TCS Guide of Convex Optimization - Day 2 Binary Search

January 31, 2023

1 Cutting Plane Methods

For any convex f and any current x, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ for all } y. \tag{1.1}$$

Let x^* be any minimizer of f. Replacing y with x^*, we have that

$$f(x) \geq f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle .$$

Therefore, we know that $\langle \nabla f(x), x^* - x \rangle \leq 0$.

Fact 1. *For any x, x^* lies in the halfspace*

$$H_x \stackrel{\text{def}}{=} \{y : \langle \nabla f(x), y - x \rangle \leq 0\}.$$

Question: How to use it to do binary search in $\mathbb{R}^n$?
Cutting Plane Framework:

- **Invariant:** Maintain $X^* \subset \Omega^{(k)}$ where X^* is the set of minimizers of f.
- Pick some large enough $\Omega^{(0)}$.
- For $k = 0, 1, \cdots$
  - If $\Omega^{(k)}$ is small enough, **return** the point with smallest function value we ever queried.
  - Pick some $x^{(k)} \in \Omega^{(k)}$.
  - Note that $X^* \subset H_{x^{(k)}}$. So, we know $X^* \subset H_{x^{(k)}} \cap \Omega^{(k)}$.
  - Pick some $\Omega^{(k+1)} \supset H_{x^{(k)}} \cap \Omega^{(k)}$.

To analyze the algorithm, the main questions we need to answer are:

1. How do we choose $\Omega^{(k+1)}$ and $x^{(k)}$?
2. How do we measure progress?
3. How quickly does the method converge?
4. How expensive is each step?

2 Center of Gravity Method

Discovered by Levin and Newman on 1965 independently on opposite side of iron curtain.

1. The algorithm does not forget any information, namely $\Omega^{(k+1)} = \Omega^{(k)} \cap H_{x^{(k)}}$. It uses the center of gravity of $\Omega^{(k)}$ for $x^{(k)}$.

$$\text{center of gravity of } \Omega = \mathbb{E}_{x \sim \Omega} x$$

2. Measure progress by $\text{vol}\Omega^{(k)}$.

3. Each step, the volume goes down by $1 - \frac{1}{e}$ factor.

4. However, the runtime for each step is horrible.

The following theorem show the volume is cut by at least $1/e$ factor each step.

Theorem 2. *(Grunbaum Theorem) Let K be a convex body in $\mathbb{R}^n$ with center of gravity z. Let H be any halfspace containing z. Then,*

$$\mathrm{vol}(K \cap H) \geq \left(\frac{n}{n+1}\right)^n \mathrm{vol}(K) \geq \frac{1}{e}\mathrm{vol}(K).$$

Proof. WLOG, $H = \{x_1 = 0\}$. Let $K_t = K \cap \{x_1 = t\}$ be the slices of K. Brunn–Minkowski theorem shows that the "volume radius"

$$r(t) \stackrel{\text{def}}{=} \mathrm{vol}(K_t)^{\frac{1}{n-1}}$$

is concave. Namely, K must be either a cone or more concave than cone in terms of volume radius. It reduces to this 1-dimension question:

Given a 1-dimension pdf p such that the mean is 0 and that $p^{\frac{1}{n-1}}$ is concave, solving it, we have

$$\mathbb{P}(t \geq 0) \geq \left(\frac{n}{n+1}\right)^n.$$

□

Exercise 3. Read a detailed proof of Grunbaum's theorem.

Remark: Since we decreases the volume by constant factor every step, to decrease the volume from unit ball to ϵ-size ball requires $n\log(1/\epsilon)$ steps. This is where the factor n comes from.

Exercise 4. Recall what is ellipsoid method. Suppose we are given a function f such that $I \preceq \nabla^2 f \preceq 2I$. Show how to modify the ellipsoid method such that it takes $O(\log(1/\epsilon))$ steps and each steps take $O(n)$ time.

- Part 1) Shows that $f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}\|y-x\|^2 \leq f(x) \leq f(x) + \nabla f(x)^\top (y-x) + \|y-x\|^2$.
- Part 2) Using that, shows that given $\nabla f(x)$, $f(x)$ and $f(x^*)$, show that x^* contains in a ball bounded away from x.
- Part 3) Using above to design a new variant of ellipsoid method that converges in $O(\log(1/\epsilon))$ steps.

