INTRODUCTORY COURSE ON OPTIMIZATION

PRANEETH NETRAPALLI.

~~CONVEX FUNCTIONS~~:

$$\min_{x} f(x)$$

Iterative methods : Start with $x_0 \to x_1 \to \cdots$

How do we access f?

Black-box models : Zeroth order $x \to f(x)$

First order $x \to \nabla f(x)$

Second order $x \to \nabla^2 f(x)$.

First-order hits a sweet spot. If f is given by a circuit with smooth components, can compute ∇f in essentially the same amount of time.

Classes of functions : Assume for now the function 'f' is very smooth.

Prototypical first order method : GRADIENT DESCENT.

1. Start with x_0
2. Do $k = 0, 1, \ldots$

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

[LATER]

Lemma : Suppose $\|\nabla^2 f\| \leq L$ and $\eta = \frac{1}{L}$. Then

$$\min_{i=1,\cdots k} \|\nabla f(x_i)\|^2 \leq \frac{(f(x_0) - f^*)\, 2L}{k}$$

Proof: $f(x_{k+1}) = f(x_k - \eta \nabla f(x_k))$

$= f(x_k) + \langle \nabla f(x_k), -\eta \nabla f(x_k) \rangle + \eta^2 \nabla f(x_k)^T \nabla^2 f(\tilde{x}_k) \cdot \nabla f(x_k)$

[by

SMOOTHNESS: $\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|$.

Lemma: If f is L-smooth, then

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \| x - y \|^2.$$

Proof: $f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \int_{\tau=0}^{1} \langle \nabla f(\tau y + (1-\tau)x) - \nabla f(x), y - x \rangle \, d\tau$

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| = \left| \int_{\tau=0}^{1} \langle \nabla f(\tau y + (1-\tau)x) - \nabla f(x), y - x \rangle \, d\tau \right|$$

$$\overset{(C\text{-}S)}{\leq} \int_{\tau=0}^{1} \| \nabla f(\tau y + (1-\tau)x) - \nabla f(x) \| \cdot \| y - x \| \, d\tau$$

$$\overset{(\text{Smoothness})}{\leq} \left[\int_{\tau=0}^{1} \tau \, d\tau \right] \cdot L \| y - x \|^2$$

$$= \frac{L}{2} \| y - x \|^2$$ ▨

Proof of GD:

$$f(x_{k+1}) \leq f(x_k) - \eta \|\nabla f(x_k)\|^2 + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$= f(x_k) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(x_k)\|^2.$$

$$\Rightarrow \|\nabla f(x_k)\|^2 \leq 2L\left[f(x_k) - f(x_{k+1})\right]$$

Summing over k, we have

$$\frac{1}{K}\sum_{i=1}^{K}\|\nabla f(x_i)\|^2 \leq 2L\left[f(x_0) - f(x_{K+1})\right]\cdot\frac{1}{K}$$

$$\leq 2L\left[f(x_0) - f^*\right]\cdot\frac{1}{K}$$

$$\Rightarrow \min_i \|\nabla f(x_i)\|^2 \leq \frac{2L\cdot\left[f(x_0) - f^*\right]}{K}. \quad \square$$

In general, this is all we can say using first order algorithms for general smooth functions.

Assumption I: $\nabla f(x) = 0 \Rightarrow$ Global optimum.

Assumption II: $f_1, f_2 \in \mathcal{F}$ then $\alpha f_1 + \beta f_2 \in \mathcal{F}$ for $\alpha, \beta \geq 0$.

Assumption III: Linear fns. $\in \mathcal{F}$.

Lemma: $\forall f \in \mathcal{F}$, we have $\boxed{f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle}$

$\mathcal{F}$ is the set of Convex functions

Proof: Let $g(y) = f(y) - f(x) - \langle \nabla f(x), y - x\rangle \in \mathcal{F}$

$\nabla g(y) = \nabla f(y) - \nabla f(x)$. $\nabla g(x) = 0 \Rightarrow x$ is global min of $g(\cdot)$.

$\Rightarrow$ Statement of Lemma.

Examples of convex functions

① Linear regression: $f(x) = \frac{1}{2}\|Ax-b\|^2$.

$$\nabla f(x) = A^T(Ax-b)$$

$$f(y) - f(x) - \langle \nabla f(x), y-x \rangle$$

$$= \frac{1}{2}\|Ay-b\|^2 - \frac{1}{2}\|Ax-b\|^2 - \langle A^T(Ax-b), y-x \rangle$$

$$= \frac{1}{2}\|A(x-y)\|^2 \geq 0.$$

② ℓ_p norms: $f(x) = |x|^p$ Say p is even.

$$\nabla f(x) = p\,x^{p-1}$$

$$g_x(y) = f(y) - f(x) - \langle \nabla f(x), y-x \rangle = |y|^p - |x|^p - p\,x^{p-1}(y-x)$$

$$= \cancel{y^p + x^{p-1}y - p\,x^{p-1}y}$$

$$\nabla g_x(y) = p\left[y^{p-1} - x^{p-1}\right] = 0 \Rightarrow y = x.$$

③ Composition: If f is convex then $g(x) = f(Ax+b)$ is convex.

$$\nabla g = A^T \nabla f(Ax+b)$$

$$g(y) - g(x) - \langle \nabla g(x), y-x \rangle = f(Ay+b) - f(Ax+b)$$

$$- \langle A^T \nabla f(Ax+b), y-x \rangle$$

$$= f(Ay+b) - f(Ax+b) - \langle \nabla f(Ax+b), (Ay+b)-(Ax+b) \rangle$$

$$\geq 0$$

(3) Max of Convex functions is also Convex.

(4) L_p^p regression : $f(x) = \|Ax-b\|_p^p$.

(5) LASSO : $f(x) = \frac{1}{2}\|Ax-b\|^2 + \|x\|_1$.

CONVEXITY : $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$

↳ ALSO : $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$

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

Theorem : If $\|\nabla f(x_t)\| \leq G$ then

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta T} + \frac{\eta G^2}{2}.$$

Proof : $\|x_{t+1} - x^*\|^2 = \|x_t - \eta \nabla f(x_t) - x^*\|^2$

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

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

$$\Rightarrow 2\eta \langle \nabla f(x_t), x_t - x^* \rangle \leq \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 + \eta^2 G^2$$

Adding and dividing by T,

$$\frac{1}{T}\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{\|x_0 - x^*\|^2}{2\eta T} + \frac{\eta G^2}{2}.$$

Use Convexity. ∎

Choosing $\eta = \frac{\|x_0 - x^*\|}{G\sqrt{T}}$ gives us,

$$\frac{1}{T}\sum_{t=1}^{T}\left[f(x_t) - f(x^*)\right] \leq \frac{\|x_0 - x^*\| G}{\sqrt{T}}.$$

Remarks: ⓪ The resulting rate is indep. of dimension.

① Need knowledge of $\|x_0 - x^*\|$ and G to set the optimal parameters. Designing parameter free algorithms is practically important. There are results but outside the scope of current course.

② All we needed from the function was

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

As long as we have a ~~notion~~ of $\nabla f(x)$ satisfying this, we are ok. We do not even need the function to be differentiable. The notion of Subgradient suffices.

Subgradient: A vector g is said to be a Subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall\, y.$$

Need access to Subgradient oracle.

Example: $f(x) = |x|$.

For $x \neq 0$, $\nabla f(x) = \text{sign}(x)$.

$x = 0$, $\nabla f(x) = [-1, 1]$. ///

Lower bounds: Is the bound that we obtained, i.e., $f(\bar{x}_T) - f(x^*) \leq \frac{\|x_0 - x^*\| \cdot G}{\sqrt{T}}$ optimal?

How do we define optimality? For any given problem, I can hardcode its output. We need to look at function classes.

Let $\mathcal{F}_{G,R,x_0} = \{ f : f$ is convex
f is Lipschitz i.e., $|f(x) - f(y)| \leq G\|x - y\|$
x^* exists and $x^* \in B(x_0, R) \}$.

Note: f is Lipschitz $\Leftrightarrow$ All subgradients are bounded.

→ {We T instead of K}

Theorem: For any class $\mathcal{F}_{G,R,x_0}$ (i.e., for any value of G, R and x_0) and any $0 \leq k \leq n-1$ | $\exists$ a function $f \in \mathcal{F}_{G,R,x_0}$ s.t

$$f(x_k) - f^* \geq \frac{GR}{2(1+\sqrt{k+1})}$$

$\forall$ optimization algorithm generating x_k s.t.

$x_k \in x_0 + \text{Lin span} \{ \nabla f(x_0), \nabla f(x_1), \cdots \nabla f(x_{k-1}) \}$.

USE 'T' INSTEAD OF 'K'

Note: Gradient descent $x_{k+1} = x_k - \eta \nabla f(x_k)$ Satisfies the hypotheses. Indeed in the black box model, this seems to be the most an algorithm can do.

Proof: WLOG $x_0 = 0$. Fix $0 \le K \le n-1$.

Pick $$f(x) = \frac{\sqrt{K+1} \cdot G}{1+\sqrt{K+1}} \max_{1 \le i \le K} x^{(i)} + \frac{G}{2(1+\sqrt{K+1})R} \cdot \|x\|^2.$$

Denote $\gamma = \frac{\sqrt{K+1} \cdot G}{1+\sqrt{K+1}}$ and $\mu = \frac{G}{(1+\sqrt{K+1})R}$. → May be later.

i) f is Convex

iii) $$x^{*(i)} = \begin{cases} -\frac{\gamma}{\mu(K+1)} & 1 \le i \le K \\ 0 & K+1 \le i \le n. \end{cases}$$

Aside: x^* is optimizes of $f \iff 0$ is a subgradient of f at x^*

Rewrite: $$f(x) = \gamma \max_{1 \le i \le K+1} x^{(i)} + \frac{\mu}{2}\|x\|^2$$

RULES FOR COMPUTING SUBGRADIENT.

$$\partial f(x) = \frac{\gamma}{K+1} \cdot (e_1 + \cdots + e_{K+1}) + \mu x$$

$= 0$. So x^* is opt of f.

$$\|x^*\| = \frac{r}{\mu\sqrt{K+1}} \Rightarrow x^* \in B\left(x_0, \frac{r}{\mu\sqrt{K+1}}\right).$$

ii) SUBgradient at x:

$$\partial f(x) = \mu x + r \cdot \operatorname{Conv}\{e_i : i \in I(x)\},$$

where $I(x) = \left\{ j : x^{(j)} = \max_{1 \le i \le K+1} x^{(i)} \right\}$.

$$1 \le j \le K+1$$

$$\|\partial f(x)\| \le \mu \|x\| + r \cdot \max_{v \in \operatorname{Conv}\{\cdots\}} \|v\|$$

$$\le \mu \cdot R + r \cdot$$

Want: $\mu R + r \le G$, $\frac{r}{\mu\sqrt{K+1}} \le R$ → Choose $r = \frac{\sqrt{K+1} \cdot G}{1+\sqrt{K+1}}$, $\mu = \frac{G}{R(1+\sqrt{K+1})}$.

Now: $\partial f(x) = \mu x + r \cdot \operatorname{Conv}\{e_i : i \in I(x)\}$.

