EE C227C Convex Optimization and Approximation Lecture 10 - 2/17/2022

Lecture 10: Adaptive Mirror Descent and Bundle Method

Lecturer: Jiantao Jiao *Scribe: Jiaqi Yang, Nikhil Deveshwar*

1 Adaptive Mirror Descent

Algorithm 1 Adaptive Mirror Descent

```
Require: convex regularizar: $\psi_t$
  $\theta^0 = 0$
  for t = 0, 1, ..., T-1 do
      $w^t \leftarrow \nabla \psi_t^*(\theta^t)$
      observe $g^t$
      $\theta^{t+1} \leftarrow \theta^t - \eta g^t$
```

Lemma 1. *Assume $\psi_{t+1}^*(\theta^t - \eta g^t) \leq \psi_t(\theta^t) - \eta\langle w^t, g^t\rangle$ then,*

$$\sum_{t=0}^{T-1} \langle g^t, w^t - u\rangle \leq \frac{\psi_0^*(\theta^0) + \psi_T(u)}{\eta} \tag{1}$$

Proof For any u:

$$\begin{aligned}
\langle u, \theta^T\rangle - \psi_T(u) &\leq \psi_T^*(\theta^T) & (2)\\
&\leq \psi_0^*(\theta^0) + \sum_{t=0}^{T-1}(\psi_{t+1}^*(\theta^{t+1}) - \psi_t^*(\theta^t)) & (3)\\
&\leq \psi_0^*(\theta^0) + \sum_{t=0}^{T-1}(-\eta\langle w^t, g^t\rangle) & (4)\\
\langle u, \theta^T\rangle + \eta\sum_{t=0}^{T-1}\langle w^t, g^t\rangle &\leq \psi_0^*(\theta^0) + \psi_T(u) & (5)
\end{aligned}$$

where

$$\theta^T = \sum_{t=0}^{T-1} -\eta g^t \tag{6}$$

Expression (4) matches our original expression from earlier: $\sum_{t=0}^{T-1}(-\eta\langle w^t, g^t\rangle)$

□

Now for the case of a particular convex regularizer: ψ_t we can define the following terms:

$$\begin{aligned}
\beta_i^t &= \sum_{s=0}^{t-1} \log(1-\eta g_i^s) && (7)\\
\psi_t(u) &\triangleq \psi_(u) + \langle u, \theta^t - \beta^t \rangle && (8)\\
\psi(u) &= \sum_{s=0}^{t-1} u_i \log u_i && (9)
\end{aligned}$$

where $\langle u, \theta^t - \beta^t \rangle$ is a correction term.
Aside: Given $\psi(u) = \sum_i u_i \log u_i$ with $u_i \geq 0$ and $\sum u_i = 1$, its convex conjugate is $\psi^*(x) = \log(\sum_i \exp(x_i))$. The convexity of this expression can be established by "Convexity calculus" with no computations:

$$\begin{aligned}
s > 0 \Rightarrow ln(s) &= \min_z [s * \exp\{z\} - z - 1] && (10)\\
\Rightarrow \ln\left(\sum_i \exp\{x_i\}\right) &= \min_z \Big[\sum_i \exp\{z\}\exp\{x_i\} - z - 1\Big] && (11)
\end{aligned}$$

The top line is the straight forward computation, while the bottom term in the minimization objective function is a convex function of $[x; z]$.
Using the definition of convex conjugate:

$$\begin{aligned}
\psi_t^*(x) &= \psi^*(x + (\beta^t - \theta^t)) && (12)\\
\psi_{t+1}^*(\theta^{t+1}) &= \psi^*(\theta^{t+1} + (\beta^{t+1} - \theta^{t+1})) && (13)\\
&= \psi^*(\beta^{t+1}) && (14)
\end{aligned}$$

This also implies $\psi_t^*(\theta^t) = \psi^*(\beta^t)$
Using convex conjugate of entropy function,

$$\begin{aligned}
\psi^*(\beta^{t+1}) &= \log\left(\sum_{i=1}^n \exp\big(\beta_i^{t+1}\big)\right) && (15)\\
&= \log\left(\sum_{i=1}^n \exp\big(\beta_i^t + \log\big(1-\eta g_i^t\big)\big)\right) && (16)\\
&= \log\left(\sum_{i=1}^n \exp\big(\beta_i^t\big)(1-\eta g_i^t)\right) && (17)\\
&= \log\left(\sum_{i=1}^n \exp\big(\beta_i^t\big) - \eta \sum_{i=1}^n g_i^t \exp(\beta_i)\right) && (18)\\
&\leq \psi^*(\beta^t) - \eta \frac{\sum_{i=1}^n g_i^t \exp(\beta_i^t)}{\sum_{i=1}^n \exp(\beta_i^t)} && (19)\\
&= \psi^*(\beta^t) - \eta \langle w^t, g^t \rangle && (20)
\end{aligned}$$

Here, $\frac{\exp(\beta_i^t)}{\sum_{i=1}^n \exp(\beta_i^t)} \to \nabla\psi^*(\beta^t)$ (gradient term w^t). This gives us the left side expression from Lemma 1. Note that we have used inequality $\log(x-y) \leq \log(x) - \frac{y}{x}$.

