Lecture 12

Approximate Caratheodory Theorem using Mirror Descent

Caratheodory's theorem (exact version)

Let $u \in \mathbb{R}^{n}$ (dimension of the space) be a point which lies in the convex hull of a set of points $\{v_1, v_2, \dots, v_m\}$.
(number of points which define the convex hull)

Thus $u \in \operatorname{conv}(\{v_1, \dots, v_m\})$.

Then u can be written as a convex combination of at most $n+1$ points of $\{v_1, \dots, v_m\}$.

That is $u = \sum_{i=1}^{m} \lambda_i v_i$, $\quad \lambda_i \geq 0 \quad \sum \lambda_i = 1$

$\#\text{non zero } \lambda_i \leq n+1$

Note that $n \ll m$

The Theorem is tight. That's why the natural question to ask is if can use less points if we permit an ε-error.

Approximate Caratheodory's Therom

for any ℓ_p-norm with $p \geq 2$, convex hull $\mathrm{conv}(\{v_1, v_2, \dots, v_m\}) \subseteq \mathbb{R}^n$ with $\|v_i\|_p \leq 1$ ($v_i \in B_p^n \; \forall i \in [m]$), and point $u \in \mathrm{conv}(\{v_1, \dots, v_m\})$ there exists $\tilde{u}$ s.t $\|u - \tilde{u}\|_p \leq \varepsilon$ where $\tilde{u}$ is in the convex hull of at most $\frac{4p}{\varepsilon^2}$ points of $\{v_1, v_2, \dots, v_m\}$. That means

just to denote a permutation of the set $\{v_1, \dots, v_m\}$

$$\tilde{u} \in \mathrm{conv}(\{\tilde{v}_1, \dots, \tilde{v}_{\frac{4p}{\varepsilon^2}}\}) \subseteq \mathrm{conv}(\{v_1, \dots, v_m\})$$

observations

1. if $\max \|v_i\|_p \leq \alpha$ then the # of points $\leq \frac{4p\alpha^2}{\varepsilon^2}$

2. $\frac{4p}{\varepsilon^2}$ is independent of the dimension n of the space.

3. the original proof of the theorem is probabilistic and non-constructive.

Optimization formulation

$$V = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & & | \end{bmatrix}, \quad x \in \Delta^m$$ ← simplex

we want to solve $\min \|Vx - u\|_p$

s.t x is sparse

$x \in \Delta^m$

Problem: we do not know how to model the sparsity requirement

Idea: solve the unconstrained problem iteratively using gradient descent/mirror descent and hope that the solution is sparse.

this idea will fail since the gradient update may not be sparse

we need to formulate the problem differently.

Reformulation using Sion's minimax theorem

Remember that the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p}+\frac{1}{q}=1$. From the dual norm definition we get that

$$\|Vx-u\|_p = \max\{\langle y, Vx-u\rangle \mid \|y\|_q \le 1\} =$$

$$= \max\{\langle y, Vx-u\rangle \mid y \in B_q^m\}$$

$$\Longrightarrow \min_{x\in\Delta^m} \|Vx-u\|_p = \min_{x\in\Delta^m} \max_{y\in B_q^m} \langle y, Vx-u\rangle =$$

$$\underset{\text{minimax theorem}}{\overset{\text{Sion's}}{=}} \max_{y\in B_q^m} \min_{x\in\Delta^m} \langle y, Vx-u\rangle =$$

$$= \max_{y\in B_q^m}\left(-\max_{x\in\Delta^m} \langle y, u-Vx\rangle\right) =$$

$$= -\min_{y\in B_q^m} \underbrace{\max_{x\in\Delta^m} \langle y, u-Vx\rangle}_{f(y)} =$$

$$= -\min_{y\in B_q^m} f(y)$$

Now the unconstrained problem is equivalent to solving:

$$\min_{y \in B^q} f(y) \quad , \quad f(y) = \max_{x \in \Delta^m} \langle y, u - Vx \rangle$$
$$= \langle y, u - Vx^y \rangle$$
$$, x^y = \arg\max_{x \in \Delta^m} \langle y, u - Vx^y \rangle$$

Let's calculate some (sub)gradients:

$x^y = \arg\max_{x \in \Delta^m} -\langle V^T y, x \rangle \Rightarrow x^y$ is the vector of all zeros except for a 1 at the coordinate $i = \arg\max \{-u_i^T y\}$

