Lecture 3

Tuesday, March 9, 2021 3:28 PM

What if the convex function is Lipschitz?

Assumptions: ① $\|\nabla f(x)\|_2 \le G \; \forall x$, ② $\|x^0 - x^*\|_2 \le D$

Theorem: Grad descent with an appropriate choice of step-size η, will in $T = O\left(\left(\frac{GD}{\varepsilon}\right)^2\right)$ iterations output a sequence of points $x^0, x^1, \ldots, x^{T-1}$ s.t.

$$f\left(\frac{1}{T}\sum_{t=0}^{T-1} x^t\right) - f(x^*) \le \varepsilon$$

Remarks ① Different convergence guarantee than the smooth setting (where we had $f(x^T) - f(x^*) \le \varepsilon$).

② worse convergence guarantee ($\frac{1}{\varepsilon}$ vs $\frac{1}{\varepsilon^2}$). This is tight ($\Omega\left(\frac{GD}{\varepsilon}\right)^2$ lower bound in the first order oracle setting).

→ What if $\|\nabla f(x)\|_\infty$ was bounded (as opposed to $\|\nabla f(x)\|_2$)?

A very general recipe/algorithm: mirror descent that studies different bounded norm constraints on the gradient.

In this lecture, will study a special case: exponential gradient descent.

$$K = \Delta_n := \left\{ p \in [0,1]^n : \sum_{i=1}^n p_i = 1 \right\} \quad \text{(simplex)}$$

$f: \Delta_n \to \mathbb{R}$ convex, $\min_{p \in \Delta_n} f(p)$

A general strategy for optimization:

Construct a sequence of approximations f_t to f.

& set $p^{t+1} := \underset{p \in \Delta_n}{\text{argmin}} \; f_t(p)$

e.g. the gradient descent for smooth convex functions:

$$f_t(x) = f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{L}{2} \|x - x_t\|_2^2$$

$$f(x) \leq f_t(x) \quad (\text{if } f \text{ is } L\text{-smooth})$$

$$f(p_t) + \langle \nabla f(p_t), p - p_t \rangle \leq f(p) \quad (\text{by convexity of } f)$$

$f_t(p)$?

What about $p_{t+1} := \underset{p \in \Delta_n}{\text{argmin}} \left(f(p_t) + \langle \nabla f(p_t), p - p_t \rangle \right)$?

Too aggressive: keeps oscillating between the vertices.

So regularize!

$$p_{t+1} := \underset{p \in \Delta_n}{\text{argmin}} \left(M(p, p_t) + f(p_t) + \langle \nabla f(p_t), p - p_t \rangle \right)$$

M is some sort of distance measure on the simplex.

KL divergence: Let $p, q \in \Delta_n$. Then

$$D(p \| q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$$

(lets say log is the natural log)

the natural log)

Nice properties: ① $D(p\|q) \geq 0$

② $D(p\|q) \geq \frac{1}{2}\|p-q\|_1^2$ (Pinsker's ineq.)

③ Jointly convex in p, q

$$D((1-\lambda)p_1 + \lambda p_2 \,\|\, (1-\lambda)q_1 + \lambda q_2) \leq (1-\lambda) D(p_1\|q_1) + \lambda D(p_2\|q_2)$$

$$p_{t+1} := \underset{p \in \Delta_n}{\operatorname{argmin}} \left\{ D(p\|p_t) + \eta \left(f(p_t) + \langle \nabla f(p_t), p-p_t \rangle \right) \right\}$$

$$= \underset{p \in \Delta_n}{\operatorname{argmin}} \left\{ D(p\|p_t) + \eta \langle \nabla f(p_t), p \rangle \right\}$$

Lemma: $q \in \Delta_n$, $g \in \mathbb{R}^n$. Let $p^* = \underset{p \in \Delta_n}{\operatorname{argmin}} \{ D(p\|q) + \eta \langle g, p \rangle \}$

