EE C227C Convex Optimization and Approximation Lecture 27 - 4/28/2022

Lecture 27: Smoothing Techniques

Lecturer: Nived Rajaraman *Scribe: Xin Zhou, Mark Oussoren, Philip Canoza*

1 Smooth Approximations of Nonsmooth Functions

As discussed in lecture four, subgradient methods find ϵ-approximate solutions to nonsmooth convex optimization problems in $O(\frac{1}{\epsilon^2})$ oracle calls. The subgradient method is poor in the sense that this bound cannot be improved upon for general black box models of objective functions; however, even simple problems like

$$\min_{x\in\mathbb{R}^n}\left\{\max_{1\le i\le k} x_i + \frac{\mu}{2}\|x\|_2^2\right\}, \qquad k=1,2,\dots,n,$$

where we understand the structure, are challenging for all numerical methods to optimize. This lecture is motivated in looking beyond black-box models and leveraging the underlying problem structure to develop more efficient schemes. In particular, we explore smooth approximations of nonsmooth functions.

Consider a convex function $f(\cdot)$ over a compact set $\mathbb{E}$ satisfying the following growth condition globally:

$$f(x) \le f(0) + L\|x\|,$$

where $\|x\| = \langle Bx, x\rangle^{1/2}$ and $B:\mathbb{E}\to\mathbb{E}^*$ is a self-adjoint positive definite linear operator (we only considered the identity map in class). Moreover, we define the Fenchel conjugate of the function by

$$f_*(s) = \max_{x\in\mathbb{E}}\{\langle s,x\rangle - f(x)\}, \qquad s\in\mathbb{E}^*.$$

This latter function is interesting in its own right and we can trivially deduce the following facts regarding its behavior.

Fact 1. *dom(f_*) $\supseteq \partial f(y)$ for all $y\in\mathbb{E}$.*

Proof For any $g_y \in \partial f(y)$, we have

$$\begin{aligned} f_*(g_y) &= \max_x\{\langle g_y, x\rangle - f(x)\} \\ &= \max_x\{\langle g_y, x-y\rangle + \langle g_y, y\rangle - (f(x)-f(y)) - f(y)\} \\ &= \langle g_y, y\rangle - f(y), \end{aligned}$$

where the last equality stems from convexity of f: $\langle g_y, x-y\rangle - (f(x)-f(y)) \le 0$. □

Fact 2. *$\forall s\in$ dom(f_*), we have $\|s\|\le L$.*

Proof

$$\begin{aligned} f_*(s) &= \max_x\{\langle s,x\rangle - f(x)\} \\ &\ge \max_x\{\langle s,x\rangle - L\|x\|\} \\ &\ge t(\|s\|^2 - L\|s\|), \end{aligned}$$

where in the last inequality we set $s = tx$ in $\{\langle s,x\rangle - L\|x\|\}$. If $\|s\| > L$, we know $\|s\|^2 - L\|s\| > 0$. Letting $t\to\infty$, we can show $f_*(s)$ is unbounded which is a contradiction. □

Fact 3. *For $x \in \mathbb{E}$, $g \in \partial f(x)$, we have $f(x) + f_*(g) = \langle g, x \rangle$.*

Proof Rewrite g as g_x, and by the proof of Fact 1, we have $f_*(g_x) = \langle g_x, x \rangle - f(x)$, that is,

$$f(x) + f_*(g_x) = \langle g_x, x \rangle.$$

□

Fact 4. *For $s \in dom(f_*)$, $f_*(s) \geq f_*(g) + \langle s - g, x \rangle$.*

Proof

$$\begin{aligned} f_*(s) &= \max_s \{\langle s, x \rangle - f(x)\} \\ &\geq \langle s, x \rangle - f(x) \\ &= f_*(g) + \langle s - g, x \rangle, \end{aligned}$$

where the last equality follows from fact 3. □

This last fact equivalently says that for all $g_x \in \partial f(x)$, $x \in \partial f_*(g_x)$, and furthermore from this fourth fact, we can construct an equivalent formulation of f using the Fenchel conjugate as follows

$$f(x) = \max_{x \in \text{dom}(f_*)} \{\langle x, s \rangle - f_*(s)\}. \tag{1}$$

Now, we can define a crude, smooth approximation of this newly formulated f by subtracting a quadratic term in $\|s\|^*$:

$$f_\mu(x) = \max_{x \in \text{dom}(f_*)} \left\{\langle x, s \rangle - f_*(s) - \frac{\mu}{2}(\|s\|^*)^2\right\},$$

where $\mu \geq 0$ is a smoothing parameter and $\|s\|^* = \langle s, B^{-1}s \rangle^{1/2}$ (again $B = I$ in lecture - so we will just refrain from using the dual norm from now on). This approximation is pointwise upper bounded by f using (1), and moreover from fact 2, we have that

$$f_\mu(x) \geq f(x) - \frac{1}{2}\mu L^2.$$

As seen in the following theorem, it can be shown that ∇f_μ is μ^{-1}−Lipschitz which affirms this is indeed a smoothing of f.

