex and closed set $\mathcal{X} \subseteq \mathcal{R}^n$ w.r.t $\|\cdot\|$. Let $D^2 = \sup_{x\in\mathcal{K}} R(x) - R(x^0)$, where $\mathcal{K} \subseteq \mathcal{X}$ is a convex set. Also, let f be convex and L-Lipschitz*

w.r.t. $\|\cdot\|$. *Then the lazy mirror descent algorithm in Algorithm 1 with* $\eta = \frac{D}{L}\sqrt{\frac{\sigma}{2T}}$ *satisfies*

$$f\left(\frac{1}{T}\sum_{t=0}^{T-1} x^t\right) - f(x^*) \leq 2DL\sqrt{\frac{2}{\sigma T}}.$$

Proof Define $\psi^t(x) = \eta\sum_{s=0}^{t-1}\langle g^s, x\rangle + R(x)$. Then $x^t \in \arg\min_{x\in\mathcal{K}} \psi^t(x)$. Since R is σ-strongly convex, ψ^t is also σ-strongly convex for all t. Then

$$\psi^{t+1}(x^{t+1}) - \psi^{t+1}(x^t) \leq \langle \nabla\psi^{t+1}(x^{t+1}), x^{t+1} - x^t\rangle - \frac{\sigma}{2}\|x^{t+1} - x^t\|^2 \leq -\frac{\sigma}{2}\|x^{t+1} - x^t\|^2,$$

where the last inequality holds by the first order optimality condition. Also, we can observe that

$$\psi^{t+1}(x^{t+1}) - \psi^{t+1}(x^t) = \psi^t(x^{t+1}) - \psi^t(x^t) + \eta\langle g^t, x^{t+1} - x^t\rangle \geq \eta\langle g^t, x^{t+1} - x^t\rangle.$$

Combining the above two inequalities together with that $\|g^t\|_* \leq L$, we can obtain that

$$\frac{\sigma}{2}\|x^{t+1} - x^t\|^2 \leq \eta\langle g^t, x^t - x^{t+1}\rangle \leq \eta L\|x^t - x^{t+1}\|.$$

where the last inequality holds by generalized Cauchy-Schwartz inequality. Therefore,

$$\|x^t - x^{t+1}\| \leq \frac{2\eta L}{\sigma}$$

and thus we can obtain that

$$\langle g^t, x^t - x^{t+1}\rangle \leq \frac{2\eta L^2}{\sigma} \tag{5}$$

again by generalized Cauchy-Schwartz inequality. Now we claim that for any $x \in \mathcal{K}$,

$$\sum_{t=0}^{T-1}\langle g^t, x^{t+1}\rangle + \frac{R(x^0)}{\eta} \leq \sum_{t=0}^{T-1}\langle g^t, x\rangle + \frac{R(x)}{\eta}. \tag{6}$$

We prove by induction on T. For $T = 0$, it directly holds since $x^0 \in \arg\min_{x\in\mathcal{K}} R(x)$. Now assume for $T-1$, we have

$$\sum_{t=0}^{T-2}\langle g^t, x^{t+1}\rangle + \frac{R(x^0)}{\eta} \leq \sum_{t=0}^{T-2}\langle g^t, x\rangle + \frac{R(x)}{\eta}.$$

In particular, setting $x = x^T$, we can obtain that

$$\sum_{t=0}^{T-2}\langle g^t, x^{t+1}\rangle + \frac{R(x^0)}{\eta} \leq \sum_{t=0}^{T-2}\langle g^t, x^T\rangle + \frac{R(x^T)}{\eta}.$$

Therefore, we can conclude for T that

$$\sum_{t=0}^{T-1}\langle g^t, x^{t+1}\rangle + \frac{R(x^0)}{\eta} \leq \langle g^{T-1}, x^T\rangle + \sum_{t=0}^{T-2}\langle g^t, x^T\rangle + \frac{R(x^T)}{\eta} \leq \sum_{t=0}^{T-1}\langle g^t, x\rangle + \frac{R(x)}{\eta},$$

where the last inequality holds by the definition of x^T. Finally, note that (6) is equivalent to

$$\sum_{t=0}^{T-1}\langle g^t, x^t - x\rangle \leq \sum_{t=0}^{T-1}\langle g^t, x^t - x^{t+1}\rangle + \frac{R(x) - R(x^0)}{\eta}$$

and apply (5), we can complete the proof. □

2 Improved Exponential Gradient Descent

2.1 Alternative Algorithm

In this section, we focus on analyzing a variant of EGD, which has the following update rule

$$w_i^{t+1} \propto w_i(1-\eta g_i^t)$$

Theorem 2 is a comparison between the regret bound of the original EGD and the above variant, which we denote by MW2. We will formally prove this theorem in the next lecture.

