EE C227C Convex Optimization and Approximation Lecture 6 - 02/03/2022

Lecture 6: Exponential Gradient Descent

Lecturer: Jiantao Jiao *Scribe: Syomantak Chaudhuri, Ezinne Nwankwo, Sheng-Jung Yu*

In this lecture, we explain the algorithm for exponential gradient descent (EGD) and prove an upper bound of it's convergence rate. Previously, we discussed the general paradigms of descent algorithms, such as when the distance function was quadratic ($D(x, x^t) = \frac{1}{2}\|x - x^t\|_2^2$). Now lets choose a particular distance function for a convex set known as Kullback-Leiber divergence over a probability simplex.

Definition 1 (Kullback-Leiber Divergence over Δ_n)**.** *For two probability distributions $p, q \in \Delta_n$, their Kullback-Leiber divergence is defined as*

$$D_{KL} := \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

Despite not being symmetric, KL divergence follows natural distance-like properties and for this definition to make sense it follows that if $q_i = 0$ then $p_i = 0$ and $D_{KL}(p, q) = 0 \Rightarrow p = q$. Additionally, from convexity, it follows that $D_{KL} \geq 0$.

Such a choice leads to the EGD algorithm. In this lecture, we will analyze EGD and provide an upper bound [1].

1 The Algorithm

Let's recap the EGD algorithm. We are trying to solve the following convex optimization problem

$$\min_{p \in \Delta_n} f(p),$$

where $f : \Delta_n \mapsto \mathbb{R}$ is a convex function over the closed and compact n-dimensional probability simplex

$$\Delta_n := \left\{ p \in [0,1]^n : \sum_{i=1}^{n} p_i = 1 \right\}.$$

For this algorithm, the update rule that now incorporates the distance function D_{KL}. It takes the form:

$$p^{t+1} := \arg\min_{p \in \Delta_n} \left\{ D_{KL}(p, p^t) + \eta \left\langle \nabla f(p^t), p \right\rangle \right\}$$

Algorithm 1 Exponential Gradient Descent (EGD)

```
Input: convex $f : \Delta_n \mapsto \mathbb{R}$
Parameters: $\eta \geq 0, T > 0$
Output: A point $\bar{p} \in \Delta_n$
Algorithm:

Set $p^0 = \frac{1}{n}\mathbb{1}$ (the uniform distribution) $\in \Delta_n$
for  $t = 0 : T - 1$ do
    $g^t := \nabla f(p^t)$
    $w^{t+1} := p_i^{t+1} e^{-\eta g_i^t}$
    $p_i^{t+1} := \frac{w_i^{t+1}}{\|w^{t+1}\|}$
return $\bar{p} = \frac{1}{T} \sum_{t=0}^{T-1} p^t$
```

It's important to note that with EGD the output is actually an average of point $\bar{p}$ as opposed to the last iterate p^{T-1}, which was the case in other gradient descent algorithms discussed in class. To illustrate why this difference matters, let's consider an example where we wish to minimize $f(x) = |x|$. The gradient of this function will either be 1 or -1 at every point, but knowing this does not give us any information about whether we are close to or far from the minimizer (0). As opposed to the case where we have a Lipschitz gradient, with EGD the gradient does not guarantee that we are close to the optimal. So instead we gather more information by visiting more points on the function and averaging.

2 Proof of Convergence

We want to show that for any $p \in \Delta_n$, it holds that

$$f(\bar{p}) - f(p) \leq \epsilon,$$

where $\bar{p} = \frac{1}{T}\sum_{t=0}^{T-1} p^t$. If we can prove the above bound, then it is also true that this results holds for the minimizer p^* of f over Δ_n.
Step 1: Bound using gradients

We first start with bounding $f(\bar{p}) - f(p)$ by it's gradient.

$$\begin{aligned} f(\bar{p}) - f(p) &\leq \left(\frac{1}{T}\sum_{t=0}^{T-1} f(p^t)\right) - f(p) \text{ (Def. of convexity)} \\ &= \frac{1}{T}\sum_{t=0}^{T-1}(f(p^t) - f(p)) \\ &\leq \frac{1}{T}\sum_{t=0}^{T-1}\langle \nabla f(p^t), p^t - p\rangle \text{ (First-order notion of convexity)} \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\langle g^t, p^t - p\rangle \end{aligned}$$

Now all we need to do is focus on providing an upper bound for $\frac{1}{T}\sum_{t=0}^{T-1}\langle g^t, p^t - p\rangle$.
Step 2: Write in terms of KL-divergence

Fix $t \in \{0, \dots, T-1\}$. We know from algorithm 1, that

$$w_i^{t+1} = p_i^t e^{-\eta g_i^t}$$

and by solving for g_i^t, we get that for all $i \in \{1, \dots, n\}$

$$g_i^t = \frac{1}{\eta}(\log\, p_i^t - \log\, w_i^{t+1}).$$

Next, we will write this in terms of gradient of the generalized negative entropy function $H(x) = \sum_{i=1}^n (x_i \log\, x_i - x_i)$ and $\nabla H(x) = [\ln\, x_1, \ln\, x_2, \dots, \ln\, x_n]^T$

$$g^t = \frac{1}{\eta}(\log\, p_i^t - \log\, w_i^{t+1}) = \frac{1}{\eta}(\nabla H(p^t) - \nabla H(w^{t+1}))$$

We can continue simplifying g^t, but for the next step to make sense, we must first define the Law of cosines for Bregman divergence.