The subgradient oracle could return any item of the subgradient set. Let us say, we are dealing with the subgradient oracle which gives $\nabla f(x) = \mu x + r e_l$ where

$$l = \min\left\{ j : x_j = \max_{1 \le i \le K+1} x_i \right\}.$$

$x_0 = 0$

$\nabla f(0) = \nabla f(x_0) = \mu \cdot 0 + \sqrt{} e_1 = \sqrt{} e_1$

$x_1 \in 0 + \text{Span}(e_1)$.

$\nabla f(x_1) = \mu x_1 + \sqrt{} \{e_1 \text{ or } e_2\}$

So, $x_2 \in 0 + \text{Span}\{\nabla f(x_0), \nabla f(x_1)\}$

$\subseteq \text{Span}\{e_1, e_2\}$.

Similarly if $x_{k-1} \in \text{Span}\{e_1, \dots e_{k-1}\}$,

then $\nabla f(x_k) \in \text{Span}\{e_1, \dots, e_k\}$

and so $x_k \in \text{Span}\{e_1, \dots e_k\}$.

On the other hand,

$$\min_{x : x \in \text{Span}\{e_1, \dots e_k\}} f(x) \geq \sqrt{} \cdot 0 + \mu \cdot 0 = 0.$$

as long as $k \leq K$.

So, $f(x_k) \geq 0$ while $f(x^*) = \frac{-\sqrt{}^2}{\mu(K+1)} + \frac{\mu}{2} \cdot \frac{\sqrt{}^2}{\mu^2 (K+1)}$

$$= -\frac{\sqrt{}^2}{2\mu(K+1)}$$

$$= -\frac{(K+1)GR}{2(K+1)\left[1+\sqrt{K+1}\right]}$$

$$= \frac{-GR}{2\left[1+\sqrt{K+1}\right]}$$

So, $f(x_k) - f(x^*) \geq \frac{GR}{2[1+\sqrt{k+1}]}$. (1)

Comparing with the performance bound of GD, we see that this is tight.

GD: $f(\bar{x}_T) - f(x^*) \leq \frac{GR}{\sqrt{T}}$.

Lower bound: $f(x_T) - f(x^*) \geq \frac{GR}{2[1+\sqrt{T+1}]}$.
for any "span" algorithm.

Running example: $f(x) = \frac{1}{2}\|Ax-b\|^2$

Linear model for (a_i, b_i): $\langle a_i, x \rangle \sim b_i$.

Given $A = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ and $b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$

we would like $x^* = \arg\min_x f(x)$

Let us suppose $\|x^*\| \leq R$.

$\|\nabla f(x)\| = \|Ax-b\| \leq \|A\| \cdot R + \|b\| \triangleq G$.

Then GD starting at $x_0 = 0$ will give

$$f(\bar{x}_T) - f(x_0^*) \leq \frac{R[\|A\| \cdot R + \|b\|]}{\sqrt{T}}$$

Can we do better for this problem?

Perhaps. The gradients exist everywhere. In fact they are Lipschitz continuous.

$$\|\nabla f(x) - \nabla f(y)\| = \|Ax - b - Ay + b\|$$

$$\leq \|A\| \cdot \|x - y\|.$$

The lower bound example we constructed does not have this property. ~~So, may~~

Theorem: Suppose f has Lipschitz gradients (also known as smoothness): $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. Suppose x^* exists. ~~and $x^* \in B_R(x_0)$~~. Then,

$$f(\bar{x}_T) - f(x_0) \leq \frac{6L \cdot \|x_0 - x^*\|^2}{T}.$$

Proof:

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

$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta \nabla f(x_t) - x^*\|^2$$

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

On the other hand,

$$f(x^*) \leq f(x_t - \eta \nabla f(x_t)) \leq f(x_t) - \eta \cdot \|\nabla f(x_t)\|^2 + L\eta^2 \|\nabla f(x_t)\|^2$$

$$\Rightarrow \eta(1 - \eta L) \|\nabla f(x_t)\|^2 \leq f(x_t) - f(x^*).$$

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

This gives, $\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - \eta\left[2 - \frac{1}{1-\eta L}\right]\left[f(x_t) - f(x^*)\right]$

$$\Rightarrow f(x_t) - f(x^*) \leq \frac{1}{\eta\left[2 - \frac{1}{1-\eta L}\right]} \left\{ \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right\}.$$

Taking a telescopic sum and using $f(x_{t+1}) \leq f(x_t)$, we have

$$f(x_T) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{T \cdot \eta\left[2 - \frac{1}{1-\eta L}\right]}.$$

Choosing $\eta = \frac{1}{3L}$, we have

$$f(x_T) - f(x^*) \leq \frac{6\,\|x_0 - x^*\|^2 \cdot L}{T}. \quad \blacksquare$$

Compare this to the non-smooth setting where we obtained

$$f(\bar{x}_T) - f(x^*) \leq \frac{GR}{\sqrt{T}}.$$

For the linear least squares $f(x) = \frac{1}{2}\|Ax - b\|^2$,
Smoothness parameter $= \|A\|^2$.

So, $$f(x_T) - f(x^*) \leq \frac{6\,\|A\|^2\,\|x_0 - x^*\|^2}{T}.$$

Is this tight? Best possible?

Lower bound for Smooth functions

$$f(x) = \frac{1}{2} x^T A x - x^T b$$

where $A = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}$

* $\|A\| \leq 4$. [by Gershgorin disk theorem]

So, $\|\nabla f(x) - \nabla f(y)\| \leq 4 \cdot \|x - y\|$.

* Let $x^{(0)} = 0$.

$$f(x) = \frac{1}{2}\left[x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + \cdots \right] - x_1 .$$

If we have $x_{k+1} = x_{k+2} = \cdots = 0$ then the minimizer over only $x_1, \ldots x_k$ will satisfy

$$\left.\begin{aligned} 2x_1 - x_2 &= 1 \\ x_3 - 2x_2 + x_1 &= 0 \\ x_4 - 2x_3 + x_2 &= 0 \\ &\vdots \\ x_{k-1} - 2x_k &= 0 \end{aligned}\right\} \quad \therefore x_{k-i} = \frac{(i+1)}{k+1} .$$

$$f_K^* = \frac{1}{2}\left[\left(\frac{K}{K+1}\right)^2 + K\cdot\frac{1}{(K+1)^2}\right] - \frac{K}{K+1}$$

$$= \frac{1}{2}\cdot\frac{K}{K+1} - \frac{K}{K+1} = \frac{-K}{2(K+1)} = -\frac{1}{2}\left\{\frac{1}{K+1}\right\} + \frac{1}{2(K+1)}.$$

$$\|x^*\|^2 = \frac{\sum_{i=0}^{K'-1}(i+1)^2}{(K'+1)^2} = \frac{(K'-1)K'(2K'-1)}{6(K'+1)^2}. \Big] \rightarrow \text{Choose } K'=2K$$

$$\approx K'$$

$$\frac{f_K^* - f^*}{L\cdot\|x^*-x^0\|^2} \geq \frac{\frac{1}{2(K+1)}}{K'} \gtrsim \frac{1}{K^2}.$$

The lower complexity bound is $\frac{1}{K^2}$.

GD upper bound $\longrightarrow \frac{1}{K}$.

Can we actually achieve $\frac{1}{K^2}$?

YES! NESTEROV'S AGD. [Quad. U.B. minimization viewpoint of GD]

In GD, we use gradient information from only the current point. The main idea is to make use of gradients from previous time steps as well.

Estimate Sequences: Start with a quadratic fn.

$\Phi_0(x) = \Phi_0^* + \frac{L}{2}\|x - v_0\|^2$, for some v_0.

*USE T instead of K

When we query the gradient at a point y_k, we get a lower bound $f(x) \geq f(y_k) + \langle \nabla f(y_k), x - y_k \rangle$

Update $\Phi_{k+1}(x) = (1-\alpha_k)\Phi_k(x) + \alpha_k\left[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle\right]$

Choosing $\delta_0 = 1$ and $\delta_{k+1} = (1-\alpha_k)\delta_k$, we have

$$\Phi_{k+1}(x) \leq \delta_{k+1}\Phi_0(x) + (1-\delta_{k+1})f(x).$$

Let $\Phi_k(x) = \Phi_k^* + \frac{\delta_k \cdot L}{2}\|x - v_k\|^2$.

Lemma: If $\exists x_k$ s.t. $f(x_k) \leq \Phi_k^*$ then

$$f(x_k) - f^* \leq \delta_k\left[\Phi_0(x^*) - f^*\right]$$

$\min_x f(x) \triangleq f^*$

Proof:
$$\begin{aligned} f(x_k) - f^* &\leq \Phi_k^* - f^* \\ &\leq \Phi_k(x^*) - f(x^*) \\ &\leq \delta_k\left[\Phi_0(x^*) - f(x^*)\right] \quad \blacksquare \end{aligned}$$

Two requirements: (1) δ_k Should decrease rapidly.

(2) Can always find x_k s.t. $f(x_k) \leq \Phi_k^*$.

Given x_k and $\phi_k(\cdot)$, let us see how to find x_{k+1} and $\phi_{k+1}(\cdot)$ satisfying

i) $f(x_{k+1}) \leq \phi_{k+1}^*$ and

ii) α_k as large as possible.

$$\phi_k(x) = \phi_k^* + \frac{d_k \cdot L}{2} \|x - v_k\|^2.$$

Let us say we update ϕ_{k+1} as

$$\phi_{k+1}(x) = (1-\alpha_k)\left[\phi_k^* + \frac{d_k \cdot L}{2}\|x - v_k\|^2\right] + \alpha_k\left[f(y_k) + \langle \nabla f(y_k), x - y_k\rangle\right]$$

Lemma: $\phi_{k+1}^* =$ see next page.

$$v_{k+1}^* = v_k - \frac{\alpha_k}{d_k(1-\alpha_k)\cdot L}\nabla f(y_k).$$

Proof: The terms involving x can be written as

$$\frac{d_{k+1}\cdot L}{2}\left[\|x\|^2 - 2\langle x, v_k\rangle + \frac{2\alpha_k}{d_{k+1}\cdot L}\langle x, \nabla f(y_k)\rangle\right]$$

$$= \frac{d_{k+1}\cdot L}{2}\left[\|x\|^2 - 2\left\langle x, v_k - \frac{\alpha_k}{d_k(1-\alpha_k)\cdot L}\nabla f(y_k)\right\rangle\right].$$

So, $v_{k+1} = v_k - \frac{\alpha_k}{d_k(1-\alpha_k)\cdot L}\nabla f(y_k)$.

- And $\Phi_{k+1}^* = (1-\alpha_k)\Phi_k^* + \frac{\alpha_k(1-\alpha_k)\cdot L}{2}\|v_k\|^2$

$+\alpha_k f(y_k) - \alpha_k \langle \nabla f(y_k), y_k \rangle$

$- \frac{\alpha_k(1-\alpha_k)\cdot L}{2}\left\| v_k - \frac{\alpha_k}{\alpha_k(1-\alpha_k)\cdot L}\nabla f(y_k)\right\|^2$

$= (1-\alpha_k)\Phi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\alpha_k(1-\alpha_k)\cdot L}\|\nabla f(y_k)\|^2$

$+ \alpha_k \langle \nabla f(y_k), v_k - y_k \rangle$. ☐

Given x_k, y_k and v_k, one way to get x_{k+1} s.t. $f(x_{k+1}) \le \Phi_{k+1}^*$, is to try GD step at one of these points.

