Gradient Descent & Mirror Descent

Recall from last time: if f is G-Lipschitz and $D = \|x_0 - x^*\|_2$ then basic GD satisfies $f(\hat{x}) - f(x^*) \le \varepsilon$ after $T = \left(\frac{GD}{\varepsilon}\right)^2$ steps.

Can skip proof if remember general idea from previous lecture

Sps f is α-strongly convex which means $f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\alpha}{2}\|y-x\|^2$ then we can get the convergence rate down to $T = \tilde{O}(1/\varepsilon)$. How?

Again: potential fn: $\Phi_t = \|x_t - x^*\|^2 \frac{1}{2\eta_{t-1}}$ but η_t changes over time. In fact: $\eta_t = \frac{1}{\alpha t}$

Amortized cost @ time t:

$$f(x_t) - f(x^*) + (\Phi_{t+1} - \Phi_t)$$

$$\cancel{\frac{\alpha t}{2}\|x_{t+1} - x^*\|^2 - \frac{\alpha(t-1)}{2}\|x_t - x^*\|^2}$$

but $x_{t+1} = x_t - \eta_t \cdot \nabla_t$

$$\Rightarrow \|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 + \|\eta_t \nabla_t\|^2 - 2\langle \eta_t \nabla_t, x_t - x^*\rangle$$

$$\frac{\alpha t}{2}\|x_{t+1} - x^*\|^2 - \frac{\alpha(t-1)}{2}\|x_t - x^*\|^2 \le \frac{\alpha}{2}\|x_t - x^*\|^2 + \frac{1}{2\eta_t}\cdot\eta_t^2\|\nabla_t\|^2 - \frac{1}{2\eta_t}2\langle \eta_t \nabla_t, x_t - x^*\rangle$$

$$\Rightarrow \text{amortized cost} \le \underbrace{f(x_t) - f(x^*) - \langle \nabla_t, x_t - x^*\rangle + \frac{\alpha}{2}\|x_t - x^*\|^2}_{\le 0 \text{ by } \alpha\text{-strong convexity of } f.} + \frac{\eta_t}{2}\|\nabla_t\|^2$$

$$\Rightarrow \cancel{\text{total cost}} \text{ average cost } \frac{1}{T}\sum_1^T f(x_t) \le f(x^*) + \underbrace{\frac{1}{T}\Phi_0 + \frac{1}{T}\sum_t^T \frac{\eta_t}{2}\|\nabla_t\|^2}$$

but $\Phi_0 = 0$ and $\|\nabla_t\| \le G$

$$\sum_{t=1}^T \frac{1}{2\alpha t} = \frac{\ln T}{2\alpha}$$

if $\hat{x} = \frac{1}{T}\sum_1^T x^t$

by convexity $\Rightarrow f(\hat{x}) - f(x^*) \le G^2 \frac{\ln T}{2\alpha T} \le \varepsilon$

if $T \ge \frac{G^2}{2\alpha} \cdot \left(\frac{1}{\varepsilon}\ln\frac{1}{\varepsilon}\right)$.

[Moral of the story: if assume strong convexity then faster convergence. Basically $O(1/\varepsilon)$ convergence time.]

BTW: for basic GD, ℓ_2 recall what guarantees it gives for experts: $T \ge \Omega\left(\frac{n}{\varepsilon^2}\right)$ since $D = 1$ but $G = \sqrt{n}$.

Good. Today do things with different norms.

- $\|\cdot\|$ <u>norm</u> if $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ satisfies ① $\|\bar{x}\| = 0 \Leftrightarrow x = 0$.
  ② $\|\alpha \bar{x}\| = |\alpha| \|x\|$
  ③ $\|\bar{x} + \bar{y}\| \le \|x\| + \|y\|$.

<u>Examples</u>: ℓ_p norms, polyhedral norms, etc.

- <u>Dual norm</u> $\|\cdot\|_*$ such that $\|x\|_* = \cancel{\max}\ \sup \{ |y^T x| \mid \|y\| \le 1 \}$.

hence $\langle x, y \rangle \le \|x\|_* \|y\|$ for all x, y [generalized C-S].