Theorem 2. *Suppose $\|g^t\|_\infty \le 1, 0 < \eta \le \frac{1}{2}$, then*

1. *(MW1) If the update rule is $w^{t+1} \propto w^t e^{-\eta g^t}$, then*

$$\sum_{t=0}^{T-1} \langle g^t, w^t - p\rangle \le \frac{\log n}{\eta} + \eta \sum_{t=0}^{T-1} \|g^t\|_\infty^2.$$

2. *(MW2) If the update rule is $w^{t+1} \propto w^t(1-\eta g^t)$, then*

$$\sum_{t=0}^{T-1} \langle g^t, w^t - p\rangle \le \frac{\log n}{\eta} + \eta \sum_{t=0}^{T-1} \mathbb{E}_{i\sim p}[(g_i^t)^2] = \frac{\log n}{\eta} + \eta \sum_{t=0}^{T-1} \langle p, (g^t)^2\rangle,$$

where $(g^t)^2$ denote the element-wise square operation.

Note that for the alternative update rule, the regret bound depends on $\mathbb{E}_{i\sim p}[(g_i^t)^2]$, where the distribution p is chosen by us and does not rely on any intermediate terms such as w^t. Therefore, if we know beforehand that the magnitude of certain coordinates of g^t are going to be small, then we can put more mass on these coordinates in p to recover a better bound. Specifically, in the case where the different coordinates in g^t are highly heterogeneous, the second guarantee can be much more stronger than the first one.

Re-imagining MW2 in the mirror descent framework In fact, the alternative algorithm (MW2) can still be cast in the mirror descent framework, because it can be understood as a form of adaptive mirror descent (see Algorithm 2), where the regularizer changes in each round t in response to previously observed vectors $g^0, \cdots, g^{t-1}$.

Recall that in MD, the update rule is derived by

$$x^{t+1} = \arg\min_{x\in\mathcal{K}} \left(D_R(x, x^t) + \eta\langle g^t, x\rangle\right) = \arg\min_{x\in\mathcal{K}} \left(R(x) - \left\langle \underbrace{\nabla R(x^t) - \eta g^t}_{(a)}, x \right\rangle\right).$$

In the last section, we have shown that term (a) in the above equation can be replaced by $\theta^{t+1} = -\eta\sum_{s=0}^t g^s$, which gives us the lazy MD algorithm. We will show in sequel that if we replace the fixed regularizer R by an adaptive regularizer ψ^t, then the MW2 update rule can be understood as an instance of lazy MD, namely

$$w^t = \arg\min_{w\in\Delta_n} \left(\psi^t(w) - \langle\theta^t, w\rangle\right) \tag{7}$$

Theorem 3. *Define $\beta_i^t = \sum_{s=0}^{t-1} \log(1-\eta g_i^s)$, $\phi(u) = \sum_{i=0}^n u_i \log u_i$, and the adaptive regularizer*

$$\psi^t(u) = \phi(u) + \langle u, \theta^t - \beta^t\rangle, \tag{8}$$

then Equation (7) holds, i.e., MW2 corresponds exactly to the adaptive mirror descent with regularizer ψ^t.

We remark that in Equation (8), the second term can be seen as performing a second-order correction to the gradient. If we do not have this correction term, i.e., $\beta_i^t = -\eta \sum_{s=0}^{t-1} g^s$, then the update rule (7) reduces to the non-adaptive EGD algorithm.