In fact if $x_{k+1} = y_k - \eta \nabla f(y_k)$ with $\eta \le \frac{1}{L}$, then $f(x_{k+1}) \le f(y_k) - \frac{\eta}{2}\|\nabla f(y_k)\|^2$.

So, $f(x_{k+1}) - \Phi_{k+1}^*$

$\le f(y_k) - \frac{\eta}{2}\|\nabla f(y_k)\|^2 - (1-\alpha_k) f(x_k)$

$-\alpha_k f(y_k) + \frac{\alpha_k^2}{2\alpha_k(1-\alpha_k)L}\|\nabla f(y_k)\|^2 - \alpha_k \langle \nabla f(y_k), v_k - y_k \rangle$

$\le (1-\alpha_k)\left[f(y_k) - f(x_k)\right]$

$- \alpha_k \langle \nabla f(y_k), v_k - y_k \rangle$.

$\eta \ge \frac{\alpha_k^2}{L\cdot\alpha_k(1-\alpha_k)}$

$$\leq (1-\alpha_k) \langle \nabla f(y_k), y_k - x_k \rangle - \alpha_k \langle \nabla f(y_k), v_k - y_k \rangle$$

$$= \langle \nabla f(y_k), y_k - (1-\alpha_k) x_k - \alpha_k v_k \rangle .$$

Remember, we have a choice in choosing y_k.

So, choose $y_k = (1-\alpha_k) x_k + \alpha_k v_k$.

$$\Rightarrow f(x_{k+1}) \leq \phi_{k+1}^{*} .$$

What are the conditions ?.. $\dfrac{\alpha_k^2}{d_k (1-\alpha_k)} \leq \dfrac{1}{2}$.

Eventually, we want the smallest d_k.

$\Rightarrow$ α_k should be as large as possible.

So, choose α_k as the largest root of

$$\frac{\alpha_k^2}{d_k (1-\alpha_k)} = \frac{1}{2} .$$

Let $a_k = \dfrac{1}{\sqrt{d_k}}$. Then

$$a_{k+1} - a_k = \frac{\sqrt{d_k} - \sqrt{d_{k+1}}}{\left(\sqrt{d_k} \cdot \sqrt{d_{k+1}}\right)} = \frac{d_k - d_{k+1}}{\sqrt{d_k d_{k+1}} \left(\sqrt{d_k} + \sqrt{d_{k+1}}\right)}$$

$$= \frac{\alpha_k d_k}{\sqrt{d_k d_{k+1}} \left(\sqrt{d_k} + \sqrt{d_{k+1}}\right)}$$

$$\geq \frac{\alpha_k}{2\sqrt{d_{k+1}}} = \frac{1}{2} \cdot \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sqrt{a_{k+1}}} \geq \frac{k}{2} \Rightarrow a_{k+1} \leq \frac{4}{k^2}.$$

Plugging this in the previous lemma gives us

$$f(x_k) - f^* \leq \frac{4}{k^2}\left[\phi_0(x^*) - f(x^*)\right]$$

$$= \frac{4}{k^2}\left[f(x_0) - f(x^*) + \frac{L}{2}\|x_0 - x^*\|^2\right].$$

For linear least squares problem $f(x) = \frac{1}{2}\|Ax-b\|^2$,

we get $f(x_k) - f^* \leq \frac{4}{k^2} \cdot \|A\|^2 \cdot \|x_0 - x^*\|^2$.

Compare to GD: $\frac{6\|A\|^2\|x_0 - x^*\|^2}{k}$.

$\frac{1}{k}$ vs $\frac{1}{k^2}$

By the lower bound argument, we also know this is optimal for gradient methods.

STRONG CONVEXITY

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \boxed{\frac{\mu}{2}\|x-y\|^2}$$

$\mu \rightarrow$ Strong Convexity parameter.

Smooth + Strongly Convex

By defn. : $\langle \nabla f(x), y-x \rangle \leq f(x) - f(y) - \frac{\mu}{2}\|x-y\|^2$.

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

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta \langle \nabla f(x_t), x_t - x^* \rangle + \eta^2 \|\nabla f(x_t)\|^2$$

Smoothness $\Rightarrow \eta(1-\eta L)\|\nabla f(x_t)\|^2 \leq f(x_t) - f(x^*)$.

Substituting, we have,

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - \eta\left[2 - \frac{\eta(1-\eta L)}{1-\eta L}\right]\left[f(x_t) - f(x^*)\right] - \eta\mu\|x_t - x^*\|^2.$$

For $\eta \leq \frac{1}{2L}$, we have

$$\|x_{t+1} - x^*\|^2 \leq (1-\eta\mu)\|x_t - x^*\|^2.$$

$$\leq (1-\eta\mu)^{t+1}\|x_0 - x^*\|^2.$$

So, $f(x_{t+1}) - f(x^*) \leq \frac{L}{2}\|x_{t+1} - x^*\|^2$

$$\leq \frac{L}{2}(1-\eta\mu)^{t+1}\|x_0 - x^*\|^2.$$

$\frac{L}{\mu} \triangleq \kappa \leq$ $(1-\eta\mu)^{t+1}\left[f(x_0) - f(x^*)\right]$

So, $f(x_t) - f(x^*) \leq K\left(1 - \frac{1}{K}\right)^t [f(x_0) - f(x^*)]$.

$O\left(K \log \frac{K}{\epsilon}\right)$ ← Important factor.

Can we improve? YES using AGD.

Reduction: Apply AGD

$$f(x_t) - f(x^*) \leq \frac{4L\|x_0 - x^*\|^2}{t^2}$$

$$\leq \frac{4K[f(x_0) - f(x^*)]}{t^2}$$

After $t = 4\sqrt{K}$ steps, we have

$$f(x_t) - f(x^*) \leq \frac{1}{4}[f(x_0) - f(x^*)]$$

ALSO TALK LOWER BOUND

For ϵ-accuracy, we need $O\left(\sqrt{K} \log \frac{1}{\epsilon}\right)$ iterations.

Improvement from K to $\sqrt{K}$.

Can also design a direct algorithm (without restarts).

~~NON-SMOOTH~~ + ~~STRONG~~ ~~CONV~~

For least squares $f(x) = \frac{1}{2}\|Ax - b\|^2$.

$$K = \frac{\sigma_{max}(AA^T)}{\sigma_{min}(AA^T)} = K(A)^2.$$

AGD: $f(x_t) - f(x^*) \leq \exp\left(-\frac{t}{\sqrt{K}}\right) \cdot [f(x_0) - f(x^*)]$.

NON-SMOOTH + STRONGLY CONVEX

Recall + LIPSCHITZ $[\|\nabla f\| \le G]$.

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

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta \langle \nabla f(x_t), x_t - x^* \rangle + \eta^2 \|\nabla f(x_t)\|^2$$

Recall by strong convexity

$$f(x^*) \ge f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle + \frac{\mu}{2} \|x_t - x^*\|^2.$$

$$\Rightarrow \|x_{t+1} - x^*\|^2 \le (1 - \eta\mu) \|x_t - x^*\|^2 - 2\eta [f(x_t) - f(x^*)] + \eta^2 G^2.$$

$$\Rightarrow f(x_t) - f(x^*) \le \frac{1}{2} \left[\frac{(1-\eta\mu)}{\eta} \|x_t - x^*\|^2 - \frac{1}{\eta} \|x_{t+1} - x^*\|^2 \right] + \frac{1}{2} \eta G^2$$

Choose $\eta_t = \frac{1}{\mu t}$.

$$\Rightarrow f(x_t) - f(x^*) \le \frac{1}{2} \left[\mu(t-1) \|x_t - x^*\|^2 - \mu t \|x_{t+1} - x^*\|^2 \right] + \frac{G^2}{2\mu t}.$$

$$\Rightarrow \sum_{t=1}^{T} f(x_t) - f(x^*) \le \frac{G^2}{2\mu} \sum_{t=1}^{T} \frac{1}{t} \le \frac{G^2 \log T}{2\mu}.$$

$$\Rightarrow f(\bar{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*)$$

$$\leq \frac{G^2 \log T}{2\mu T}.$$

Lower bounds : Consider a non smooth fn $f(x)$

Consider $f_\epsilon(x) = f(x) + \frac{\epsilon}{2}\|x - x_0\|^2$.

If f is G- Lipschitz then in $B_D(x_0)$,

f_ϵ is $G + \epsilon D$ - Lipschitz.

Furthermore, $f_\epsilon(\hat{x}) = f(\hat{x}) + \frac{\epsilon}{2}\|\hat{x} - x_0\|^2$

$\geq f(x^*) + \frac{\epsilon}{2}\|x^* - x_0\|^2$

$\geq f^* + \frac{\epsilon D^2}{2}$.

for $\hat{x} \triangleq \arg\min_x f_\epsilon(x)$

and $f(\hat{x}) \leq f(x^*) + \frac{\epsilon}{2}\|x^* - x_0\|^2$

$\leq f^* + \frac{\epsilon}{2} D^2$.

If for Strongly convex, we have Sub optimality $\frac{1}{T^\alpha}$ then, by choosing $\epsilon \propto T^{-\alpha/2}$, we have

Sub optimality $\frac{\epsilon D^2}{2} + \frac{G^2 \log T}{2 \cdot \epsilon \cdot T^\alpha}$

$= \left[\frac{D^2}{2} + \frac{G^2 \log T}{2}\right] \cdot T^{-\alpha/2}$.

So, $O(\frac{1}{T})$ is the optimal rate achievable for non-Smooth + Strongly convex functions.

PROJECTION & GRADIENT MAPPING

$\min_{x \in \mathcal{X}} f(x)$ $\quad$ $\mathcal{X}$ is a simple convex set.

E.g., $\mathbb{R}^n_{\geq 0}$.

Simple means that we have access to a projection oracle.

Given a point w, $P_{\mathcal{X}}(w) = \arg\min_{x \in \mathcal{X}} \|x - w\|^2$.

Examples: i) $\mathcal{X} = B_1(0)$: $P_{\mathcal{X}}(w) = w$ if $\|w\| \leq 1$

$= \frac{w}{\|w\|}$ if $\|w\| > 1$.

ii) $\mathcal{X} = \mathbb{R}^n_{\geq 0}$: $P_{\mathcal{X}}(w) = w \odot \mathbb{1}_{\{w \geq 0\}}$.

Pythagoras theorem: $\langle w - P_{\mathcal{X}}(w), y - P_{\mathcal{X}}(w) \rangle \leq 0$

and $\|w - y\|^2 \geq \|P_{\mathcal{X}}(w) - y\|^2$ $\quad \forall y \in \mathcal{X}$.