$\to$ this guy (which is a dual certificate of the new optimization formulation) is sparse

we abuse notation even if it is a subgradient

$$\nabla f(y) = u - Vx^y =$$
$$= u - u_i$$

<u>Proof</u>

$$f(z) \geq f(y) + \langle u - Vx^y, z - y \rangle \Rightarrow$$
$$\Rightarrow f(z) \geq \cancel{f(y) - \langle u - Vx^y, y \rangle} + \langle u - Vx^y, z \rangle \Rightarrow$$
$$\Rightarrow \max_{x \in \Delta^m} \langle z, u - Vx \rangle \geq \langle z, u - Vx^y \rangle \quad \blacksquare$$

M.D theorem

$\eta = \dfrac{\sqrt{a\, D_\varphi(x^*, x_0)}}{L\sqrt{T}}$

$D_\varphi(x, y) \geq a\|x-y\|_q^2$

f is L-Lipschitz w.r.t $\|\cdot\|_p$ then with an appropriate stepsize* we get

$$\frac{1}{T}\sum \langle \nabla f(y_t), y_t - y \rangle \leq 2\frac{\sqrt{D_\varphi(x^*, x_0)}\, L}{\sqrt{aT}}, \ \forall y \in B_q$$

in the original statement we have

$$\frac{1}{T}\sum \left(f(y_t) - f(y^*)\right)$$

because of the first order condition and because y^* is a minimizer

$$\frac{1}{T}\sum \langle \nabla f(y_t), y_t - y \rangle = \frac{1}{T}\sum \langle u - V_{i(t)}, y_t - y \rangle =$$

$$= \frac{1}{T}\sum \Big(\underbrace{\langle u - V_{i(t)}, y_t \rangle}_{f(y_t)} - \langle u - V_{i(t)}, y \rangle\Big) =$$

$$= \frac{1}{T}\sum f(y_t) - \frac{1}{T}\sum \langle u - V_{i(t)}, y \rangle \geq$$

$$\geq -\frac{1}{T}\sum \langle u - V_{i(t)}, y \rangle, \text{ since } f(y_t) \geq 0$$

$\left(f(y) = \max_{x \in \Delta^m} \langle y, u - Vx \rangle \geq 0\right.$ because $\exists\, x'$ s.t $u = Vx'$

So, using mirror descent we can get a bound of the form

$$-\frac{1}{T}\sum \langle u - v_{i(t)}, y\rangle \leq \bigcirc \quad \forall y \in B_q$$

$$\Rightarrow \langle y, \frac{1}{T}\sum v_{i(t)} - u\rangle \leq \bigcirc \quad \forall y \in B_q$$

by definition of the dual norm

$$\left\| \frac{1}{T}\sum v_{i(t)} - u \right\|_p \leq \bigcirc$$

$\left(\frac{1}{p} + \frac{1}{q} = 1 \right)$ (just a reminder)

Now it remains to calculate the Lipschitz parameter L and find a function φ s.t. $D_\varphi(x,y) \geq a\|x-y\|_q^2$

① we have $\|\nabla f(y_t)\|_p = \|u - v_{i(t)}\|_p \leq$

$\leq \|u\|_p + \|v_{i(t)}\|_p \leq 2$ (since $v_i \in B^q$ and u belongs to the convex hull of $\{v_1, \dots, v_m\}$

so $L = 2$

② For $1 \le q \le 2$, $\psi(y) = \frac{1}{2}\|y\|_q^2$, $y \in B^q$

(why should we try this function?

well. In the case of the classic l_2-norm

we have $\psi(y) = \frac{1}{2}\|y\|_2^2 \Rightarrow D_\varphi(x,y) = \frac{1}{2}\|x-y\|_2^2$

and now we need a bound of the

form $D_\varphi(x,y) \geq a\|x-y\|_q^2$)

For the domain B^q we have

$D_\varphi(x,y) \geq \frac{q-1}{2}\|x-y\|_q^2$ and

$D_\psi(y^*,0) \leq \max_{y \in B^q} D_\varphi(y,0) \leq 1/2$ (for $q=2$)

so $D_\varphi(x,y) \geq a\|x-y\|_q^2$, $a = \frac{q-1}{2} =$

$$\underline{\underline{\frac{1}{q}+\frac{1}{p}=1}} \quad \frac{\frac{p}{p-1}-1}{2} = \frac{p-p+1}{2(p-1)} = \frac{1}{2(p-1)}$$

So to get ε-close we need

$$T = \frac{4L^2 D_\varphi(y^*,0)}{a\varepsilon^2} \leq \frac{4 \cdot 2^2 \cdot 1/2}{\frac{1}{2(p-1)}\varepsilon^2} = O(P/\varepsilon^2) \text{ iterations}$$

↳ which is also the # of v_i's that we need to use