Now, we evaluate right side original expression from Lemma 1: $\frac{\psi_0^*(\theta^0) + \psi_T(u)}{\eta}$

$$\begin{aligned}
\psi_0^*(\theta^0) &= \log\left(\sum_{i=1}^n \exp(0)\right) = \log(n) && (21)\\
\psi_T(u) &= \psi(u) + \langle u, \theta^T - \beta^T \rangle && (22)
\end{aligned}$$

Where

$$\theta^T = -\eta \sum_{s=0}^{T-1} g^s \tag{23}$$

$$\theta^T - \beta^T = \sum_{s=0}^{T-1} (-\eta g^s - \log(1 - \eta g^s)) \tag{24}$$

and where the i-th entry $\le \eta^2 \left|g_i^s\right|^2$ by applying the inequality $-x - x^2 \le \ln(1-x)$ for $|x| \le 1/2$.

2 Bundle Method

2.1 Kelley's Method

Let $f(x)$ be a convex (non-smooth) function, $\{x_i : i = 1, 2, \cdots\}$ be a sequence of points. We consider the task of finding the minimum $f^\star = \min_{x \in G} f(x)$, where G is a convex compact set of interest. The bundle is defined to be a sequence of affine forms, each of which is a global lower bound of the function $f(\cdot)$. Formally, it is the right-hand side of the following equation:

$$f(x) \ge f(x_j) + \langle \boldsymbol{\nabla} f(x_j), x - x_j \rangle. \tag{25}$$

Because the affine forms serve as global lower bounds of the function, so does their maximum. Therefore, Kelley's method is proposed to iteratively compute the minimum $f^\star$, as detailed in Algorithm 2.

Algorithm 2 Kelley's Method

```
Require: initial point $x_1$
  for $i \leftarrow 1, 2, \cdots$ do
      Let $x_{i+1} \in \arg\min_{x \in G} f_i(x)$, where
          $f_i(x) = \max_{1 \le j \le i} \{f(x_j) + \langle \nabla f(x_j), x - x_j \rangle\}$.   (26)
```

The intuition behind Kelley's method is that, by choosing the maximum over the lower bounds of the function, we could discover a new point that hopefully has the "highest uncertainty", and thus we query at that point in the next iteration, with the hope that the new lower bound at that point could give us more information about the landscape of the function, which helps us to find the minimum.

Unfortunately, this idea does not work even for mild function classes such as the convex Lipschitz functions. Specifically, to find x such that $f(x) \le f^\star + \epsilon$, one may need $\Omega(1/\epsilon^{(n+1)/2})$ iterations, where n is the dimension. Here, we intuitively explain why the Kelley's method could be slow. This is because the $\arg\min$ over $f_i(\cdot)$ operation in the Kelley's method are too brittle, because $f_i(\cdot)$ is a maximum of affine forms, which is piece-wise linear, and as a result, there could be big jump in the $\arg\min$ operation with a slightest perturbation.

We conclude this section with the following example illustrating why Kelley's method would be slow.

Example 2 ([1], Example 8.1)**.** Consider optimizing the convex function

$$f(x) = \max\{0, -1 + 2\varepsilon + \|x\|\}, \tag{27}$$

$$\partial f(x) = \begin{cases} \{0\}, & \|x\| < 1 - 2\varepsilon, \\ \{\frac{x}{\|x\|}\}, & \|x\| > 1 - 2\varepsilon, \\ \mathrm{conv}\{0, \frac{x}{\|x\|}\}, & \|x\| = 1 - 2\varepsilon, \end{cases} \tag{28}$$

over the unit ball $G = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ using Kelley's method with initial point $x_1 = 0$. The goal is to find an ε-optimal solution, so our stopping criterion is $f(x_{i+1}) - f_i(x_{i+1}) \leq \varepsilon$.

After the first iteration, we have $f_1 = 0$, so $x_2 \in \arg\min_{x\in G} 0$ could be arbitrary. Suppose it returns a solution with $\|x_2\| = 1$. Then $f(x_2) = 2\varepsilon$ and $f_2(x) = \max\{f_1(x), 2\varepsilon + \langle x_2, x - x_2\rangle\} = \max\{0, -1 + 2\varepsilon + \langle x_2, x\rangle\}$. Note that $f_2(x)$ is non-zero only on a sphere cap $S_2 = \{x \in G : \langle x, x_2\rangle > 1 - 2\varepsilon\}$, so $x_3 \in \arg\min_{x\in G} f_2(x) = G \setminus S_2$. Subsequently, we have

$$x_{i+1} \in \arg\min_{x\in G} f_i(x) = G \setminus (\bigcup_{j=2}^{i} S_j). \tag{29}$$