Then $p^* = \dfrac{w^*}{\|w^*\|_1}$, where $w_i^* = q_i \, e^{-\eta g_i}$

Proof: $D(p\|q) + \eta\langle g, p\rangle - D(p^*\|q) - \eta \langle g, p^* \rangle$

$$= D(p\|q) + \eta\langle g,p\rangle - \sum_i p_i^* \log\left(\frac{p_i^*}{q_i}\right) - \eta\langle g, p^*\rangle$$

$$= D(p\|q) + \eta\langle g,p\rangle - \sum_i p_i^* \log\left(\frac{e^{-\eta g_i}}{\|w^*\|_1}\right) - \eta\langle g, p^*\rangle$$

$$= D(p\|q) + \eta\langle g,p\rangle + \log(\|w^*\|_1)$$

$$D(p\|p^*) = \sum_i p_i \log\left(\frac{p_i}{p^*_i}\right)$$

$$= \sum_i p_i \log\left(\frac{p_i \quad \|w^*\|_1}{q_i\, e^{-\eta g_i}}\right)$$

$$= D(p\|q) + \log(\|w^*\|_1) + \eta \langle g, p\rangle$$

Hence $D(p\|q) + \eta\langle g,p\rangle - D(p^*\|q) - \eta\langle g, p^*\rangle$

$$= D(p\|p^*) \geq 0$$

↳ Jensen's ineq. & Concavity of log

Algorithm (Exponential gradient descent):

$p^0 := \frac{1}{n}\mathbb{1}$ (uniform distribution)

for $t = 0, 1, \ldots T-1$

$g^t := \nabla f(p^t)$

$w_i^{t+1} := p_i^t \, e^{-\eta g_i^t}$

$p_i^{t+1} := w_i^{t+1} / \|w^{t+1}\|_1$

return $\bar{p} = \frac{1}{T}\sum_{t=0}^{T-1} p^t$

Assumption: $\|\nabla f(p)\|_\infty \leq G \;\; \forall p \in \Delta_n$

Thm: with an appropriate choice of η & $T = \Theta\left(\frac{G^2 \log(n)}{\varepsilon^2}\right)$

EGD returns $\bar{p}$ s.t.

$$f(\bar{p}) - f(p^*) \leq \varepsilon$$

$$\left(p^* = \underset{p\in\Delta_n}{\arg\min}\, f(p)\right)$$

Jensen's ineq.

Proof: $f(\bar{p}) - f(p^*) \overset{\text{Jensen's ineq.}}{\leq} \left(\frac{1}{T}\sum_{t=0}^{T-1} f(p^t)\right) - f(p^*)$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\left(f(p^t) - f(p^*)\right)$$

$$\overset{\text{Convexity of } f}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} \langle \nabla f(p^t), p^t - p^* \rangle$$

$$= \boxed{\frac{1}{T}\sum_{t=0}^{T-1} \langle g^t, p^t - p^* \rangle}$$

Strategy: measure decrease in the KL divergence to the optimal solution

$$D(p^* \| p^t) - D(p^* \| p^{t+1})$$

$$= D(p^* \| p^t) - \sum_i p^*_i \log\left(\frac{p^*_i}{p^{t+1}_i}\right)$$

$$= D(p^* \| p^t) - \sum_i p^*_i \log\left(\frac{p^*_i \, \|w^{t+1}\|_1}{p^t_i \, e^{-\eta g^t_i}}\right)$$

$$= \eta \langle g^t, p^* \rangle - \log\left(\|w^{t+1}\|_1\right)$$

$$\|w^{t+1}\|_1 = \sum_i p^t_i \, e^{-\eta g^t_i}$$

$$\log\left(\sum_i p^t_i e^{-\eta g^t_i}\right) \overset{\text{Jensen's}}{\geq} \sum_i p^t_i \log\left(e^{-\eta g^t_i}\right) = -\eta \langle g^t, p^t \rangle$$

↳ want to lower bound the decrease in KL divergence

& hence need an upper bound on this term

$$\boxed{\underline{\text{Useful fact}}: \quad e^{-x} \leq 1 - x + \frac{e x^2}{2} \quad \forall\ |x| \leq 1}$$