Theorem 5 (Lemma 6.1.2 in (1))**.** *The function f_μ is differentiable on $\mathbb{E}$, and for all $x_1, x_2 \in \mathbb{E}$,*

$$\|\nabla f_\mu(x_1) - \nabla f_\mu(x_2)\|_2 \leq \frac{1}{\mu}\|x_1 - x_2\|_2.$$

Proof For $i = 1, 2$, define the points

$$s_i^* = \operatorname*{arg\,max}_{s \in \text{dom}(f_*)} \{\langle x_i, s \rangle - f_*(s) - \frac{\mu}{2}\|s\|_2^2\}$$

which are uniquely determined as f_μ is strongly concave. By first-order optimality conditions, there exists vectors $\tilde{x}_i \in \partial f_*(s_i^*)$, such that for all $s \in \text{dom}(f_*)$,

$$\langle s - s_i^*, x_i - \tilde{x}_i - \mu s_i^* \rangle \leq 0.$$

Letting $g(x,s) = \langle x, s\rangle - f^*(s) - \frac{\mu}{2}\|s\|_2^2$ implies this can be rewritten as

$$\langle s - s_i^*, \nabla g(x_i, s_i^*)\rangle \leq 0.$$

Then subbing in $s = s_i^*$ yields two inequalities,

$$\langle s_2^* - s_1^*, x_1 - \tilde{x}_1 - \mu s_1^*\rangle \leq 0, \text{ and } \langle s_1^* - s_2^*, x_2 - \tilde{x}_2 - \mu s_2^*\rangle \leq 0, \tag{2}$$

that when combined together yield

$$\langle s_2^* - s_1^*, (x_2 - \tilde{x}_2 - \mu s_2^*) - (x_1 - \tilde{x}_1 - \mu s_1^*)\rangle \leq 0 \iff \langle s_1^* - s_2^*, \mu(s_1^* - s_2^*)\rangle \leq \langle s_1^* - s_2^*, \mu(s_1^* - s_2^*)\rangle. \tag{3}$$

From which, we deduce that

$$\begin{aligned}
\mu\|s_1^* - s_2^*\|_2^2 &= \langle s_1^* - s_2^*, \mu s_1^* - \mu s_2^*\rangle \\
&\leq \langle s_1^* - s_2^*, x_1 - \tilde{x}_1 - (x_2 - \tilde{x}_2)\rangle && \textit{(3) above} \\
&= \langle s_1^* - s_2^*, x_1 - x_2\rangle - \langle s_1^* - s_2^*, \tilde{x}_1 - \tilde{x}_2\rangle \\
&\leq \langle s_1^* - s_2^*, x_1 - x_2\rangle && \textit{(2) above} \\
&\leq \|s_1^* - s_2^*\|_2 \|x_1 - x_2\|_2 && \textit{Cauchy-Schwarz}
\end{aligned}$$

which implies that $\|s_1^* - s_2^*\|_2 \leq \mu^{-1}\|x_1 - x_2\|_2$. By lemma 3.1.10 and theorem 3.1.14 in (1), we have that

$$\nabla f_\mu(x_i) = s_i^*, \text{ for } i = 1, 2.$$

and thus

$$\|\nabla f_\mu(x_1) - \nabla f_\mu(x_2)\|_2 \leq \frac{1}{\mu}\|x_1 - x_2\|_2.$$

□

For nonsmooth convex f, $f = f_{**}$ and moreover

$$f_{**} \approx \max_{s \in \mathrm{dom}(f_*)} \langle x, s\rangle - f_*(s) - \frac{\mu}{2}\|s\|_2^2 = f_\mu(x).$$

If $\mu = O(\epsilon)$ and f satisfies the growth condition, we have $|f(x) - f_\mu(x)| = O(\epsilon)$.

2 Minimax model of an Objective Function

Let us explore a candidate example for smoothing - the classic minimax problem whose structure is given by

$$\min_{x \in Q_1} f(x)$$

where $Q_1 \subseteq \mathbb{E}_1$,

$$f(x) = \hat{f}(x) + \max_{u \in Q_2} \left\{ \langle Ax, u\rangle - \hat{\phi}(u) \right\},$$

and $Q_2 \subseteq \mathbb{E}_2$ (Q_1 and Q_2 are bounded closed convex sets). Additionally $A : \mathbb{E}_1 \to \mathbb{E}_2^*$ is a linear operator and, $\hat{\phi}, \hat{f}$ are convex and continuous. In lecture, we discussed the case where f is convex, $\hat{f} = 0$, $\hat{\phi}(u) = f_*(u)$, and A is identity. Then the objective function reduces to the form

$$f(x) = \max_{u \in Q_2} \{\langle x, u\rangle - f_*(u)\} = f_{**}(x).$$

Nesterov continues discussion here on the adjoint problem and examples of using prox-functions to simplify objective functions in section 6.1.2 of (1), but we pivot our discussion now towards a worked out concrete example of smoothing.