As a result, the algorithm would not stop until $G \setminus (\bigcup_{j=2}^{i} S_j)$ is an empty set. Let $\nu_n = \mathsf{Vol}_{n-1}(\partial G)$, where $\mathsf{Vol}_n(\cdot)$ denotes the volume in $\mathbb{R}^n$. We note that

$$\mathsf{Vol}_{n-1}(\partial G) = \nu_n = 2\nu_{n-1}\int_0^1 (1-t^2)^{\frac{n-1}{2}}\,\mathrm{d}t \geq \nu_{n-1}\int_0^1 2t(1-t^2)^{\frac{n-1}{2}}\,\mathrm{d}t \geq \frac{2}{n+1}\nu_{n-1}, \tag{30}$$

$$\begin{aligned}\mathsf{Vol}_{n-1}(\partial S_i) &= \nu_{n-1}\int_{1-2\varepsilon}^1 (1-t^2)^{\frac{n-1}{2}}\,\mathrm{d}t \\ &\leq \nu_{n-1}(1-(1-2\varepsilon)^2)^{\frac{n-1}{2}}(1-(1-2\varepsilon)) \leq \nu_{n-1}(4\varepsilon - 4\varepsilon^2)^{\frac{n+1}{2}}, \end{aligned}\tag{31}$$

so there would be at least

$$\frac{\mathsf{Vol}_{n-1}(\partial G)}{\mathsf{Vol}_{n-1}(\partial S_i)} \geq \frac{2}{n+1}(\frac{1}{4\varepsilon})^{\frac{n+1}{2}}. \tag{32}$$

2.2 Level Method

We fix the instability issue in Kelley's method by adding a quadratic regularizer, which borrows insight from the mirror descent algorithm. Specifically, we replace the rule of choosing x_{i+1} in Algorithm 2 with the following equation:

$$x_{i+1} = \arg\min_G\{f_i(x) + \frac{d_i}{2}\left\|x - x_i^+\right\|^2\}, \tag{33}$$

where x_i^+ may not be x_i. We call d_i the prox weight and x_i^+ the prox center. The objective in (33) is similar to the mirror descent. Indeed, in mirror descent, d_i is like a constant. Furthermore, we observe that when $d_i = 0$ it reduces to the Kelley's method. When $d_i \to +\infty$ the point x_i does not move. Therefore, we need to cleverly choose d_i to seek a balance. Unfortunately, we do not really know how to choose d_i.

Level method is a smart way of choosing d_i. In fact, it somehow eliminates the need of explicitly choosing d_i. This is achieved by implicit mapping the parameter space of d_i to another parameter space. Note that this is in analogous to how the Lagrange's method of multipliers solves the constrained optimization problem. Indeed, the optimal Lagrangian multiplier can be seen as an implicit mapping from the constraints to the multipliers, which circumvents the need of explicitly choosing the multiplier.

Same story here. Level method implicitly sets the parameters d_i. Here, we formally describe the level method in Algorithm 3.

We make the following observations to the level method.

Lemma 3. *1. $f_1^- \leq f_2^- \leq \cdots \leq f^\star$.*

2. $f_1^+ \geq f_2^+ \geq \cdots \geq f^\star$.

3. Define $\Delta_i = f_i^+ - f_i^-$. Then $\Delta_1 \geq \Delta_2 \geq \cdots \geq 0$.

Proof

Algorithm 3 Level Method

```
Require: initial point $x_1$
  for $i \leftarrow 1, 2, \cdots$ do
    Solve $f_i^- = \min_{x \in G} f_i(x)$, where $f_i(x)$ is defined in (26). Let $f_i^+ = \min_{1 \leq j \leq i} f(x)$
    Form the level $\ell_i = (1 - \lambda) f_i^- + \lambda f_i^+$
    Set $x_{i+1} = \Pi_{Q_i}(x_i)$, where $Q_i = \{x \in G \mid f_i(x) \leq \ell_i\}$.
```

1. It is clear that $f_i^- \leq f^\star$ because $f_i(x) \leq f(x)$ for every $x \in G$. We have $f_i^- \leq f_{i+1}^-$ because $f_{i+1}(\cdot)$ has more terms to maximize over than $f_i(\cdot)$ by (26).
2. $f_i^+ \geq f^\star$ because $f^\star$ is global minimum. $f_i^+ \geq f_{i+1}^+$ because f_{i+1}^+ has more terms to minimize over, compared to f_i^+.
3. It follows immediately from parts 1 and 2.

□

Finally, we emphasize that the projection operator Π_{Q_i} in Algorithm 3 can be solved efficiently when G is well-shaped. For example, when G is a polytope, the projection essentially finds a point in Q_i that is closest in Euclidean distance to x_i, which reduces to applying convex and linear constraints to a polytope and solving a convex optimization problem.

References

[1] A. Belloni, "Lecture notes for iap 2005 course introduction to bundle methods," *Operation Research Center, MIT, Version of February*, vol. 11, 2005.