<u>We know</u>: $\|g^t\|_\infty \leq G$

$$\Rightarrow \quad e^{-\eta g_i^t} \leq 1 - \eta g_i^t + \frac{e \eta^2 G^2}{2} \quad \text{if} \quad \eta \leq \frac{1}{G}.$$

(using the above fact)

$$\Rightarrow \quad \sum_i p_i^t e^{-\eta g_i^t} \leq 1 - \eta \langle g^t, p^t \rangle + \frac{e\eta^2 G^2}{2}$$

$$\Rightarrow \quad \log\left(\sum_i p_i^t e^{-\eta g_i^t} \right) \leq -\eta \langle g^t, p^t \rangle + \frac{e\eta^2 G^2}{2}$$

$$\left(\log(1+y) \leq y \right)$$

Hence
$$D(p^* \| p^t) - D(p^* \| p^{t+1}) \geq \eta \langle g^t, p^t - p^* \rangle - \frac{e\eta^2 G^2}{2}$$

summing up & telescoping

$$D(p^* \| p^0) - D(p^* \| p^T) \geq \eta \sum_{t=0}^{T-1} \langle g^t, p^t - p^* \rangle - \frac{e\eta^2 G^2 T}{2}$$

$$\frac{1}{T} \sum_{t=0}^{T-1} \langle g^t, p^t - p^* \rangle \leq \frac{D(p^* \| p^0) - D(p^* \| p^T)}{\eta T} + \frac{e\eta G^2}{2}$$

$\log(n)$ $D(p^* \| p^0)$

$$\leq \frac{\overset{\log(n)}{D(p^* \| p^0)}}{\eta T} + \frac{e\eta G^2}{2}$$

$$D(p^* \| p^0) = \log(n) - H(p^*) \leq \log(n)$$

$$\hookrightarrow \sum_i p_i^* \log\left(p_i^* / \tfrac{1}{n}\right) = \log(n) + \sum_i p_i^* \log(p_i^*)$$

$$= \log(n) - \underbrace{\left(\sum_i p_i^* \log\left(1/p_i^*\right)\right)}_{\geq 0}$$

$$\leq \frac{\log(n)}{\eta T} + \frac{e\eta G^2}{2}$$

pick η so that $\frac{\log(n)}{\eta T} = \frac{e\eta G^2}{2} \Rightarrow \eta = \Theta\left(\frac{\sqrt{\log(n)}}{G\sqrt{T}}\right)$

plugging this η

$$\Rightarrow \Theta\left(\frac{G\sqrt{\log(n)}}{\sqrt{T}}\right) \longrightarrow \text{we want this} \leq \varepsilon$$

so pick $T = \Theta\left(\frac{G^2 \log(n)}{\varepsilon^2}\right)$

This completes the proof.

Actually can replace p^* with any p & g^t's need not be gradients of same f.

Algorithm: Multiplicative weights update (MWU)

Algorithm: Multiplicate weights update (MWU)
essentially the same as EGD

Just that same oracle is providing these g^t's.

$$\frac{1}{T} \sum_{t=0}^{T-1} \langle g^t, p^t \rangle - \underset{p \in \Delta_n}{\operatorname{argmin}} \left(\frac{1}{T} \sum_{t=0}^{T-1} \langle g^t, p \rangle \right) \leq \varepsilon$$

Assumption: $\|g^t\|_\infty \leq G$.

Example setting:
- n experts in the stock market.
- Not sure which expert to follow

On day t, expert i incurs a loss of g_i^t.

Use MWU to predict their credibility from previous days data.

p^t only depends on $g^1, \ldots, g^{t-1}$

on day t, invest probabilistically acc. to p^t.

total expected loss (average over the days): $\frac{1}{T} \sum_{t=0}^{T-1} \langle g^t, p^t \rangle$

Guarantee! $$\frac{1}{T} \sum_{t=0}^{T-1} \langle g^t, p^t \rangle - \underset{p \in \Delta_n}{\operatorname{argmin}} \frac{1}{T} \sum_{t=0}^{T-1} \langle g^t, p \rangle \leq \varepsilon$$

regret