3 Extra: From Volume to Function Value

In the last section, we talk about how to decrease the volume. But how about how to decrease the objective value? The following lemma shows the volume and the objective value are essentially same.

Theorem 5. *Let $x^{(k)}$ be the sequence of points produced by the cutting plane framework for a convex function f. Let $\mathcal{R}$ be a mapping from subsets of $\mathbb{R}^n$ to non-negative numbers satisfying*

1. *(Linearity) For any set $S \subseteq \mathbb{R}^n$, any vector y and any scalar $\alpha \geq 0$, we have $\mathcal{R}(\alpha S + y) = \alpha \mathcal{R}(S)$ where $\alpha S + y = \{\alpha x + y : x \in S\}$.*
2. *(Monotonic) For any set $T \subset S$, we have that $\mathcal{R}(T) \leq \mathcal{R}(S)$.*

Then, we have that

$$\min_{i=0,1,\cdots,k-1} f(x^{(i)}) - \min_{y \in \Omega^{(0)}} f(y) \leq \frac{\mathcal{R}(\Omega^{(k)})}{\mathcal{R}(\Omega^{(0)})} \cdot \left(\max_{z \in \Omega^{(0)}} f(z) - \min_{x \in \Omega^{(0)}} f(x)\right) .$$

Remark 6. We can think $\mathcal{R}(\Omega)$ as some way to measure the size of Ω. It can be radius, mean-width or any other way to measure "size". For the center of gravity method, we use $\mathcal{R}(\Omega) = \mathrm{vol}(\Omega)^{\frac{1}{n}}$ for which we have proved volume decrease before. We raise the volume to power $1/n$ to satisfy linearity.

Proof. For simplicity, we define $\Omega = \Omega^{(0)}$. Let x^* be any minimizer of f over Ω. For any $\alpha > \frac{\mathcal{V}(\Omega^{(k)})}{\mathcal{V}(\Omega)}$ and $S = (1-\alpha)x^* + \alpha\Omega$, by the linearity of $\mathcal{V}$, we have that

$$\mathcal{V}(S) = \alpha\mathcal{V}(\Omega) > \mathcal{V}(\Omega^{(k)})$$

Therefore, S is not a subset of $\Omega^{(k)}$ and hence there is a point $y \in S\backslash\Omega^{(k)}$. y is not in $\Omega^{(k)}$. This means it is separated by the gradient at some step $i \le k$, namely for some $i \le k$, we have

$$\nabla f(x^{(i)})^\top (y - x^{(i)}) > 0.$$

By the convexity of f, it follows that $f(x^{(i)}) \le f(y)$. Since $y \in S$, we have $y = (1-\alpha)x^* + \alpha z$ for some $z \in \Omega$. Thus, the convexity of f implies that

$$f(x^{(i)}) \le f(y) \le (1-\alpha)f(x^*) + \alpha f(z).$$

Therefore, we have

$$\min_{i=1,2,\cdots k} f(x^{(i)}) - \min_{x\in\Omega} f(x) \le \alpha\left(\max_{z\in\Omega} f(z) - \min_{x\in\Omega} f(x)\right).$$

Since this holds for any $\alpha > \frac{\mathcal{V}(\Omega^{(k)})}{\mathcal{V}(\Omega)}$, we have the result. □

Combining with the last section, we have that the center of gravity method satisfies

$$\min_{i=0,1,\cdots,k-1} f(x^{(i)}) - \min_{y\in\Omega^{(0)}} f(y) \le \left(1 - \frac{1}{n+1}\right)^k \cdot \left(\max_{z\in\Omega^{(0)}} f(z) - \min_{x\in\Omega^{(0)}} f(x)\right) .$$

4 What do we know now?

4.1 Upper bound

$\Omega^{(k)}$	$x^{(k)}$	Iter up to log	Cost/Iter
$\Omega^{(k+1)} = \Omega^{(k)} \cap H^{(k)}$	Center of gravity	$n\log(1/\epsilon)$	n^3
$\Omega^{(k+1)}$ =smallest ellipsoid containing $\Omega^{(k)} \cap H^{(k)}$	Center of ellipsoid	$n^2\log(1/\epsilon)$	n^2
$\Omega^{(k+1)} = \Omega^{(k)} \cap H^{(k)}$	Center of John ellipsoid	$n\log(1/\epsilon)$	$n^{2.878}$
Some polytope having only $O(n)$ constraints	Volumetric center	$n\log(n/\epsilon)$	n^2

All the above algorithms measure progresses by volume of something. So, the total runtime for convex optimization is $n^3\log(1/\epsilon)$.