Proof: ~~$\|y - w\|^2 \geq \|P_{\mathcal{X}}(w) - w\|^2$~~

~~$\Rightarrow \|y - P_{\mathcal{X}}(w)\|^2 - 2$~~ Since $\mathcal{X}$ is a convex set,

$P_{\mathcal{X}}(w) + \alpha(y - P_{\mathcal{X}}(w)) \in \mathcal{X}$ for $\alpha \in [0, 1]$.

$$\|w - P_X(w)\|^2 \le \|w - P_X(w) - \alpha(y - P_X(w))\|^2$$

$$= \|w - P_X(w)\|^2 + \alpha^2 \|y - P_X(w)\|^2 - 2\alpha \langle w - P_X(w), y - P_X(w) \rangle$$

$$\Rightarrow \langle w - P_X(w), y - P_X(w) \rangle \le \frac{\alpha}{2} \|y - P_X(w)\| \quad \forall \alpha > 0.$$

Taking $\alpha \to 0$ finishes the proof. ∎

~~NON SMOOTH GD~~

~~$\|x_{t+1} - x^*\|^2$~~ PGD: $x_{t+1} = P_X\left(x_t - \eta \nabla f(x_t)\right)$

$$\|x_{t+1} - x^*\|^2 \le \|x_t - \eta \nabla f(x_t) - x^*\|^2$$

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

Same argument: $\frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*)$

$$\le \frac{\|x_0 - x^*\|^2}{\eta T} + \eta G^2.$$

SMOOTH GD [Gradient mapping]

$$x_{t+1} = P_{\mathcal{X}}\left(x_t - \eta \nabla f(x_t)\right)$$

$$= \arg\min_{x \in \mathcal{X}} \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|^2.$$

Let $\hat{f}_\eta(x_t; x) \triangleq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|^2$.

$\longrightarrow$ $\eta \le \frac{1}{L}$

So, we have $x_{t+1} = \arg\min_{x \in \mathcal{X}} \hat{f}_\eta(x_t; x)$.

Use $\hat{f}_\eta(x_t; x_t)$

Since $\hat{f}_\eta(x_t; \cdot)$ is $\frac{1}{\eta}$ strongly convex, we have

$$\hat{f}_\eta(x_t; x) \ge \hat{f}_\eta(x_t; x_{t+1}) + \frac{1}{2\eta} \|x - x_{t+1}\|^2$$

$\hat{f}_\eta(x_t; x) \ge f(x)$

$$= f(x_{t+1}) + \langle \nabla f(x_t), x - x_{t+1} \rangle + \frac{1}{2\eta} \|x_{t+1} - x_t\|^2 + \frac{1}{2\eta} \|x - x_{t+1}\|^2.$$

$$f(x) \ge f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$$

$$= f_\eta(x_t; x) - \frac{1}{2\eta} \|x - x_t\|^2$$

$$\ge f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2\eta} \|x_{t+1} - x_t\|^2 + \frac{1}{2\eta} \|x - x_{t+1}\|^2 - \frac{1}{2\eta} \|x - x_t\|^2.$$

$$\hat{f}_\eta(x_t; x) \geq \hat{f}_\eta(x_t; x_{t+1}) + \frac{1}{2\eta}\|x - x_{t+1}\|^2$$

$$\geq f(x_{t+1}) + \frac{1}{2\eta}\|x - x_{t+1}\|^2.$$

Also,

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

$$= \hat{f}_\eta(x_t; x) - \frac{1}{2\eta}\|x - x_t\|^2$$

$$\geq f(x_{t+1}) + \frac{1}{2\eta}\|x - x_{t+1}\|^2 - \frac{1}{2\eta}\|x - x_t\|^2.$$

Firstly, $f(x_t) \geq f(x_{t+1}) + \frac{1}{2\eta}\|x_t - x_{t+1}\|^2$

Secondly,

$$\|x^* - x_{t+1}\|^2 \leq \|x^* - x_t\|^2 - 2\eta\left[f(x_{t+1}) - f(x^*)\right].$$

Taking a telescopic sum and averaging, we have

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta T}.$$

Since $\eta = \frac{1}{L}$, we have

$$f(\bar{x}_T) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2T}.$$

□

We can similarly extend the accelerated gradient descent algorithm also to the constrained setting with simple constraints. This gives convergence rate of $O(\frac{1}{T^2})$.

NON-EUCLIDEAN SETTINGS

Can we obtain similar results when the Lipschitz constant, smoothness, strong convexity etc. are all w.r.t. to a non-Euclidean norm, say $\|\cdot\|_1$?

E.g., $$\min_{x : \|x\|_1 \le 1} \frac{1}{2} x^T A x - x^T b$$

where $\|A\|_\infty$ and $\|b\|_\infty$ are bounded.

If we use the previous results, the

Lipschitz parameter $= \max_x \|Ax - b\|$ Could be $\sqrt{d}$

Smoothness parameter $= \|A\|$ Could be d.

Can we avoid such dependence?

This is not just looseness of the results. The actual behavior of GD is this.

How can we change the algorithm depending on the structure of the problem?

$$GD: \quad x_{t+1} = \arg\min_x \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|_2^2$$

$$\downarrow$$

$$x_{t+1} = \arg\min_x \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|_1^2$$

(1) Computing such projections could be hard in general since the squared term couples diff. coordinates.

(2) What would be the potential function that we can use to track the progress of GD?

BREGMAN DIVERGENCES

Given a fn. ϕ which is 1-Strongly Convex w.r.t. $\|\cdot\|$ i.e., $\phi(y) \geq \phi(x) + \langle \nabla\phi(x), y - x \rangle + \frac{1}{2}\|x - y\|^2$,

Bregman divergence

$$D_\phi(x; y) \triangleq \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle.$$

Since ϕ is 1- strongly convex, $D_\phi(\cdot, y)$ is also 1- strongly convex.

MD: $x_{t+1} = \arg\min_{x \in \mathcal{X}} f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x; x_t)$

We will first consider the non-smooth setting: $\|\nabla f(x)\|_* \leq G$ (Dual norm.)

So, we have

$$D_\phi(x; x_{t+1}) = \phi(x) - \phi(x_{t+1}) - \langle \nabla \phi(x_{t+1}), x - x_{t+1} \rangle$$

$$= \phi(x) - \phi(x_{t+1}) - \langle \nabla \phi(x_{t+1}), x_t - x_{t+1} \rangle - \langle \nabla \phi(x_{t+1}), x - x_t \rangle$$

$$\geq \phi(x) - \phi(x_t) + \frac{1}{2}\|x_t - x_{t+1}\|^2 - \langle \nabla \phi(x_{t+1}), x - x_t \rangle.$$

By the update rule for x_{t+1}, we have

$$\langle \nabla f(x_t) + \frac{1}{\eta}\nabla \phi(x_{t+1}) - \frac{1}{\eta}\nabla \phi(x_t), x - x_{t+1} \rangle \geq 0 \quad \forall x \in \mathcal{X}$$

$$\Rightarrow \langle \nabla \phi(x_{t+1}), x - x_{t+1} \rangle \geq \langle \nabla \phi(x_t), x - x_{t+1} \rangle - \eta \langle \nabla f(x_t), x - x_{t+1} \rangle.$$

So, $D_\phi(x; x_{t+1}) = \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_{t+1}), x - x_{t+1} \rangle$

$$\leq \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_t), x - x_{t+1} \rangle + \eta \langle \nabla f(x_t), x - x_{t+1} \rangle$$

$$\leq \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x_{t+1} - x_t \rangle - \langle \nabla\phi(x_t), x - x_{t+1} \rangle + \eta \langle \nabla f(x_t), x - x_{t+1} \rangle$$

$$= \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x - x_t \rangle + \eta \langle \nabla f(x_t), x - x_{t+1} \rangle.$$

$$= D_\phi(x; x_t) + \eta \langle \nabla f(x_t), x - x_t \rangle + \eta \langle \nabla f(x_t), x_t - x_{t+1} \rangle.$$

$$\leq D_\phi(x; x_t) - \eta [f(x_t) - f(x)] + \eta \| \nabla f(x_t) \|_* \| x_t - x_{t+1} \|.$$

Since $x_{t+1} = \underset{x \in X}{\arg\min}\ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x; x_t)$

we have

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{\eta} D_\phi(x_{t+1}; x_t) \leq 0.$$

$$\Rightarrow D_\phi(x_{t+1}; x_t) \leq -\eta \langle \nabla f(x_t), x_{t+1} - x_t \rangle$$

$$\leq \eta \|\nabla f(x_t)\|_* \cdot \|x_{t+1} - x_t\|.$$

Since $D_\phi(\cdot; x_t)$ is 1-Strongly convex, we have

$$\frac{1}{2}\|x_{t+1} - x_t\|^2 \leq \eta \cdot G \cdot \|x_{t+1} - x_t\|.$$

So, $\|x_{t+1} - x_t\| \leq 2\eta G$.

Plugging this back, we obtain

$$D_\phi(x; x_{t+1}) \leq D_\phi(x; x_t) - \eta [f(x_t) - f(x)] + 2\eta^2 G^2.$$

Taking a telescopic sum, we get.

$$\frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x) \leq \frac{D_\phi(x; x_0)}{\eta T} + 2\eta G^2$$

Taking $\eta = \sqrt{\frac{D_\phi(x; x_0)}{2GT}}$ and by convexity, we have

$$f(\bar{x}_T) - f(x) \leq G \sqrt{D_\phi(x; x_0)} \cdot \sqrt{\frac{2}{T}}.$$

□

Example:- Let $f(x) = \frac{1}{2} x^T A x - x^T b$. over ~~$\|x\|_1 \le 1$~~ simplex
$x_i \ge 0$
$\sum x_i = 1$

$\|A\|_\infty \le 1$ and $\|b\|_\infty \le 1$

$\Rightarrow \|\nabla f(x)\| = \|Ax - b\|_\infty \le 2$.

Take $\phi(x) = \sum_i x_i \log x_i = -H(x)$
↓ entropy.

$D_\phi(x;y) = \sum_i x_i \log \frac{x_i}{y_i} - \sum y_i \log \frac{y_i}{y_i}$

$- \sum (\log y_i - 1)(x_i - y_i)$

$= \sum_i x_i \log \frac{x_i}{y_i} = KL(x \| y)$.

Pinsker's inequality: $KL(x\|y) \ge 2 \|x-y\|_1^2$.

Gradient step:

$x_{t+1} = \underset{x \in X}{\operatorname{argmin}} \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x; x_t)$

$= \underset{x \in X}{\operatorname{argmin}} \langle w, x \rangle + \sum_i x_i \log \frac{x_i}{x_{t_i}}$

$\triangleq \eta \nabla f(x_t)$.

$$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_i x_i \log \frac{x_i}{x_{t,i} \exp(-\omega_i)}$$

Let $Z = \dfrac{x_t \odot \exp(-\omega)}{\sum_i x_{t,i} \exp(-\omega_i)} \to \alpha$

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_i x_i \log \frac{x_i}{z_i} - x_i \log \alpha$$

$$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \; KL(x \| z) - \log \alpha \quad [\because \Sigma x_i = 1]$$

$$= z. \quad [\because z \in \mathcal{X}].$$

So, the gradient update is

$$x_{t+1} = \frac{x_t \odot \exp(-\eta \nabla f(x_t)_i)}{\sum_i x_{t,i} \exp(-\eta \nabla f(x_t)_i)}$$

Our result says,

$$f(\bar{x}_T) - f(x^*) \leq \frac{\sqrt{8 D_\phi(x^*; x_0)}}{\sqrt{T}}$$

If we choose $x_0 = \begin{bmatrix} 1/d \\ \vdots \\ 1/d \end{bmatrix}$ then $D_\phi(x^*; x_0)$

$$= KL\left(x^*, \begin{bmatrix} 1/d \\ \vdots \\ 1/d \end{bmatrix}\right).$$

$$= \log d - H(x^*) \leq \log d.$$

So, $f(\bar{x}_T) - f(x^*) \leq \sqrt{\frac{8 \log d}{T}}$.

