Lecture 1

Monday, February 22, 2021 2:55 PM

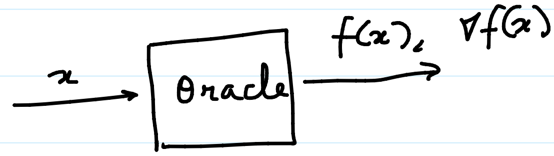
$\min_{x \in K} f(x)$, $K \subseteq \mathbb{R}^n$ convex set, $f: \mathbb{R}^n \to \mathbb{R}$ convex function

concretely, find x' s.t. $f(x') \le \min_{x \in K} f(x) + \varepsilon$.

can also allow some error in the membership in K.

<u>Input models for f:</u>

① <u>First order oracle:</u>

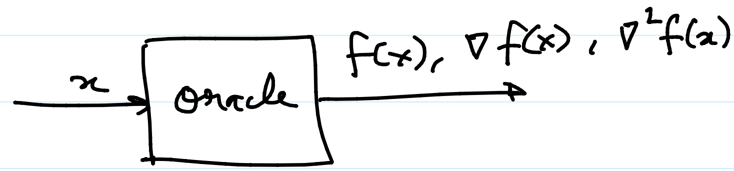


quite popular: typically gradient computation is pretty cheap

e.g. back propagation for neural networks.

example algorithm: gradient descent

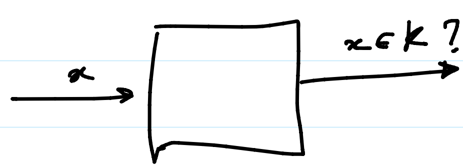
② <u>Second order oracle!</u>



example algorithm: Newton's algorithm

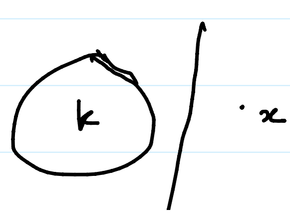
<u>Input models for K:</u>

① <u>Membership oracle:</u>

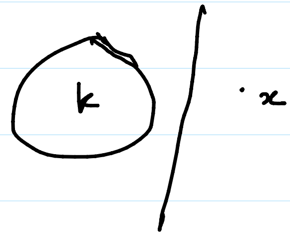


② <u>Separation oracle:</u>

if $x \notin K$, then output a separating hyperplane i.e hyperplane which separates x from K.



The hyperplane
separates x from K.



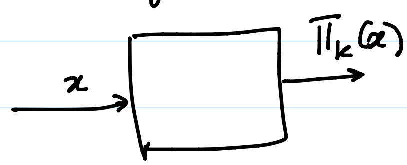
Example: $K = \{ x \in \mathbb{R}^n : \langle a_i, x \rangle \leq b_i \quad \forall\, i \in [m] \}$

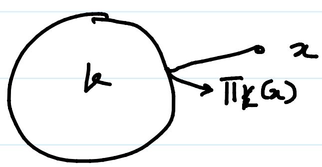
easy to implement both the oracles in time poly(n,m).

sometimes can implement these oracles even when m is exponentially larger than n.

e.g. matching polytope, spanning tree polytope

③ **Projection oracle:**





Parameters governing the running times for f

① Dimension n

② Smoothness of function f

Def: f is said to be L-smooth if the gradient of f is L-Lipschitz.

$$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \, \| x - y \|_2 \quad \forall\, x, y$$

L-Lipschitz, $\quad \|\nabla f(x) - \nabla f(y)\|_2 \le L \|x-y\|_2 \quad \forall x, y$

equivalently,

- $\|\nabla^2 f(x)\| \le L \quad \forall x$.
- $f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|x-y\|_2^2$

exercise

③ <u>Lipschitzness of f</u>: f is said to be G-Lipschitz if

$$|f(x) - f(y)| \le G \|x-y\|_2 \quad \forall x, y$$

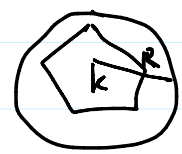
equivalently:

$$\|\nabla f(x)\|_2 \le G \quad \forall x$$

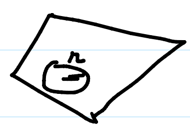
④ size of the optimal sol^n: $\quad x^* = \arg\min_{x \in K} f(x)$, $\quad \|x^*\|_2 \le D$

<u>Parameters for K</u>:

① Bounding radius R: $\quad K \subseteq B(0,R)$.