<u>Examples</u>: ℓ_1/ℓ_∞, ℓ_p/ℓ_q.

<u>for finite dim spaces, $(\|\cdot\|_*)_* = \|\cdot\|$.</u>

<u>Recall</u>: given K, (convex)
$K^* = \{ y \mid x^T y \le 1 \ \forall x \in K \}$.

The basic idea is that we should interpret GD in the "right" norm.

<u>Ok</u>: facts: the gradient should be considered in the "dual" space.

if f is G-Lipschitz wrt $\|\cdot\|$, then we really say that $\|\nabla f\|_* \le G$.
↑ dual norm.

So the ~~[illegible]~~ gradients lie in the "dual" space.

And hence doing $x_{t+1} \leftarrow x_t - \eta \nabla_t$ does not make any sense
(unless space is self-dual, like ℓ_2).

<u>Insight of Nemirovski & Yudin</u>:

map x to dual space (how?)
do a gradient step there (makes sense now)
map it back.

picture
(we'll populate it over the lecture)

<u>Setup:</u> Let $K \subseteq \mathbb{R}^n$ be the feasible region of the primal space

- Let $h(x): \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable, strictly convex fn on $\mathbb{R}^n$

  then $\nabla h(x): \mathbb{R}^n \to \mathbb{R}^n$ will be a map from the primal $\to$ dual space.

- Assume $\nabla h(\mathbb{R}^n) = \mathbb{R}^n$

$$\begin{tikzcd} y_t \arrow[d, "-\eta_t \nabla f(x_t)"'] & x_t \arrow[l, bend right, "\nabla h(\cdot)"'] \\ y_{t+1} \arrow[r, bend right, "(\nabla h)^{-1}(\cdot)"'] & x'_{t+1} \end{tikzcd}$$

So now: $y_{t+1} = \overset{\nabla h(x_t)}{\overset{\shortparallel}{y_t}} - \eta_t \nabla f(x_t)$

let x'_{t+1} be preimage of y_{t+1}.

x'_{t+1} may be outside K.

So "project" x'_{t+1} back onto K (using the primal metric, and the fn $h(\cdot)$)

How?

Define the Bregman divergence wrt function $g(\cdot)$ as follows:

$$D_g(y \| x) = g(y) - g(x) - \langle \nabla g(x), y - x \rangle$$ is the "slack" in the convexity.

by convexity ≥ 0. by strict convexity > 0.

Use knowledge of g: we can reserve g for gradient, which is more common!!

- <u>Examples</u>: $g_1 = \frac{1}{2}\sum x_i^2 = \frac{1}{2}\|x\|^2$. then $\nabla g(x) = x$. this is the identity map.

  $g_2 = \sum x_i (\ln x_i)$. $\Rightarrow$ $\nabla g(x) = (1 + \ln x_i)_i$

  $(\nabla g)^{-1}(y) = e^{(y_i - 1)}$

and $D_{g_1}(y \| x) = \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x\|^2 - \langle x, y - x \rangle = \frac{1}{2}\|x - y\|^2$.

$D_{g_2}(y \| x) = KL(y \| x) - \sum y_i + \sum x_i = KL(y \| x)$ for $x, y \in \Delta_n$.

why? $\sum y_i \ln y_i - \sum x_i \ln x_i - \langle \nabla g(x), y - x \rangle$

$= \sum_i (y_i - x_i)(\ln x_i)$

$= \sum_i y_i \ln x_i + \sum y_i - \sum x_i - \sum x_i \ln x_i$

$= \sum y_i \ln \frac{y_i}{x_i} - \sum y_i + \sum x_i$

So the picture is now complete

$$x_{t+1} = \arg\max_{x \in K} \{ D_h(x \| x'_{t+1}) \}.$$

<u>Thm</u>: Let h be α-strongly convex wrt $\|\cdot\|$ ← primal norm. i.e. $g(y) \geq g(x) + \langle \nabla g(x), y-x \rangle + \frac{\alpha}{2}\|y-x\|^2$ ← that norm.