Lemma 2 (Law of Cosines for Bregman Divergence)**.** *Let $R : K \mapsto \mathbb{R}$ be a convex, differentiable function and let $x, y, z \in K$. Then*

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

where $D_R(x, y) = R(x) - R(y) - \langle \nabla R(y), x - y \rangle$ is the Bregman divergence induced by R and it is $D_R \geq 0$ since R is convex.

In the special case of $R = \|x\|_2^2$, by the above formula we can write

$$\begin{aligned} 2\langle y - z, y - x \rangle &= \|x - y\|^2 + \|y - z\|^2 - \|x - z\|^2 \\ \Rightarrow \|x - z\|^2 &= \|x - y\|^2 + \|y - z\|^2 - 2\langle y - z, y - x \rangle \end{aligned}$$

This is very similar to the law of cosines in Euclidean space.

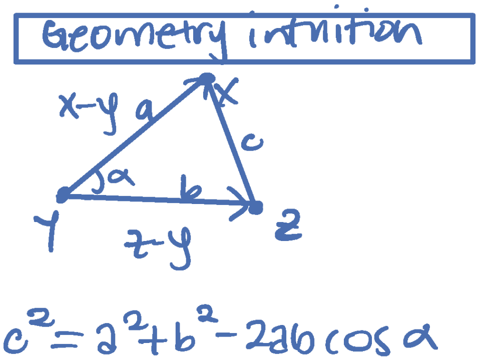

Figure 1: The geometric interpretation of the law of cosines equation in Eucliden space.

Going back to bounding our gradient, with this we can write

$$\begin{aligned} \langle g^t, p^t - p \rangle &= \frac{1}{\eta} \langle \nabla H(p^t) - \nabla H(w^{t+1}), p^t - p \rangle \\ &= \frac{1}{\eta} (D_H(p, p^t) + D_H(p^t, w^{t+1}) - D_H(p, w^{t+1})), \end{aligned}$$

where D_H is some distance function induced by the entropy function H. We set $D_H = D_{KL}$ (See theorem 1).
Step 3: Using the Pythagorean theorem to get a telescoping sum

To further simplifies the gradient, we use the following lemma:

Lemma 3. Generalized Pythagoream Theorem
Let $R : K \mapsto \mathbb{R}$ be a convex, differentiable function and let $S \subseteq K$ be a closed convex subset of K. Let $x, y \in S$ and $z \in K$ such that

$$y = \arg\min_{u \in S} D_R(u, z).$$

Then

$$D_R(x, y) + D_R(y, z) \leq D_R(x, z).$$

The geometric interpretation is that z is a point outside the convex set S, and we define a projection of z to the set S to be $y = \arg\min_{u \in S} D_R(u, z)$, the nearest point in S with respect to the Bregman Distance of defined by R.

Let's say $R = \|x\|_2^2$, the theorem becomes

$$\|x - y\|_2^2 + \|y - z\|_2^2 \leq \|x - z\|_2^2,$$

which shows that the angle between the vectors $x - y$ and $z - y$ is an obtuse angle, as shown in Figure 2

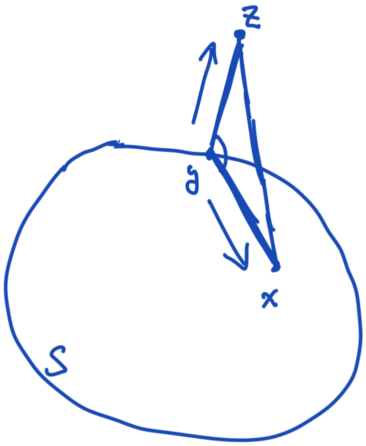

Figure 2: The geometric interpretation of the General Pathegorean Theorem in Eucliden space.

We can see the Generalized Pathagorean Theorem can deal with points outside the convex set. In our previous derivation, p and p^t are both in Δ_n, while w^{t+1} may be outside Δ_n because we didn't normalize it. Thus, we can apply Generalized Pathagorean Theorem to get a relation for w^{t+1}:

$$\begin{aligned}
& \eta \sum_{t=0}^{T-1} \langle g^t, p^t - p \rangle \\
= \quad & \eta * \frac{1}{\eta} \sum_{t=0}^{T-1} \langle \nabla H(p^t) - \nabla H(w^{t+1}), p^t - p \rangle \\
\leq \quad & \sum_{t=0}^{T-1} (D_H(p, p^t) + D_H(p^t, w^{t+1}) - (D_H(p, p^{t+1}) + D_H(p^{t+1}, w^{t+1}))) \\
= \quad & \sum_{t=0}^{T-1} (D_H(p, p^t) - (D_H(p, p^{t+1})) + \sum_{t=0}^{T-1} (D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}))) \\
= \quad & D_H(p, p^0) - D_H(p, p^T) + \sum_{t=0}^{T-1} (D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}))) \\
\leq \quad & D_H(p, p^0) + \sum_{t=0}^{T-1} (D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}))), \qquad (1)
\end{aligned}$$

where for the last inequality, we use the fact that $D_H(p, p^T) \geq 0$.