Proof [of Theorem 3]

$$\begin{aligned}\arg\min_{w\in\Delta_n}\left(\psi^t(w) - \langle\theta^t, w\rangle\right) =& \arg\min_{w\in\Delta_n}\left(\phi(w) + \langle w, \theta^t - \beta^t\rangle - \langle\theta^t, w\rangle\right)\\ =& \arg\min_{w\in\Delta_n}\left(\phi(w) - \langle w, \beta^t\rangle\right).\end{aligned}$$

Using Lagrange multipliers to incorporate the constraint $w \in \Delta_n$, we can see immediately that the above constrained optimization problem has solution

$$w_i^t \propto_i e^{\beta_i^t} = e^{\sum_{s=0}^{t-1}\log(1-\eta g_i^s)} = \prod_{s=0}^{t-1}(1-\eta g_i^s), \tag{9}$$

which is exactly the update rule of MW2. □

We will now show why MW2 can provide a much better bound than MW1 by considering a special case.

Proposition 4. *For any η and T, there exists a sequence of $(g^t)_{t=0}^{T-1}$ and $p \in \Delta_n$ such that the updates in (MW1) results in $\sum_{t=0}^{T-1}\langle g^t, w^t - p\rangle = \Omega(\sqrt{T})$.*

Proof To show the above proposition, we will construct two sequences of $(g^t)_{t=0}^{T-1}$ and show that for every η, MW1 will perform $\Omega(\sqrt{T})$ on at least one of the sequences. Let $n = 2$ and $p = \left[1,\ 0\right]^\top$.

Sequence 1 Let $g^t = \left[0,\ (-1)^t\right]^\top$. Notice that $\sum_{t=0}^{T-1} g_2^t$ will be either 0 or 1, depending on T being odd or even. When t is even, $w_t = \left[\frac{1}{2},\ \frac{1}{2}\right]^\top$. Otherwise, when t is odd, $w_t = \left[\frac{1}{1+\exp(-\eta)},\ \frac{1}{1+\exp(\eta)}\right]^\top$. If $\eta \le 1$ then:

$$\begin{aligned}\sum_{t=0}^{T-1}\langle g^t, w^t - p\rangle &= \sum_{t=0}^{T-1} g_2^t w_2^t\\ &= \sum_{t=0}^{\lfloor\frac{T-1}{2}\rfloor}\left(\frac{1}{2} - \frac{1}{1+\exp(\eta)}\right)\\ &= \lfloor\frac{T-1}{2}\rfloor\left(\frac{1}{2} - \frac{1}{1+\exp(\eta)}\right)\\ &\ge \lfloor\frac{T-1}{2}\rfloor\left(\frac{1}{2} - \frac{1}{2+2\eta}\right)\\ &\ge \frac{1}{4}\lfloor\frac{T-1}{2}\rfloor\eta.\end{aligned}$$

The first inequality follows from the fact that $\exp(x) \ge 1 + x$ when $|x| \le 1$ and the second is due to the bound being minimized when $\eta = 1$. This term is increasing in η, so we have a lower bound of $\frac{1}{4}\lfloor\frac{T-1}{2}\rfloor\min(1, \eta)$.

Sequence 2 Let $g^t = \begin{bmatrix} 0, & 1 \end{bmatrix}^\top$. Then $w_2^t = \frac{1}{1+\exp((t-1)\eta)}$ and thus when $t \leq \lceil \frac{1}{\eta} \rceil$, $w_2^t \geq \frac{1}{1+e}$. In this case,

$$\begin{aligned}
\sum_{t=0}^{T-1} \langle g^t, w^t - p \rangle &= \sum_{t=0}^{T-1} w_2^t \\
&\geq \sum_{t=0}^{\min(T-1, \frac{1}{\eta})} w_2^t \\
&\geq \sum_{t=0}^{\min(T-1, \frac{1}{\eta})} \frac{1}{1+e} \\
&= \frac{1}{1+e} \min(T-1, \frac{1}{\eta})
\end{aligned}$$

Now consider two possible cases. One is when $\eta \geq 1/\sqrt{T-1}$, then the first sequence will give a regret of $\Omega(\sqrt{T})$. The second is when $\eta \leq 1/\sqrt{T-1}$, then we get a regret of $\Omega(1/\sqrt{T})$ on the second sequence. Therefore, there will always be a regret of $\Omega(\sqrt{T})$ in any case. □

Notice that MW2 will not suffer in this special case, achieving an asymptotically constant regret. To see this, consider the bound on MW2 shown in Theorem 2. Since we could chose p, we have put all the weight on g_1^t of the sequences we introduced, which is always equal to 0. This gives a constant bound of $\frac{\log n}{\eta}$.