Then Mirror descent has property that after T steps

$$\sum_t f_t(x_t) \leq \sum_t f_t(x^*) + \frac{h(x^* \| x_0)}{\eta} + \frac{\eta}{2\alpha} \sum \| \nabla_t \|_*^2$$

$\nabla f_t(x_t)$

[Hmm: good?

Sps $h = \frac{1}{2}\|x\|_2^2$. Clearly strongly convex wrt $\|\ \|_2$. Also $(\|\ \|_2)^* = \|\ \|_2$.

$D_h(a\|b) = \frac{1}{2}\|a-b\|_2^2$ $\Rightarrow$ RHS $\leq \frac{1}{2\eta}\|x_0 - x^*\|_2^2 + \frac{\eta}{2\alpha} G^2$ as in last lecture.

<u>Now sps</u>: $h = \sum x_i \ln x_i$ and sps $K = \Delta_n$ probability simplex.

then $D_h(y\|x) = KL(y\|x)$.

$$\Rightarrow \text{RHS} \leq \frac{1}{\eta}(KL(x^* \| x_0)) + \frac{\eta}{2\alpha} \sum \| \nabla_t \|_*^2$$

Hang on! What norm? Use ℓ_1 norm (dual is ℓ_∞).

And h is α-strongly convex wrt L_1 norm. [Pinsker's Inequality] for $\alpha =$ (Same as saying that $KL(y\|x) \geq \frac{\alpha}{2}\|y-x\|_1^2$ $\alpha = \frac{1}{\ln 2}$).

<u>Now</u>: if we do experts, then $\nabla_t = \nabla f(x_t) = \ell_t \in [-1,1]^N$ and their ℓ_∞ norm ≤ 1.

<u>Moreover</u>: distance is measured in $KL(x^* \| x_0) \leq \ln N$.

$$\Rightarrow \frac{\ln N}{\eta} + \frac{\eta}{O(1)} \cdot T$$ (as shown earlier)

What does MD do in the L_1/L_∞ case?

given $x_i^t \in \Delta_n$. find $y_i^t = \nabla h(x_i^t) = (1 + \ln x_i^t)$

Now: add in $-\eta \ell_i^t$ to each coord.

Finally: $(\nabla h)^{-1}(y_i^{t+1}) = e^{(\cancel{1} + \ln x_i^t - \eta \ell_i^t \cancel{-1})}$ for inverse map

$$= x_i^t \cdot e^{-\eta \ell_i^t} \qquad \text{[Hedge!!]}.$$

Oh, maybe outside the probability simplex Δ_n. so do Bregman Projection.

What does this do? Rescale $x_i^{t+1} = \dfrac{x_i^t \cdot e^{-\eta \ell_i^t}}{\sum_j x_j^t e^{-\eta \ell_j^t}}$! [Hedge again !!]

If you want to see a proof of this projection fact

Can ignore if you want to read rest of proof!

Lemma: $\underset{b \in \Delta_n}{\operatorname{argmin}} \; D_h(b \| a) = \dfrac{a}{\|a\|_1}$. for $h = \sum_i x_i \ln x_i$

Pf: $\underset{b}{\operatorname{argmin}} D_h(b\|a) = \underset{b \in \Delta_n}{\operatorname{argmin}} \left\{ \sum_i b_i \ln \frac{b_i}{a_i} - \sum_i b_i + \sum a_i \right\}$

use Lagrange multiplier λ:

$$F(b, \lambda) = \sum_i b_i \ln \frac{b_i}{a_i} - \sum b_i + \sum a_i + \lambda\left(\sum_i b_i - 1\right).$$

In this unconstrained problem, look at where $\partial F / \partial b_i = 0$.

$$\ln b_i/a_i + 1 - 1 + \lambda = 0 \Rightarrow b_i = a_i \cdot e^{-\lambda}.$$

$\partial F/\partial \lambda = 0$ at $\sum_i b_i = 1 \Rightarrow \sum_i a_i e^{-\lambda} = 1 \Rightarrow \lambda = \ln \sum_i a_i$

$$\Rightarrow b_i = a_i \cdot e^{-\ln \sum a_i} = \frac{a_i}{\|a\|_1}.$$

Finaly the proof of Thm: (same principle as earlier ones). Assume first that $K = \mathbb{R}^n$ so no projection required.