Compare with ~~gr~~ Euclidean gradient descent.

$$f(\bar{x}_T) - f(x^*) \leq \frac{\left[\|A\|_{1\to 2} + \|b\|\right]}{\sqrt{T}}.$$

If $\|A\|_\infty \leq 1$, it is still possible that

$\|A\|_{1\to 2}$ could still be $\sqrt{d}$

and $\|b\|_\infty \leq 1$, it is still possible that $\|b\| \approx \sqrt{d}$.

MIRROR DESCENT + SMOOTHNESS.

Smoothness: $\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|$.

($\|\cdot\|_*$: Dual norm; $\|x-y\|$: Original norm.)

~~$x_{t+1} = \underset{x \in X}{\operatorname{argmin}} \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x; x_t)$.~~

~~Firstly $f(x_{t+1}) \leq f(x_t) +$~~

Smoothness again gives us a quadratic upper bound.

$$f(y) = f(x) + \int_{\alpha=0}^{1} \langle \nabla f(x+\alpha(y-x)), y-x \rangle \, d\alpha$$

$$= f(x) + \langle \nabla f(x), y-x \rangle + \int_{\alpha=0}^{1} \langle \nabla f(x+\alpha(y-x)) - \nabla f(x), y-x \rangle \, d\alpha$$

$$\leq f(x) + \langle \nabla f(x), y-x \rangle + \int_{\alpha=0}^{1} \| \nabla f(x+\alpha(y-x)) - \nabla f(x) \|_* \, \|y-x\| \, d\alpha$$

[SMOOTHNESS]

$$\leq f(x) + \langle \nabla f(x), y-x \rangle + \int_{\alpha=0}^{1} L \cdot \alpha \, \|y-x\| \cdot \|y-x\| \, d\alpha$$

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

Since $x_{t+1} = \arg\min_{x \in \mathcal{X}} f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x; x_t)$

we have

$f(x_t) - f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{\eta} D_\phi(x_{t+1}; x_t)$

[s.c. of D_ϕ] $\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2\eta} \|x_t - x_{t+1}\|^2$

$$\geq f(x_{t+1}) + \left(\frac{1}{2\eta} - L\right) \|x_t - x_{t+1}\|^2.$$

So, mirror descent is a descent algorithm.

Furthermore, from previous computations, we have

$$D_\phi(x; x_{t+1}) \le D_\phi(x; x_t) + \eta \langle \nabla f(x_t), x - x_{t+1} \rangle \pm \eta \langle \nabla f(x_t), x_t - x_{t+1} \rangle$$

By optimality def. of x_{t+1}, and S.C. of D_ϕ and smoothness, we have

$$\le D_\phi(x; x_t) + \eta \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \eta \langle \nabla f(x_t) - \nabla f(x_{t+1}), x - x_{t+1} \rangle$$

$$= D_\phi(x; x_t) + \eta \langle \nabla f(x_t), x - x_t \rangle + \eta \langle \nabla f(x_t), x_t - x_{t+1} \rangle$$

$$= D_\phi(x; x_t) + \eta \langle \nabla f(x_t), x - x_t \rangle + \eta \langle \nabla f(x_t) - \nabla f(x_{t+1}), x_t - x_{t+1} \rangle + \eta \langle \nabla f(x_{t+1}), x_t - x_{t+1} \rangle$$

$\leq D_\phi(x; x_t) - \eta\left[f(x_t) - f(x)\right]$

$- \eta\left[f(x_t) - f(x_{t+1})\right]$

$= D_\phi(x; x_t) + \eta \langle \nabla f(x_t), x - x_t \rangle$

$+ \eta \langle \nabla f(x_t), x_t - x_{t+1} \rangle$

$\leq D_\phi(x; x_t) - \eta\left[f(x_t) - f(x)\right]$

$+ \eta \langle \nabla f(x_{t+1}), x_t - x_{t+1} \rangle$

$+ \eta \langle \nabla f(x_t) - \nabla f(x_{t+1}), x_t - x_{t+1} \rangle$

$\leq D_\phi(x; x_t) - \eta\left[f(x_t) - f(x)\right]$

$- \eta\left[f(x_t) - f(x_{t+1})\right]$

By ~~def. of x_{t+1} and~~ Strong Convexity of ϕ, we have

$\phi(x_{t+1}) \geq \phi(x_t) + \langle \nabla \phi(x_t), x_{t+1} - x_t \rangle$

$+ \frac{1}{2} \|x_t - x_{t+1}\|^2.$

So, from the previous computations, we have

$$D_\phi(x; x_{t+1}) = \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_{t+1}), x - x_{t+1} \rangle$$

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

$$\leq \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x - x_t \rangle + \eta \langle \nabla f(x_t), x - x_{t+1} \rangle - \frac{1}{2}\|x_t - x_{t+1}\|^2.$$

$$= D_\phi(x; x_t) + \eta \langle \nabla f(x_t), x - x_{t+1} \rangle - \frac{1}{2}\|x_t - x_{t+1}\|^2.$$

$$= D_\phi(x; x_t) + \eta \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle - \frac{1}{2}\|x_t - x_{t+1}\|^2 + \eta \langle \nabla f(x_t) - \nabla f(x_{t+1}), x_t - x_{t+1} \rangle.$$

Now, $\langle \nabla f(x_t), x_t - x_{t+1} \rangle$

$= \underbrace{\langle \nabla f(x_{t+1}), x_t - x_{t+1} \rangle}_{\leq f(x_{t+1}) - f(x_t) \leq 0} + \underbrace{\langle \nabla f(x_t) - \nabla f(x_{t+1}), x_t - x_{t+1} \rangle}_{\leq L\|x_t - x_{t+1}\|^2}$

So, $D_\phi(x; x_{t+1}) \leq D_\phi(x; x_t) - \eta[f(x_t) - f(x)] - \left(\frac{1}{2} - \eta L\right) \|x_t - x_{t+1}\|^2$.

Since $\eta < \frac{1}{2L}$, we have

$D_\phi(x; x_{t+1}) \leq D_\phi(x; x_t) - \eta[f(x_t) - f(x)]$.

Telescoping and averaging gives us

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x) \leq \frac{D_\phi(x; x_0)}{\eta T}.$$ ∎

Infact we can extend all the Euclidean algorithms to the non Euclidean setting

→ Gradient descent
→ Accelerated gradient descent
→ Faster rates with Strong Convexity, Smoothness etc.

FASTER ALGORITHMS FOR STRUCTURED NON-SMOOTH PROBLEMS

~~$\min_x \frac{1}{2} x^T A x + x^T b + \|x\|_1$~~

LASSO : $\min_x \frac{1}{2}\|Ax-b\|^2 + \underbrace{d}_{=1}\|x\|_1$.

$g(x)$ $\quad$ $h(x)$

The function $f(x) = \frac{1}{2}\|Ax-b\|^2 + \|x\|_1$ is not smooth. So, the best rate we get using GD is $O(\frac{1}{\sqrt{T}})$. Can we do better?

Main idea: The nonsmooth part $\|x\|_1$ is very simple.

Algorithm : $x_{t+1} = \arg\min_x g(x_t) + \langle \nabla g(x_t), x - x_t \rangle + h(x) + \frac{1}{2\eta}\|x - x_t\|^2$.

Notice x not x_t

Implementability: $x_{t+1} = \arg\min_x \langle \nabla g(x_t), x \rangle + \|x\|_1 + \frac{1}{2\eta}\|x - x_t\|^2$.

Decouples into 'd' independent problems, one for each direction.

$$i: \quad \arg\min_{x_i} \nabla g(x_t)_i \cdot x_i + |x_i| + \frac{1}{2\eta}(x_i - x_{t_i})^2$$

$$= \arg\min_{x_i} \frac{1}{2\eta}(x_i - x_{t_i} + \eta \nabla g(x_t)_i)^2 + |x_i| .$$

$$\Rightarrow \quad x_{t+1_i}^{*} = \frac{1}{1+\eta}\left[x_{t_i} - \eta \nabla g(x_t)_i\right]$$

$$\Rightarrow \quad x_{t+1} = \frac{1}{1+\eta}\left(x_t - \eta \nabla g(x_t)\right) .$$

Convergence rate:

Let $\hat{f}_\eta(x; x_t) = g(x_t) + \langle \nabla g(x_t), x - x_t \rangle + h(x) + \frac{1}{2\eta}\|x - x_t\|^2 .$

$$f(x) = g(x) + h(x)$$

$$\geq g(x_t) + \langle \nabla g(x_t), x - x_t \rangle + h(x)$$

$$= \hat{f}_\eta(x; x_t) - \frac{1}{2\eta}\|x - x_t\|^2$$

$$\geq \hat{f}_\eta(x_{t+1}; x_t) + \frac{1}{2\eta}\|x - x_{t+1}\|^2 - \frac{1}{2\eta}\|x - x_t\|^2$$

~~$\geq f(x_{t+1}) + \frac{1}{2\eta}\|x_t - x_{t+1}\|^2 - \frac{1}{\eta}\langle x - x_t, x_t - x_{t+1}\rangle$~~

$$\geq f(x_{t+1}) + \frac{1}{2\eta}\|x - x_{t+1}\|^2 - \frac{1}{2\eta}\|x - x_t\|^2 .$$

Reorganizing, we have

$$\|x - x_{t+1}\|^2 \leq \|x - x_t\|^2 - 2\eta \left[f(x_{t+1}) - f(x) \right].$$

So, $$\frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x) \leq \frac{\|x - x_0\|^2}{2\eta T}.$$

$$\Rightarrow f(\bar{x}_T) - f(x) \leq \frac{2L\|x - x_0\|^2}{T}. \quad \blacksquare$$

[Since $\eta = \frac{1}{2L}$]

Can again extend Nesterov's AGD to get a rate of $O(\frac{1}{T^2})$.

STOCHASTIC ALGORITHMS

Oracle models: So far we assumed we had access to EXACT gradients of the function. $x \longrightarrow \nabla f(x)$.

In many cases this may not be reasonable

For e.g., in machine learning

$$\min_x \underbrace{\frac{1}{2} \mathbb{E}_{(a,b)} \left[(a^T x - b)^2 \right]}_{= f(x)}.$$

$$\nabla f(x) = \mathbb{E}\left[(a^T x - b) a \right]$$

↓

Cannot compute expectations.

But if we sample (a_i, b_i) and compute $\hat{\nabla} f(x) = (a_i^T x - b_i) a_i$ then $\mathbb{E}[\hat{\nabla} f(x)] = \mathbb{E}[(a_i^T x - b_i) a_i]$

$= \nabla f(x)$.