② $B(0,r) \subseteq K$



- $\min_{x \in \mathbb{R}^n} f(x)$

<u>Most basic algorithm</u>: Gradient descent

$f: \mathbb{R}^n \to \mathbb{R}$, $\quad \nabla f: \mathbb{R}^n \to \mathbb{R}^n$, $\quad \nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$.

$$\langle \nabla f(x), h \rangle = \frac{d}{dt} f(x+th)\Big|_{t=0}$$

What is the best direction to move in - that decreases the function value the most?

$$\min_{h : \|h\|_2 = 1} \langle \nabla f(x), h \rangle \quad , \qquad |\langle \nabla f(x), h \rangle| \underset{\text{Cauchy-Schwarz}}{\leq} \|\nabla f(x)\|_2 \, \|h\|_2 = \|\nabla f(x)\|_2$$

$$\text{for } h = - \frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

$$\text{then } \langle \nabla f(x), h \rangle = - \|\nabla f(x)\|_2$$

Gradient descent: Start with $x_0 = 0$, for T iterations,

$$x_{t+1} = x_t - \eta \nabla f(x_t) .$$

Assumptions: f is L-smooth, $x^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x)$, $\|x^*\|_2 \leq D$

Theorem: For an appropriate choice of η, gradient descent after $T = O\left(\frac{LD^2}{\varepsilon}\right)$ iterations outputs x_T s.t. $f(x_T) - f(x^*) \leq \varepsilon$.

Sanity checks: ① $g(x) = c f(x)$, $D_g = D_f$, $L_g = c L_f$

② $g(x) = f(\lambda x)$, $D_g = \frac{D_f}{\lambda}$, $L_g = \lambda^2 L_f$

Pf: Strategy: As long as $f(x_t) - f(x^*) \geq \varepsilon$, then $\|x_t - x^*\|_2^2$ will drop by a significant amount.

$$\|x_{t+1} - x^*\|_2^2 - \|x_t - x^*\|_2^2$$

$$= \|x_t - x^* - \eta \nabla f(x_t)\|_2^2 - \|x_t - x^*\|_2^2$$

$$= \|x_t - x^* - \eta \nabla f(x_t)\|_2^2 - \|x_t - x^*\|_2^2$$

$$= -2\eta \langle \nabla f(x_t), x_t - x^* \rangle + \eta^2 \|\nabla f(x_t)\|_2^2$$

want to bound this in terms of $f(x_{t+1}) - f(x^*)$.

We know that f is L-smooth.

$$f(x_{t+1}) - f(x_t) \le \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|_2^2$$

$$= -\eta \|\nabla f(x_t)\|_2^2 + \frac{L}{2}\eta^2 \|\nabla f(x_t)\|_2^2$$

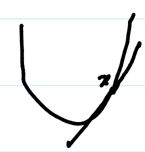
$$= \left(-\eta + \frac{L\eta^2}{2}\right)\|\nabla f(x_t)\|_2^2$$

$-1 + \eta L = 0$

$\eta = \frac{1}{L}$

$$= -\frac{\|\nabla f(x_t)\|_2^2}{2L}$$

$$\|\nabla f(x_t)\|_2^2 \le 2L\left(f(x_t) - f(x_{t+1})\right) \quad \text{— } 1$$



<u>f is convex:</u> $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

exercise

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

$$-\langle \nabla f(x_t), x_t - x^* \rangle \le -\left(f(x_t) - f(x^*)\right) \quad \text{— } 2$$

continuing from before:

$$-2\eta \langle \nabla f(x_t), x_t - x^* \rangle + \eta^2 \|\nabla f(x_t)\|_2^2$$

by 1 & 2

$$\le -2\eta\left(f(x_t) - f(x^*)\right) + 2\eta^2 L\left(f(x_t) - f(x_{t+1})\right)$$

by ① & ② $\leq -2\eta\,(f(x_t) - f(x^*)) + 2\eta^2 L\,(f(x_t) - f(x_{t+1}))$

$= -\frac{2}{L}(f(x_t) - f(x^*)) + \frac{2}{L}(f(x_t) - f(x_{t+1}))$

$= -\frac{2}{L}(f(x_{t+1}) - f(x^*))$

Hence we get, $\|x_{t+1} - x^*\|_2^2 - \|x_t - x^*\|_2^2 \leq -\frac{2}{L}(f(x_{t+1}) - f(x^*))$

if $f(x_{t+1}) - f(x^*) \geq \varepsilon$, then $\|x_{t+1} - x^*\|_2^2 - \|x_t - x^*\|_2^2 \leq -\frac{2\varepsilon}{L}$

Initially : $\|x_0 - x^*\|_2^2 \leq D^2$, $\|x_t - x^*\|_2^2 \geq 0$.

Hence after $T = \frac{LD^2}{2\varepsilon}$ iterations, it must be the case that $f(x_t) - f(x^*) \leq \varepsilon$ for some $t \in \{1, .., T\}$.

$f(x_t)$ is monotonically decreasing

$\Rightarrow f(x_T) - f(x^*) \leq \varepsilon$ after $T = \frac{LD^2}{2\varepsilon}$ iterations.

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

Projected Gradient descent: $x_0 = 0$, $x_{t+1} = \Pi_K(x_t - \eta \nabla f(x_t))$ do this for T iterations

Next lecture: ① Analysis of projected gradient descent

② Application to maximum flow problem