$$\Phi_t = \frac{1}{\eta} D_h(x^* \| x_t).$$

Amortized cost:

$$f_t(x_t) - f_t(x^*) + \underbrace{(\Phi_{t+1} - \Phi_t)}$$

$$= \frac{1}{\eta}\Big[D_h(x^* \| x_{t+1}) - D_h(x^* \| x_t)$$

$$= \frac{1}{\eta}\Big[\cancel{h(x^*)} - h(x_{t+1}) - \langle \underbrace{\nabla h(x_{t+1})}_{y_{t+1} = y_t - \eta \nabla_t}, x^* - x_{t+1} \rangle - (\cancel{h(x^*)} - h(x_t) - \langle \underbrace{\nabla h(x_t)}_{y_t}, x^* - x_t \rangle)\Big]$$

$$= \frac{1}{\eta}\Big[\underbrace{h(x_t) - h(x_{t+1}) - \langle \underbrace{y_t}_{\nabla h(x_t)}, x_t - x_{t+1} \rangle}_{\leq -\frac{\alpha}{2}\|x_t - x_{t+1}\|^2} + \eta \langle \nabla f_t(x_t), x^* - x_{t+1} \rangle \Big]$$

by strong convexity of h

$$\leq f_t(x_t) - f_t(x^*) + \langle \nabla f_t(x_t), x^* - x_{t+1} \rangle - \frac{\alpha}{2\eta}\|x_t - x_{t+1}\|^2$$

$$\leq \underbrace{\langle \nabla f_t(x_t), x_t - x_{t+1} \rangle}_{\text{by convexity of } f_t} - \frac{\alpha}{2\eta}\|x_t - x_{t+1}\|^2$$

by C-S $\leq \|\nabla_t\|_* \|x_t - x_{t+1}\| \leq \frac{1}{2}\Big[\frac{\eta}{\alpha}\|\nabla_t\|_*^2 + \frac{\alpha}{\eta}\|x_t - x_{t+1}\|^2\Big]$

cancels with

$$\leq \frac{\eta}{2\alpha}\|\nabla_t\|_*^2$$

$\Rightarrow$ Total cost: $\sum_t f_t(x_t) - \sum_t f_t(x^*) \leq \Phi_0 + \frac{\eta}{2\alpha}\sum_t \|\nabla_t\|_*^2$

$$= \frac{D_h(x^* \| x^0)}{\eta} + \frac{\eta}{2\alpha}\sum_t \|\nabla_t\|_*^2.$$

- Almost same as previous proof, just couched in a different language.
- For constrained case, need that $D_h(x^* \| x_{t+1}) \leq D_h(x^* \| x'_{t+1})$ if $x^* \in K$, x'_{t+1} not. ← "Extended Pythagoras Thm" for Bregman divergences.

More about Mirror Descent

Btw: the maps ∇h and $(\nabla h)^{-1}$ are the "mirror" maps.

the function h is called a "distance generating" function.

"Prox GD": nowadays we view MD also as follows.

$$x_{t+1} = \underset{x \in K}{\operatorname{argmin}} \left\{ \eta \langle \nabla f(x_t), x \rangle + D_h(x \| x_t) \right\}$$

if we move opposite to gradient the term is negative

but moving too far may cause this divergence term to sharply rise.

e.g. if $D_h(x \| x_t) = \frac{1}{2}\|x - x_t\|^2$

then $\underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \eta \nabla_t^T x + \frac{1}{2}\|x - x_t\|^2 \right\}$

is given by setting its gradient to 0

i.e. $\eta \nabla_t + (x - x_t) = 0$

$\Rightarrow x = x_t - \eta \nabla_t$ as desired. ∎

Exercise: prove that prox GD $\equiv$ MD as defined on previous pages.