2.1 Matrix Games

Consider the following saddle point problem:

$$\min_{x\in\Delta_n}\max_{u\in\Delta_m}\{\langle Ax,u\rangle+\langle c,x\rangle+\langle b,u\rangle\}$$

where $A:\mathbb{R}_n\to\mathbb{R}_m$, and Δ_n denotes the simplex on $\mathbb{R}^n$:

$$\Delta_n=\left\{x\in\mathbb{R}^n_+:\sum_{i=1}^n x_i=1\right\}.$$

In this game, the row and column players have their respective non-smooth minimization problems:

$$\min_{x\in\Delta_n} f(x):f(x)=\langle c,x\rangle+\max_{u\in\Delta_m}\{\langle Ax,u\rangle+\langle b,u\rangle\}$$
$$\max_{u\in\Delta_m}\phi(u):\phi(u)=\langle b,u\rangle+\min_{x\in\Delta_n}\{\langle Ax,u\rangle+\langle c,x\rangle\}$$

The non-smoothness of the objective comes from the simplex constraint. To see how this game can be optimized with the tools discussed above, we look at the closely related problem and its dual

$$\begin{aligned} f(x)&=\max_{1\le j\le m}|\langle a_j,x\rangle-b_j|,\\ f^*(u)&=\max_x\left\{\langle u,x\rangle-\max_{1\le j\le m}|\langle a_j,x\rangle-b_j|\right\}.\end{aligned}$$

We manipulate the dual form below

$$\begin{aligned} f^*(u)&=\max_x\left\{\langle u,x\rangle-\max_{1\le j\le m}|\langle a_j,x\rangle-b_j|\right\}\\ &=\max_x\left\{\langle u,x\rangle-\max_{s:||s||_1\le 1}\left\{\sum_{j=1}^m s_j\left(\langle a_j,x\rangle-b_j\right)\right\}\right\}\\ &=\max_x\min_{s:||s||_1\le 1}\left\{\langle u,x\rangle-\sum_{j=1}^m s_j\left(\langle a_j,x\rangle-b_j\right)\right\}\end{aligned}$$

where the first line is equivalent since we are simply picking out the largest index, and the second line we pulled the max out and turned it into a min due to the minus sign on the term. Now we can use Sion's Minimax theorem to further simplify the expression

$$\begin{aligned} f^*(u)&=\max_x\min_{s:||s||_1\le 1}\left\{\langle u,x\rangle-\sum_{j=1}^m s_j\left(\langle a_j,x\rangle-b_j\right)\right\}\\ &=\min_{s:||s||_1\le 1}\max_x\left\{\langle u,x\rangle-\sum_{j=1}^m s_j\left(\langle a_j,x\rangle-b_j\right)\right\}\\ &=\min_{s:||s||_1\le 1}\left\{\sum_{j=1}^m s_jb_j:As=u\right\}\end{aligned}$$

where we now have an alternate way of writing the dual. However, this is not the best choice, since we still have a L1 loss constraint that is not smooth.

We may consider a better choice of $\hat{\phi}$. We take our original constraint and instead manipulate it:

$$\begin{aligned} f(x) &= \max_{1 \le j \le m} |\langle a_j, x \rangle - b_j| \\ &= \max_{u^1, u^2 \in \mathbb{R}^{2m}} \left[\sum_{j=1}^{m} \left(u_j^1 - u_j^2 \right) [\langle a_j, x \rangle - b_j] \text{ subject to } \sum_{j=1}^{m} u_j^1 + u_j^2 = 1 \right] \end{aligned}$$

where we have written f as a linear function under the maximum over the entire domain and not just the simplex. This alternate way of writing gives the form we are interested in:

$$f(x) = \hat{f}(x) + \max_{x \in Q_2} \left\{ \langle Ax, u \rangle - \hat{\phi}(u) \right\}$$

where $\langle Ax, u \rangle$ corresponds with the $\langle a_j, x \rangle$ term and $\hat{\phi}(u)$ corresponds with b_j term and thus smoothing the objective. In (1), Nesterov employed theorem 6.1.3 from here to yield complexity bounds of $O(\frac{1}{\epsilon})$ which is much faster than naive subgradient methods as stated in the beginning.

3 Summary

To summarize, we looked at the smooth approximation to nonsmooth functions using the Fenchel dual representation. We wrote our objectives in the form

$$f(x) = \hat{f} + \max_{u} \left\{ \langle Ax, u \rangle - \hat{\phi}(u) \right\}$$

which is a nice functional form to optimize. A more in-depth treatment can be found in chapters three and six of (1).

References

[1] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer, 2018.