Theorem 7. *(Jiang Lee Song Wong 2020) Given a convex function f and a convex set K such that we can compute the gradient of f and the separating hyperplane for K. Suppose we know the minimizer $\|x^*\|_2 \le 1$, then we can find $x \in K$ such that*

$$f(x) \le \min_{x\in K} f(x) + \epsilon(\max_{x\in B} f(x) - \min_{x\in B} f(x))$$

in $n\log(\frac{n}{\epsilon vol(K)})$ iterations, each iteration takes $O(n^2)$ time plus 1 call to the oracle.

4.2 Lower bound

Sorry, I don't know exactly. The classical proof is $\Omega(n\log(1/\epsilon))$ for the question of finding a point in convex set. For convex optimization, I can only come up with $\frac{n\log\frac{1}{\epsilon}}{\log n}$ at this last minute. The function is this:

$$f(x) = \|x - x^*\|_\infty.$$

Note that the gradient of x is almost always 1-sparse vector. So, it contains just $\log n$ bit of information. Encoding a ϵ-approximate solution for x^* requires $n\log\frac{1}{\epsilon}$ bit. Hence the lower bound.

5 Discussion

Among different proofs above, one common theme is to prove that a convex set K is approximated by some ellipsoid. Here are two of such theorem:

Theorem 8. *Given a convex set $K \subset \mathbb{R}^n$. Let E be the ellipsoid such that $x_0 \stackrel{\text{def}}{=} \mathbb{E}_{x\sim K} x = \mathbb{E}_{x\sim E} x$ and $\mathbb{E}_{x\sim K} xx^\top = \mathbb{E}_{x\sim E} xx^\top$. Then, we have that*

$$\sqrt{\frac{n+1}{n}}(E - x_0) \subset K - x_0 \subset \sqrt{n(n+1)}(E - x_0).$$

Theorem 9. *Given a convex set $K \subset \mathbb{R}^n$. Let E be largest volume ellipsoid in K with center x_0 , we have*

$$E - x_0 \subset K - x_0 \subset n(E - x_0).$$

6 Extra: What if we need exact solution?

Theorem 10. *(Jiang 2022) Given an explicit lattice Λ and some convex set Ω, we can find an exact minimizer of the problem*

$$\min_{x\in\Lambda\cap\Omega} f(x)$$

using $O(n \log n + \log \frac{\text{vol}\Omega}{\det \Lambda})$ calls to ∇f. In particular, if $\Lambda = \mathbb{Z}^n$ and Ω is the unit ball, the number of calls is $O(n \log n)$ calls. (Unfortunately, the runtime is sub-exponential.)

Idea: If the current region $\Omega^{(k)}$ is very narrow on some direction, then $\Omega^{(k)} \cap \Lambda$ should be lower dimension and hence we can reduce the dimension by 1 by recursing on that lower dimensional subspace.

Suppose for simplicity that the current region Ω is a unit ball, the following lemma shows when we can reduce the dimension.

Theorem 11. *If $v \in \Lambda^* \backslash \{0\}$ such that $\|v\|_2 < \frac{1}{2}$, then, we have*

$$\Lambda \cap \{\|x - x_0\|_2 \le 1\} \subset \{x : v^\top x = [v^\top x_0]\}$$

where $[v^\top x_0]$ is the rounding of $v^\top x_0$ to the closest integer.

Proof. First, we want to prove that for any $x_1, x_2 \in \Lambda$, we have

$$v^\top (x_1 - x_2) = 0.$$

Note that $x_1 - x_2 \in \Lambda$ and $v \in \Lambda^*$, by the definition of dual lattice, we have $v^\top (x_1 - x_2) \in \mathbb{Z}$. Next, we note that $\|x_1 - x_2\|_2 \le 2$ and hence

$$|v^\top (x_1 - x_2)| \le \|v\|_2 \|x_1 - x_2\|_2 < 1.$$

Hence, we have the claim. This proves that

$$\Lambda \cap \{\|x - x_0\|_2 \le 1\} \subset \{x : v^\top x = t\}$$

for some $t \in \mathbb{Z}$.