In the remainder of this lecture, we formally introduce the adaptive mirror descent algorithm and its performance guarantee. Before we begin this topic, we must introduce the concept of the convex conjugate.

2.2 Convex Conjugate

Definition 5 (Convex conjugate)**.** *Given a function $\psi : \mathcal{K} \mapsto \mathbb{R}$, we define the convex conjugate (also called the Fenchel-Legendre dual) ψ^* as:*

$$\psi^*(y) \triangleq \sup_{x \in \mathcal{K}} (\langle x, y \rangle - \psi(x)).$$

Properties of the convex conjugate We now discuss some important properties of ψ^*:

1. ψ^* is always convex, regardless if ψ is convex or not.
2. If ψ is a closed proper convex function, then the biconjugate $\psi^{**} = \psi$.
3. If ψ is convex and differentiable, then $\nabla \psi^* = (\nabla \psi)^{-1}$.

The convex conjugate is useful since via the first property, we know it is always convex even when ψ is not, making it potentially easier to manipulate.

The third property reveals the argument maximizer relationship. To see this relationship, consider $\psi^*(y)$ for a fixed y. To get $\psi^*(y)$, we take the supremum over a set of functions of the form $\langle x, y \rangle - \psi(x)$. Since $\psi(x)$ is convex, $-\psi(x)$ is concave. Therefore, obtaining $\psi^*(y)$ is a convex problem, as we are taking a supremum over a set of concave functions. Let x^* be the maximizer for a fixed y. Then taking the gradient of $\psi^*(y)$ with respect to x we obtain:

$$y = \nabla \psi(x^*).$$

This implies that the maximizer of the expression $\langle x, y \rangle - \psi(x)$ is $x^* = (\nabla \psi)^{-1}(y)$. By the second property, we see that $\psi(x) = \sup_{y \in \mathcal{Y}} (\langle y, x \rangle - \psi^*(y))$. By using the same logic as above, we get that for fixed x, the maximizing argument is $y^* = (\nabla \psi^*)^{-1}(x)$. Now recall that to take the gradient of a supremum, we must first

find the maximizing argument for the expression, and then take the gradient of the maximizing expression. Taking the gradient of $\psi^*(y)$ we get:

$$\nabla\psi^*(y) = x^* = (\nabla\psi)^{-1}(y),$$

which is exactly the third property.

2.3 Adaptive Mirror Descent

Consider the following algorithm:

Algorithm 2 Adaptive Mirror Descent

Initialize $\theta^0 = 0$.
  for $t = 0, 1, \cdots, T-1$ **do**
    Choose $w^t \leftarrow \nabla\psi^{t^*}(\theta^t)$
    Update $\theta^{t+1} \leftarrow \theta^t - \eta g^t$
  end for

To analyze this, we must first introduce the following lemma:

Lemma 6. *Suppose ψ^t is convex and satisfies the following property: $\psi^{t+1}(\theta^t - \eta g^t) \leq \psi^{t^*}(\theta^t) - \eta\langle w^t, g^t\rangle$. Then Algorithm 2 satisfies regret:*

$$\sum_{t=0}^{T-1}\langle g^t, w^t - p\rangle \leq \frac{\psi^{0^*}(\theta^0) + \psi^T(p)}{\eta}.$$

This lemma states that if our convex regularizer has some nice properties, then we can use it to derive a bound. If we are able to compute the convex conjugate in closed form, then we can easily analyze the right hand side. The regret bound is controlled by $\psi^T(p)$, with the p chosen by us.

Via property 3, we can get an intuition on why Algorithm 2 is right. We use $\nabla\psi^*$, which is actually $(\nabla\psi)^{-1}$, which we use to solve Equation (7).

Next lecture we will prove the lemma, and show that we can get a closed form for ψ^*, allowing us to practically use the lemma. For further details on adaptive mirror descent, we refer the reader to [1].

References

[1] J. Steinhardt and P. Liang, "Adaptivity and optimism: An improved exponentiated gradient algorithm," in *International Conference on Machine Learning*. PMLR, 2014, pp. 1593–1601.