So, we may assume we have access to $\hat{\nabla} f(x)$ [Stochastic] s.t. $\mathbb{E}[\hat{\nabla} f(x)] = \nabla f(x)$.

The simplest setting, from an analytical point of view is if we assume independent and bounded variance noise.

$$\mathbb{E}\left[\| \hat{\nabla} f(x) - \nabla f(x) \|^2 \right] \leq \sigma^2 \quad \forall x.$$

Stochastic gradient descent [Robbins & Monro 1952]

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

$$\mathbb{E}\left[\| x_{t+1} - x \|^2 \right] = \mathbb{E}\left[\| x_t - x \|^2 \right] - 2\eta_t \mathbb{E}\left[\langle \hat{\nabla} f(x_t), x_t - x \rangle \right] + \eta_t^2 \mathbb{E}\left[\| \hat{\nabla} f(x_t) \|^2 \right]$$

$$\leq \mathbb{E}\left[\| x_t - x^* \|^2 \right] - 2\eta_t \mathbb{E}\left[\langle \nabla f(x_t), x_t - x \rangle \right] + \eta_t^2 \left[G^2 + \sigma^2 \right].$$

$$\leq \mathbb{E}\left[\|x_t - x\|^2\right] - 2\eta_t \mathbb{E}\left[f(x_t) - f(x)\right] + \eta_t^2 \left[G^2 + \sigma^2\right].$$

Rearranging and telescoping,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[f(x_t)\right] - f(x) \leq \frac{\mathbb{E}\left[\|x_0 - x\|^2\right]}{\eta T} + \eta\left[G^2 + \sigma^2\right]$$

If $\eta = \frac{\|x_0 - x\|}{\sqrt{T}\cdot\sqrt{G^2+\sigma^2}}$, then

$$\mathbb{E}\left[f(\bar{x}_T)\right] - f(x) \leq \frac{2\cdot\sqrt{G^2+\sigma^2}\cdot\|x_0 - x\|}{\sqrt{T}}.$$

STRONG CONVEXITY

$$\langle \nabla f(x_t), x_t - x\rangle \geq \left[f(x_t) - f(x)\right] + \frac{M}{2}\|x - x_t\|^2.$$

$$\Rightarrow \mathbb{E}\left[\|x_{t+1} - x\|^2\right] \leq (1 - \eta_t M)\,\mathbb{E}\left[\|x_t - x\|^2\right] - 2\eta_t\left[\mathbb{E}(x_t) - f(x)\right] + \eta_t^2\left[G^2 + \sigma^2\right]$$

Choosing $\eta_t = \frac{1}{\mu t}$.

$$2\,\mathbb{E}[f(x_t) - f(x)] \le \mu(t-1)\,\mathbb{E}[\|x_t - x\|^2] - \mu t\,\mathbb{E}[\|x_{t+1} - x\|^2] + \eta_t [G^2 + \sigma^2]$$

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[f(x_t)] - f(x) \le \frac{(G^2+\sigma^2)\log T}{\mu T}.$$

With smoothness: (No S.C.)

We already have

$$\mathbb{E}[\|x_{t+1} - x\|^2] \le \mathbb{E}[\|x_t - x\|^2] - 2\eta\,\mathbb{E}[f(x_t) - f(x)] + \eta^2\left[\mathbb{E}[\|\nabla f(x_t)\|^2] + \sigma^2\right]$$

By smoothness, $\|\nabla f(x_t)\|^2 \le L \cdot [f(x_t) - f(x)]$.

So, $\mathbb{E}[\|x_{t+1} - x\|^2] \le \mathbb{E}[\|x_t - x\|^2] - \eta(2 - \eta L)\,\mathbb{E}[f(x_t) - f(x)] + \eta^2\sigma^2$.

Telescoping and averaging,

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} f(x_t)\right] - f(x) \le \frac{\|x_0 - x\|^2}{\eta(2-\eta L)T} + \frac{\eta\sigma^2}{2-\eta L}.$$

$$\eta \le \min\left\{\frac{1}{L}, \frac{1}{\sigma\|x_0 - x^*\|\sqrt{T}}\right\} \Rightarrow \frac{2\sigma\|x_0 - x\|}{(2-\eta L)\sqrt{T}}.$$

STOCHASTIC MIRROR DESCENT [TALK ABOUT IT]

Is additive independent noise ~~a good~~ with bounded variance a good assumption?

* LOT of work in the Stochastic Approximation Literature for more general noise i.e., Markov structure and noise not growing fast.

E.g., $\mathbb{E}\left[\| \nabla f(x) - \hat{\nabla} f(x) \|^2 \right] \leq L \|x\|^2$.

* However the bounds are usually not ~~very~~ fully non-asymptotic [dependence on all the parameters not clear].

Primary reason for the resurgence of Stochastic algorithms in large scale machine learning

FAST COMPUTATION.

Empirical risk minimization

$$f(x) = \frac{1}{2n} \sum_{i=1}^{n} (a_i^T x - b_i)^2. \qquad a_i, x \in \mathbb{R}^d.$$

* We would like to $\min_x f(x)$.

$$\nabla f(x) = \frac{1}{n} \sum_{i=1}^{n} (a_i^T x - b_i) a_i .$$

One gradient Computation : $O(nd)$ time.

Total number of GD steps : $O(\kappa \log(\frac{1}{\epsilon}))$

↓

Condition number $= \frac{\sigma_{max}(A^TA)}{\sigma_{min}(A^TA)}$.

Total time : $O(nd\kappa \log \frac{1}{\epsilon})$.

Can we do better?

SGD : $\hat{\nabla} f(x) = (a_i^T x - b_i) a_i$

Variance : $\mathop{\mathbb{E}}_i \left[\| \hat{\nabla} f(x) - \nabla f(x) \|^2 \right] = \mathop{\mathbb{E}}_i \Big[\| (a_i^T x - b_i) a_i - \frac{1}{n} \sum_{i=1}^{n} (a_i^T x - b_i) a_i \|^2 \Big]$

$$\sigma^2 \approx \Big\{ \max_i \|a_i\|^2 \Big\} \|x - x^*\|^2 \Big]$$

$$+ f(x^*) \cdot \max_i \|a_i\|^2.$$

SGD : One gradient computation : $O(d)$ time

SGD steps : $\frac{\sigma^2}{\epsilon^2}$

$nd\kappa \log \frac{1}{\epsilon}$ vs $\frac{d\sigma^2}{\epsilon^2}$.

$\frac{nd}{\sqrt{\epsilon}}$

↳ n does not appear.

They have tradeoffs between various parameters.

SVRG: Use SGD but with decreasing variance. → How?

Let $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$.

Let each $f_i(x)$ be L-smooth: $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|$.

Suppose further that $f(\cdot)$ is μ-strongly convex.

Compute $\nabla f(x_0) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_0)$.

Instead of $x_{t+1} = x_t - \eta \nabla f_i(x_t)$, do

$$x_{t+1} = x_t - \eta \underbrace{\left[\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0) \right]}$$

$$\mathbb{E}_i[\cdot] = \nabla f(x_t).$$

$$\text{Var} = \mathbb{E}_i \left[\| \nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0) - \nabla f(x_t) \|^2 \right]$$

$$\# = \mathbb{E}_i \left[\| \nabla f_i(x_t) - \nabla f_i(x_0) \|^2 \right] + \| \nabla f(x_0) - \nabla f(x_t) \|^2$$

~~$\leq 2 \cdot L \cdot \|x_0 - x_t\|^2$~~

$$\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] = \mathbb{E}\left[\|x_t - \eta\{-\nabla f_i(x_0) + \nabla f_i(x_t) + \nabla f(x_0)\} - x^*\|^2\right]$$

$$= \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta\, \mathbb{E}\left[\langle \nabla f_i(x_t), x_t - x^* \rangle\right]$$

$$+ \eta^2\, \mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0)\|^2\right]$$

$$\leq \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta\left\{\mathbb{E}[f(x_t)] - f(x^*)\right\}$$

$$+ \eta^2 L\left\{\mathbb{E}[f(x_0)] - f(x^*)\right\}$$

~~$+2\eta^2 L \|x_0 - x_t\|^2$~~ . (?)

$$\mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x_0)\|^2\right]$$

$$\leq 2\,\mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_i(x_0) - \nabla f_i(x^*)\|^2\right]$$

$$\mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2\right] \leq ?$$

Consider $g_i(x) = f_i(x) - \langle \nabla f_i(x^*), x \rangle$

$$\Rightarrow g_i(x^*) \leq g_i\left[x_t - \eta \nabla f_i(x_t)\right] + \eta \nabla f_i(x^*)\rangle$$

~~$\leq f_i(x_t + \eta \nabla f_i(x_t)) - \langle \nabla f_i(x^*), x_t - \eta \nabla f_i(x_t) \rangle$~~

$$\leq g_i(x_t) - \eta(1 - \eta L)\, \|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2$$

$$g_i(x^*) \le f_i(x_t) - \eta \|\nabla f_i(x_t)\|^2 + \eta^2 L \|\nabla f_i(x_t)\|^2 - \langle \nabla f_i(x^*), x_t - \eta \nabla f_i(x_t) \rangle.$$

$$\frac{1}{n}\sum_i g_i(x^*) = f(x^*) \le f(x_t) - \frac{\eta(1-\eta L)}{n} \sum_i \|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2$$

$$\Rightarrow \sum_i \|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2 \le \frac{1}{\eta(1-\eta L)} [f(x_t) - f(x^*)].$$

$\to L$

So, $\mathbb{E}[\|x_{t+1} - x^*\|^2] \le \mathbb{E}[\|x_t - x^*\|^2]$

$$- \eta(2 - 8L\eta)[f(x_t) - f(x^*)]$$
$$+ \eta^2 L [f(x_0) - f(x^*)].$$

Telescoping, we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[f(x_t)] - f(x^*) \le \frac{\|x_0 - x^*\|^2}{\eta T} + \eta L [f(x_0) - f(x^*)]$$

$$\le \left[\frac{1}{\mu \eta T} + \eta L\right][f(x_0) - f(x^*)].$$

Choose $\eta = \frac{1}{10L}$ and $T > 20K$.

$$\Rightarrow \mathbb{E}[f(\bar{x}_T)] - f(x^*) \le \frac{2}{3}\{f(x_0) - f(x^*)\}.$$

Get run time complexity of $O\left[(nd + dK)\log\frac{1}{\epsilon}\right]$

GD : $O\left(nd \log\frac{1}{\epsilon}\right)$

SGD : $O\left(\frac{d\sigma^2}{\epsilon^2}\right)$.

IMPORTANCE SAMPLING :

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n}\left(a_i^T x - b_i\right)^2.$$

Smoothness parameter $L = \max_i \|a_i\|^2$.

Instead let us sample each i w.p. $\propto \|a_i\|^2$.

then $f(x) = \sum_{i=1}^{n} p_i \left\{\frac{\sum_j \|a_j\|^2}{2n\|a_i\|^2}(a_i^T x - b_i)^2\right\}$.

$$p_i = \frac{\|a_i\|^2}{\sum_j \|a_j\|^2}.$$

Here the Smoothness parameter $= \frac{\sum_j \|a_j\|^2}{n}$.