Now, it suffices to prove that $t = [v^\top x_0]$. Suppose there is some point $\overline{x} \in \Lambda \cap \{\|x - x_0\|_2 \le 1\}$ is non-empty. Note that $|v^\top (\overline{x} - x_0)| < \frac{1}{2}$. Since $\overline{x}$ in that set, we have $v^\top \overline{x} = t$ by definition. Hence, we have $|t - v^\top x_0| < \frac{1}{2}$. Since $t \in \mathbb{Z}$, we have $t = [v^\top x_0]$. □

To get the precise version, we need to find $v \in \Lambda^*$ such that $\|v\|_M < \frac{1}{2}$ for some matrix M depending on the current region. Using this, we can have the following algorithm:

- Loop
  - If the shortest vector v of Λ^* is smaller than $\frac{1}{10n}$ in some norm,
    - Reduce the dimension along normal vector v. Update Λ and Ω.
  - Else

$*$ Cut Ω using gradient oracle at the center of gravity of Ω.

He showed that the potential $\Phi = \text{vol}\Omega / \det \Lambda$

- Decreases a constant factor for every cut.
- Increases by $O(n)$ factor for every dimension reduction.
- $\Phi \geq \frac{1}{n^n}$ before every cut (Minkowski first theorem)

This proves the theorem.

Remark: This theorem is particular important as it gives the fastest (and probably tight) strongly polynomial runtime for submodular minimization.

7 Various Tips

7.1 Computing Gradient

To apply the cutting plane method, we need gradient. It turns out the cost of computing the gradient is essentially same as the cost of compute the function.

Theorem 12 (Auto diff)**.** *Suppose we have an algorithm to compute f exactly in time T, then, by modifying the algorithm, we can compute ∇f exactly in time $O(T)$.*

7.2 Computing Gradient

Imagine you have k machines. Each machine uses resources $(r_1, r_2, \cdots, r_k)$ and produces goods $(g_1, g_2, \cdots, g_\ell)$. We represent the machine by simply a vector $a = (r_1, r_2, \cdots, r_k, g_1, g_2, \cdots, g_\ell)$.

Say we have vectors $a_1, a_2, \cdots, a_n \in \mathbb{R}^{k+\ell}$, the polytope

$$K = \{\sum_{i=1}^{n} \lambda_i a_i \in \mathbb{R}^{k+\ell} | 0 \leq \lambda_i \leq 1 \text{ for all } i\}$$

represents the set of possible feasible (input, output) pairs. Many production problems can be written as

$$\min_{z \in K} f(z).$$

Suppose f is simple that you can do whatever on it, what is the best way to apply cutting plane method?

Problem: It is a bit expensive to simply check in a vector $z \in K$ or not, it involves solving an LP.

Definition 13. Let the dual of f defined by

$$f^*(\theta) = \max_x \theta^\top x - f(x).$$

Fact 14. *It is know that $f^{**} = f$.*

Hence, we can write the problem by

$$\begin{aligned}
\min_{z \in K} f(z) &= \min_{z \in K} f^{**}(z) \\
&= \min_{z \in K} \max_\theta z^\top \theta - f^*(\theta) \\
&= \max_\theta \min_{z \in K} z^\top \theta - f^*(\theta)
\end{aligned}$$

Why this is better? Note that we have

$$\nabla(\min_{z \in K} z^\top \theta - f^*(\theta)) = \arg\min_{z \in K} z^\top \theta - \arg\max_x (\theta^\top x - f(x))$$

First, note that $\arg\min_{z \in K} z^\top \theta$ is simple because

$$\begin{aligned}\min_{z \in K} z^\top \theta &= \min_{0 \le \lambda_i \le 1 \text{ for all } i} \sum_{i=1}^n \lambda_i a_i^\top \theta \\ &= \sum_{i=1}^n \min(a_i^\top \theta, 0)\end{aligned}$$

and the minimizer simply given by some explicit formula. Second, the second part is simple because we assume the f is "simple".