Step 4: Using Pinsker's inequality and bounded gradient to bound the remaining terms

To proceed, we apply Law of Cosines on $D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}))$:

$$\begin{aligned} D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}) &= D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}) + D_H(p^{t+1}, p^t) - D_H(p^{t+1}, p^t) \\ &= \langle \nabla H(p^t) - \nabla H(w^{t+1}), p^t - p^{t+1} \rangle - D_H(p^{t+1}, p^t) \\ &= \eta \langle g^t, p^t - p^{t+1} \rangle - D_H(p^{t+1}, p^t) \end{aligned}$$

Here, we cannot do too much on it unless additional assumptions are introduced. Thus, we introduce the following two lemmas: The Pinsker's Inequality and The Holder's Inequality.

Lemma 4. Pinsker's Inequality *For every $x, y \in \Delta_n$ we have*

$$D_{KL}(x, y) \geq \frac{1}{2} \|x - y\|_1^2.$$

Pinsker's Inequality show that the negative entropy function is 1-strongly convex with respect to the l_1 norm. This can be easily observed by the definitions of negative entropy function and the strongly convexity.

$$D_{KL}(x, y) = H(x) - H(y) - \langle \nabla H(y), x - y \rangle \geq \frac{1}{2} \|x - y\|_1^2,$$

where the definition for l_1 norm is $\|x\|_1 = \sum_1^n |x_i|$.

Lemma 5. Holder's Inequality

$$\langle x, y \rangle \leq \|x\| \|y\|_*$$

The definition of the dual norm is $\|y\|_* = \sup\{\langle x, y \rangle \mid \|x\| \leq 1\}$

$$D_H(p^t, w^{t+1}) - D_H(p^{t+1}) \leq \eta \langle g^t, p^t - p^{t+1} \rangle - \frac{1}{2} \|p^{t+1} - p^t\|_1^2$$

Using Holder's Inequality and the assumption that the gradient is bounded by G, we can get

$$\begin{aligned} \langle g^t, p^t - p^{t+1} \rangle &\leq \|g^t\|_\infty \|p^t - p^{t+1}\|_1 \\ &\leq G \|p^t - p^{t+1}\|_1 \\ &= G \|p^{t+1} - p^t\|_1 \end{aligned}$$

Combining the above two lemmas, we can bound the expressions by the l_1 norm of $p^t - p^{t+1}$

$$\begin{aligned} D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}) &\leq \eta \|g^t\|_\infty \|p^t - p^{t+1}\|_1 - \frac{1}{2} \|p^{t+1} - p^t\|_1^2 \\ &\leq \eta G \|p^{t+1} - p^t\|_1 - \frac{1}{2} \|p^{t+1} - p^t\|_1^2 \end{aligned}$$

where $G = \max_t \|g^t\|_\infty$.

Now, denoting $\|p^{t+1} - p^t\|_1$ as z, the expression above is $\eta G z - \frac{1}{2} z^2$; this is a quadratic which has a maximum value of $\frac{(\eta G)^2}{2}$ achieved at $z = \eta G$. Therefore, we can upper bound the expression as

$$D_H(p^t, w^{t+1}) - D_H(p^{t+1}, w^{t+1}) \leq \frac{(\eta G)^2}{2}$$

Hence, Eq (1) gives

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

The expression on right is minimized by setting $\eta = \sqrt{\frac{2 D_H(p, p^0)}{G^2 T}}$. In order to get a faster convergence, initial distribution p^0 is usually taken 'centrally‘ so $D_H(p, p^0)$ is not too big for any distribution p. Also note that we used a value of η which involves the quantity G, which we need to know beforehand!

3 Mirror Descent

The general update rule is of form

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

Let w^{t+1} be a point such that $\nabla R(w^{t+1}) = \nabla R(x^t) - \eta g^t$. Note that defining w^{t+1} has the underlying assumption that $\nabla R(\cdot)$ is a bijective map and its range is $\mathbb{R}^n$.

$$\begin{aligned} x^{t+1} &= \arg\min_{x \in K} \left\{ R(x) - \langle \nabla R(w^{t+1}), x \rangle \right\} \\ &= \arg\min_{x \in K} \left\{ D_R(x, w^{t+1}) \right\} \end{aligned}$$

So intuitively, we are projecting w^{t+1} back to domain K with respect to the the distance $D_R(\cdot, \cdot)$.

References

[1] N. K. Vishnoi, “Algorithms for convex optimization,” *Cambridge University Press*, vol. 1, no. 28, Aug. 2020. [Online]. Available: https://convex-optimization.github.io/ACO-v1.pdf