Remaining calculations in SVRG go through.

Total time $O\left[\left(nd + d\cdot\left(\frac{\sum_j \|a_j\|^2}{n\mu}\right)\right)\log\frac{1}{\epsilon}\right]$.

NONCONVEX OPTIMIZATION

NP-hard in general. Can only find ~~locally~~ ~~opt~~ STATIONARY PTS.

$f(x)$

FOSP: $\nabla f = 0$

SOSP: $\nabla f = 0$ and $\nabla^2 f \succcurlyeq 0$.

Approximate versions: $\|\nabla f\| \leq \epsilon_g$

and $\nabla^2 f \succcurlyeq -\epsilon_H \, Id$.

Let us first focus on approximate FOSP.

Lemma: If $f(\cdot)$ is L-Smooth, then GD finds an ϵ-FOSP in $O\left(\frac{L\{f(x_0)-f^*\}}{\epsilon^2}\right)$ iterations.

Proof: Already done in the first lecture.

Finite sum setting: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

Can we get SVRG-style complexity?

Algorithm: Start with $\tilde{x}_0$.

Compute $\tilde{x}_j =$ and $\nabla f(\tilde{x}_j)$

Inner loop { For $k = 1, \dots m$

$$x_{k+1} = x_k - \eta \nabla f_i(x_k) + \eta \nabla f_i(\tilde{x}_j) - \eta \nabla f(\tilde{x}_j).$$

We have $\mathbb{E}[f(x_{k+1})] = \mathbb{E}\Big[f(x_k - \eta \nabla f_i(x_k) + \eta \nabla f_i(\tilde{x}_j) - \eta \nabla f(\tilde{x})) \Big]$

$$\leq f(x_k) - \eta \| \nabla f(x_k) \|^2 + \eta^2 L \, \mathbb{E}\Big[\| \nabla f_i(x_k) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x}) \|^2 \Big]$$

What is $\mathbb{E}\Big[\| \nabla f_i(x_k) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x}) - \nabla f(x_k) \|^2 \Big]$?

$$\leq 4L^2 \, \mathbb{E}\Big[\| x_k - \tilde{x} \|^2 \Big].$$

$$x_k - x_{k+1} = \eta \nabla f_i(x_k) - \eta \nabla f_i(\tilde{x}) + \eta \nabla f(\tilde{x}).$$

$$\mathbb{E}\left[\|x_k - x_{k+1} - \eta \nabla f(x_k)\|^2\right] \leq 4\eta^2 L^2 \, \mathbb{E}\left[\|x_k - \tilde{x}\|^2\right].$$

and

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] + 4\eta^2 L^3 \, \mathbb{E}\left[\|x_k - \tilde{x}\|^2\right]$$

$$\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \geq \mathbb{E}\left[\|\nabla f(x_k)\|\right]^2$$

$$\geq \frac{1}{\eta^2}\mathbb{E}\left[\left\{\|x_k - x_{k+1}\| - \|x_k - x_{k+1} - \eta \nabla f(x_k)\|\right\}^2\right]$$

Let $v_k = \nabla f_i(x_k) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x})$.

$$x_{k+1} = x_k - \eta v_k.$$

$$\mathbb{E}[f(x_{k+1})] = \mathbb{E}[f(x_k - \eta v_k)]$$

$$\leq \mathbb{E}\left[f(x_k) - \eta \langle \nabla f(x_k), v_k \rangle + \frac{\eta^2 L}{2}\|v_k\|^2\right]$$

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

$$+ \frac{\eta^2 L}{2} \cdot 4L^2 \|x_k - \tilde{x}\|^2.$$

$$\leq f(x_k) - \frac{\eta}{2}\|\nabla f(x_k)\|^2 + 2\eta^2 L^3 \|x_k - \tilde{x}\|^2.$$

$$\mathbb{E}\left[\|x_{k+1}-\tilde{x}\|^2\right] = \mathbb{E}\left[\|x_k-\eta v_k-\tilde{x}\|^2\right]$$

$$= \mathbb{E}\left[\|x_k-\tilde{x}\|^2\right] - 2\eta\langle \nabla f(x_k), x_k-\tilde{x}\rangle + \eta^2\mathbb{E}\left[\|v_k\|^2\right]$$

$$\le \mathbb{E}\left[\|x_k-\tilde{x}\|^2\right] + \eta\left[\frac{1}{\beta}\|\nabla f(x_k)\|^2 + \beta\|x_k-\tilde{x}\|^2\right] + \eta^2\|\nabla f(x_k)\|^2 + 2\eta^2L^2\|x_k-\tilde{x}\|^2.$$

$$= \left(1+\eta\beta+2\eta^2L^2\right)\|x_k-\tilde{x}\|^2 + \left(\eta^2+\frac{\eta}{\beta}\right)\|\nabla f(x_k)\|^2$$

$$\mathbb{E}\left[f(x_{k+1})\right] + c_{k+1}\,\mathbb{E}\left[\|x_{k+1}-\tilde{x}\|^2\right]$$

$$\le f(x_k) - \left[\frac{\eta}{2} - \frac{\eta}{\beta}c_{k+1} - \eta^2c_{k+1}\right]\|\nabla f(x_k)\|^2 + \left\{2\eta^2L^3 + c_{k+1}\left(1+\eta\beta+2\eta^2L^2\right)\right\}\|x_k-\tilde{x}\|^2.$$

Pick $\alpha_{k+1} = \frac{1}{4c_{k+1}}$. Then,

$$\mathbb{E}\left[f(x_{k+1})\right] + c_{k+1}\,\mathbb{E}\left[\|x_{k+1}-\tilde{x}\|^2\right]$$

$$\le f(x_k) - \frac{\eta}{4}\left[1-4\eta c_{k+1}\right]\|\nabla f(x_k)\|^2 + \left\{2\eta^2L^3 + c_{k+1}\left(1+4\eta c_{k+1}+2\eta^2L^2\right)\right\}\|x_k-\tilde{x}\|^2.$$

If $\eta < \frac{1}{2L}$, then

$$\mathbb{E}\left[f(x_{K+1})\right] + C_{K+1}\,\mathbb{E}\left[\|x_{K+1} - \tilde{x}\|^2\right]$$

$$\leq f(x_K) + C_K \|x_K - \tilde{x}\|^2 - \left[\frac{\eta}{2} - \frac{\eta}{\beta} C_{K+1} - \eta^2 C_{K+1}\right] \|\nabla f(x_K)\|^2.$$

where $C_{K} = C_{K+1}(1 + \eta\beta + 2\eta^2 L^2) + 2\eta^2 L^3$.

If we let $C_{m+1} = 0$ then

$$C_0 = 2\eta^2 L^3 \cdot \frac{(1+\eta\beta+2\eta^2L^2)^{m+1} - 1}{\eta\beta + 2\eta^2 L^2}$$

$\eta = \frac{\mu_0}{L n^\alpha}$

$\beta = \frac{L}{n^{\alpha/2}}$

$m = \frac{n^{3\alpha/2}}{3\mu_0}$

$$\leq \frac{2\mu_0^2 L}{n^{2\alpha}} \cdot \frac{10}{\frac{\mu_0}{n^{3\alpha/2}}} = \frac{20\mu_0 L}{n^{\alpha/2}}.$$

On the other hand,

$$\min_K \frac{\eta}{2} - \frac{\eta}{\beta} C_{K+1} - \eta^2 C_{K+1} = \frac{\eta}{2} - \frac{\eta}{\beta} C_0 - \eta^2 C_0$$

$$= \eta\left[\frac{1}{2} - 20\mu_0 - \frac{20\mu_0^2}{n^{3\alpha/2}}\right]$$

$$\geq \frac{\eta}{4}.$$

Taking a telescopic sum, we have

$$\frac{1}{m}\sum_{k=0}^{m} \|\nabla f(x_k)\|^2 \leq \frac{4\left[f(\tilde{x}) - f(x_m)\right]}{m\eta}$$

$$= \frac{4\left[f(\tilde{x}) - f^*\right]}{m^{\alpha/2}}$$

Taking further sum over outer loops, we have

$$\frac{1}{mT}\sum_{\substack{k=0 \\ t=1,\ldots,T}}^{m} \|\nabla f(x_k^t)\|^2 \leq \frac{4\left[f(x_0) - f^*\right]}{(mT)\cdot\eta}$$

With n calls, we get for $\alpha = 2/3$ and $T=1$,

that $\epsilon = \frac{1}{n^{1/3}}$

If we continue further, for larger T, we get

nT calls $\longrightarrow$ $\epsilon = \frac{1}{Tn^{1/3}}$

$\longrightarrow O\left(n + \frac{n^{2/3}}{\epsilon}\right)$.

Focus on finding SOSPs

Importance: Many problems such as matrix completion, robust PCA, ... all SOSPs are good.

Give example of PCA.

Want: $\|\nabla f(x)\| \leq \epsilon_G$
and $\nabla^2 f(x) \geq -\epsilon_H \cdot Id$.

Smoothness = Gradient Lipschitz.

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|.$$

Hessian Lipschitz: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \rho \|x - y\|$.

Hessian Lipschitz is necessary for SOSPs.

Gradient descent can get stuck at FOSPs that are not SOSPs.

However this can happen only for a measure zero set of bad initialization.

However there are still situations where escaping saddle points can take exponential time.

Perturbation to the rescue.

If $\|\nabla f(x_t)\| \le \epsilon$ and perturbation not added for some time, then

$$x_t \leftarrow x_t - \eta \xi_t$$

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

Descent lemma: $f(x_{t+1}) \le f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2$.

Improve or localize: $\|x_t - x_0\| \le \sqrt{2\eta t [f(x_0) - f(x_t)]}$

Lemma: If $\tilde{x}$ is a saddle pt. i.e., $\|\nabla f(\tilde{x})\| \le \epsilon$ and $\lambda_{\min}(\nabla^2 f(\tilde{x})) \le -\sqrt{\epsilon}$.

Then letting $x_0 = \tilde{x} + \eta \xi$ $\left(\xi \sim \mathrm{Unif}(B_0(r))\right)$ and running GD for $\mathcal{T}$ iterations then w.p. $\ge 1 - \delta$,

$$f(x_{\mathcal{T}}) - f(\tilde{x}) \le -\mathcal{F} \quad \text{where}$$

$$\eta = \frac{1}{\ell}, \quad r = \frac{\epsilon}{400}, \quad \alpha = c \log\left(\frac{d\ell\Delta_f}{f\epsilon\delta}\right)$$

$$\mathcal{T} = \frac{\ell}{\sqrt{\epsilon}}, \quad \mathcal{F} = \frac{1}{500\alpha^3}\sqrt{\frac{\epsilon^3}{\rho}}, \quad \mathcal{S} = \frac{1}{4\alpha}\sqrt{\frac{\epsilon}{\rho}}.$$

$$f(x_\tau) - f(\tilde{x}) = f(x_\tau) - f(x_0) + f(x_0) - f(\tilde{x}).$$

w.h.p. $f(x_\tau) - f(x_0)$ is small.

$\leq -\mathcal{F}$.

(1) Set of bad points from where GD does not escape quickly.

(2) Volume of that set is small

$\Leftarrow$ Any two points which are separated in the -ve direction do not belong to the set.

(3) x_0 • ↕ $= \eta r_0 e_1$ • x_0' then $\min\{f(x_\tau) - f(x_0), f(x_\tau') - f(x_0')\} \leq -\mathcal{F}$.

Proof: If not, by improve or localize,

$$\|x_\tau - \tilde{x}_0\| \leq \sqrt{2\eta\tau\mathcal{F} + \eta r} \leq \zeta$$

and $\|x_\tau' - \tilde{x}\| \leq \zeta$.

Denote $\hat{x}_t \triangleq x_t - x_t'$.

$$\hat{x}_{t+1} = x_{t+1} - x'_{t+1}$$

$$= \hat{x}_t - \eta \left[\nabla f(x_t) - \nabla f(x'_t) \right]$$

$$= (I - \eta H)\hat{x}_t - \eta \Delta_t \hat{x}_t .$$

$$\nabla f(x_t) = \nabla f(\tilde{x}) + \int_{\theta=0}^{1} \nabla^2 f(\theta x_t + (1-\theta)\tilde{x}) \, d\theta$$

$$\nabla f(x_t) = \nabla f(x'_t) + \int_{\theta=0}^{1} \nabla^2 f(x'_t + \theta(x_t - x'_t)) (x_t - x'_t) \, d\theta$$

$$\hat{x}_{t+1} = (I - \eta H)\hat{x}_t - \eta \Delta_t \hat{x}_t , \text{ where}$$

$$\Delta_t \triangleq \int_{\theta=0}^{1} \left(\nabla^2 f(x'_t + \theta(x_t - x'_t)) - H \right) d\theta .$$

$$= \underbrace{(I - \eta H)^{t+1} \hat{x}_0}_{\triangleq p(t+1)} - \eta \underbrace{\sum_{\tau=0}^{t} (I - \eta H)^{t-\tau} \Delta_\tau \hat{x}_\tau}_{\triangleq q(t+1)} .$$

Claim: $\|q(t)\| \leq \frac{\|p(t)\|}{2} \quad \forall t \in \mathcal{T}$.

Proof: By induction.

Base case: $q(0) = 0$ so, true.

By Hessian Lipschitz, $\|\Delta_\tau\| \le \rho \max\{\|x_t - \tilde{x}\|, \|x'_t - \tilde{x}\|\}$

$\le \rho\varepsilon$

and $\|\hat{x}_\tau\| \le \|p(\tau)\| + \|q(\tau)\|$

(By induction hyp.) $\le 2\|p(\tau)\| = 2(1+\eta\gamma)^\tau \eta r_0$

$$\|q(t+1)\| \le \eta \sum \|(I-\eta H)^{t-\tau}\| \|\Delta_\tau\| \cdot \|\hat{x}_\tau\|$$

$$\le 2\eta\rho\varepsilon \sum_{\tau=0}^{t} (1+\eta\gamma)^t \eta r_0$$

$$\le 2\eta\rho\varepsilon\tau \|p(t+1)\|.$$

By choice, $2\eta\rho\varepsilon\tau \le \frac{1}{2}$. ∎

This implies $\max\{\|x_\tau - \tilde{x}\|, \|x'_\tau - \tilde{x}\|\}$

$$\ge \frac{1}{2}\|\hat{x}_\tau\| \ge \frac{1}{4}\|p(\tau)\|$$

$$= \frac{(1+\eta\gamma)^\tau \eta r_0}{4} > \varepsilon.$$

Contradiction. ∎

CONVEX - CONCAVE MINIMAX OPT.

$$\min_{x \in X} \max_{y \in Y} f(x, y)$$

$f(\cdot, y)$ is Convex $\forall y$

$f(x, \cdot)$ is Concave $\forall x$.

Example applications :

(1) Constrained optimization

Find x

S.t. $f_i(x) \leq 0$ $\quad i = 1, \dots m$

$$\longrightarrow \min_x \max_{i=1,\dots m} f_i(x)$$

$$= \min_x \max_{d_i \geq 0, \sum_i d_i = 1} \sum_{i=1}^{m} d_i f_i(x)$$

(2) Non smooth functions can sometimes be written as smooth minimax problems.

$$\|x\|_1 \equiv \max_{-1 \leq d_i \leq 1} \sum_i d_i x_i$$

(3) Zero Sum games.

Exercise 1: $g(x) \triangleq \max_y f(x,y)$ is a convex function since $f(\cdot, y)$ is convex $\forall y$.

Danskin's theorem: If $f(x,y)$ is differentiable with respect to x $\forall y \in Y$ and if $\frac{\partial f}{\partial x}$ is continuous w.r.t. y $\forall x$, then

$$\nabla g(x) = \text{Conv. hull}\left\{ \frac{\partial f(x,y)}{\partial x} : y \in \underset{z \in Y}{\operatorname{argmax}} f(x,z) \right\}.$$

($\nabla g(x)$: subgradient)

One way to solve the minimax problem is to do subgradient descent on $g(x)$.

① Find $y_t \in \underset{y}{\operatorname{argmax}} f(x_t, y)$

② $\nabla g(x_t) = \nabla_x f(x_t, y_t)$

③ $x_{t+1} = x_t - \eta \nabla g(x_t)$.

Issues: ① very time consuming to find y_t

② Requires f to be smooth.

Non-Smooth setting: Gradient descent ascent
↳ $\frac{1}{\sqrt{t}}$ Conv. rate.

Smooth setting: Mirror-Prox
↳ $\frac{1}{t}$ Conv. rate.

~~Weak duality: $\min_{\hat{x}} f(\hat{x}, y) \leq f(\hat{x}, \hat{y}) \leq \max_{x} f(x, \hat{y})$.~~

Weak duality: $\min_{x} f(x, \hat{y}) \leq f(\hat{x}, \hat{y}) \leq \max_{y} f(\hat{x}, y)$.

If $y^* \in \arg\max_{\hat{y} \in Y} \left[\min_{x} f(x, \hat{y}) \right]$

and $x^* \in \arg\min_{\hat{x} \in X} \left[\max_{y} f(\hat{x}, y) \right]$

then $\max_{\hat{y}} \min_{x} f(x, \hat{y}) = \min_{x} f(x, y^*)$

$\leq f(x^*, y^*)$

$\leq \max_{y} f(x^*, y)$

$= \min_{\hat{x}} \max_{y} f(\hat{x}, y)$. ∎

So, $\max_{y} \min_{x} f(x,y) \leq \min_{x} \max_{y} f(x,y)$.

Strong duality: If X and Y are compact and $f(\cdot,\cdot)$ is convex-concave, then

$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y).$$

A point $(\bar{x}, \bar{y})$ is said to be an ϵ-primal dual pair if

$$\max_{y \in Y} f(\bar{x}, y) - \min_{x \in X} f(x, \bar{y}) \leq \epsilon.$$

Exercise: If $(\bar{x}, \bar{y})$ is an ϵ-primal dual pair of ~~$f(x,y)$~~ $\min_x \max_y f(x,y)$, then

$\bar{x}$ is an ϵ-approx ~~sol~~ of minimizer of $g(x) \triangleq \max_y f(x,y)$

and $\bar{y}$ is an ϵ- " maximizer of $h(y) \triangleq \min_x f(x,y)$.

Gradient descent ascent:

$$x_{t+1} = x_t - \eta \nabla_x f(x_t, y_t)$$

$$y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t)$$

$\|\nabla_x f(x_t, y_t)\| \le G$

$\|\nabla_y f(x, y)\| \le G$

$\text{diam}(X) \le R$

$\text{diam}(Y) \le R$

Fix x.

$$\|x_{t+1} - x\|^2 = \|x_t - x\|^2 - 2\eta \langle \nabla_x f(x_t, y_t), x_t - x \rangle + \eta^2 \|\nabla_x f(x_t, y_t)\|^2$$

$$\le \|x_t - x\|^2 - 2\eta [f(x_t, y_t) - f(x, y_t)] + \eta^2 G^2$$

$$\|y_{t+1} - y\|^2 = \|y_t - y\|^2 + 2\eta \langle \nabla_y f(x_t, y_t), y_t - y \rangle + \eta^2 \|\nabla_y f(x_t, y_t)\|^2$$

$$\le \|y_t - y\|^2 - 2\eta [f(x_t, y) - f(x_t, y_t)] + \eta^2 G^2.$$

Adding,

$$\|x_{t+1} - x\|^2 + \|y_{t+1} - y\|^2 \le \|x_t - x\|^2 + \|y_t - y\|^2 - 2\eta [f(x_t, y) - f(x, y_t)] + \eta^2 G^2.$$

Telescoping and reorganizing,

$$\frac{1}{T+1}\sum_{t=0}^{T} f(x_t, y) - f(x, y_t)$$

$$\leq \frac{\|x_0 - x\|^2 + \|y_0 - y\|^2 - \|x_{T+1} - x\|^2 - \|y_{T+1} - y\|^2}{2\eta(T+1)}$$

$$+ \eta G^2.$$

By Convexity

$$f(\bar{x}_T, y) \leq \frac{1}{T+1}\sum_{t=0}^{T} f(x_t, y)$$

where $\bar{x}_T \triangleq \frac{1}{T+1}\sum_{t=0}^{T} x_t$.

Similarly $f(x, \bar{y}_T) \geq \frac{1}{T+1}\sum_{t=0}^{T} f(x, y_t)$.

So, $f(\bar{x}_T, y) - f(x, \bar{y}_T)$

$$\leq \frac{\|x_0 - x\|^2 + \|y_0 - y\|^2}{2\eta(T+1)} + \eta G^2$$

$$\max_y f(\bar{x}_T, y) - \min_x f(x, \bar{y}_T) \leq \frac{R^2}{\eta(T+1)} + \eta G^2$$

$$\eta = \frac{R}{G\sqrt{T+1}} \longrightarrow \leq \frac{2RG}{\sqrt{T+1}}.$$

Gradient descent algorithm: $x_{t+1} = x_t - \eta \nabla f(x_t)$

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

Exercise: Proximal algorithm has convergence rate $\frac{1}{T}$ for [even non-smooth] convex functions.

Exercise: Given a point x, an L-smooth function $f(\cdot)$, and $\eta \le \frac{1}{2L}$ we can find z satisfying $z = x - \eta \nabla f(z)$ up to an accuracy of ϵ in $O(\log \frac{1}{\epsilon})$ steps.

(conceptual)

Mirror-Prox:

$$x_{t+1} = x_t - \eta \nabla_x f(x_{t+1}, y_{t+1})$$

$$y_{t+1} = y_t + \eta \nabla_y f(x_{t+1}, y_{t+1})$$

Let $\bar{x}^+ = x - \eta \nabla_x f(\bar{x}, \bar{y})$

$\bar{y}^+ = y - \eta \nabla_y f(\bar{x}, \bar{y})$

Similarly, $\tilde{x}^+$ and $\tilde{y}^+$.

$$\|\bar{x}^+ - \tilde{x}^+\| + \|\bar{y}^+ - \tilde{y}^+\| \le 2\eta L \left[\|\bar{x} - \tilde{x}\| + \|\bar{y} - \tilde{y}\|\right].$$

These iterations can quickly find the CMP step.