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

(1) We know the exact form: $f(x) = \frac{1}{2} x^T A x - x^T b$

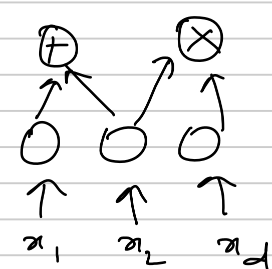
$$x^* = A^{-1} b$$

(2) $x \xrightarrow[\text{oracle}]{\text{Black box}} f(x)$] Zeroth order oracle

$x \longrightarrow f(x), \nabla f(x)$] First order oracle

$x \longrightarrow f(x), \nabla f(x), \nabla^2 f(x)$] Second order oracle.

Thm: If $f(\cdot)$ is represented by a circuit with simple nodes then the $\nabla f(x)$ can be computed in time linear in the circuit size.



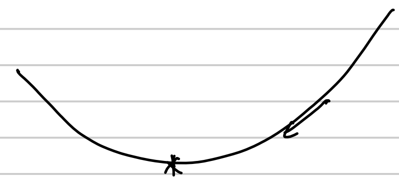
First order oracles can be implemented very efficiently.

We will focus only on first order algorithms.

OUTLINE

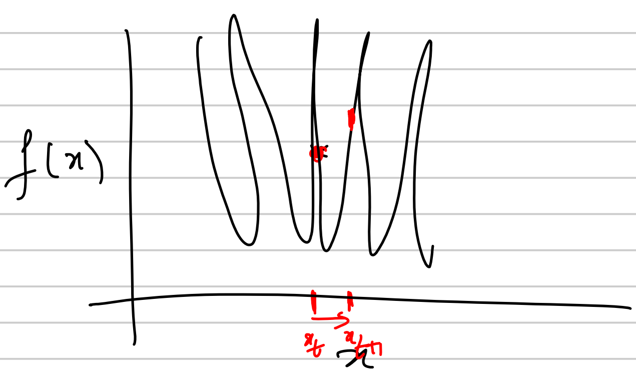
1. CONVEX OPTIMIZATION (DETERMINISTIC)
2. STOCHASTIC CONVEX OPTIMIZATION
3. NONCONVEX OPT. (DET & STOCH.)
4. CONSTRAINED / MINIMAX OPT.

ACCESS TO $x \rightarrow f(x), \nabla f(x)$



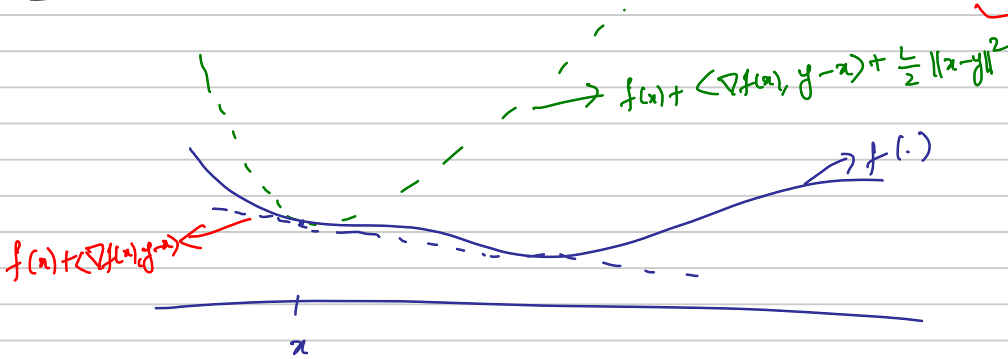
Gradient descent: Start with x_0

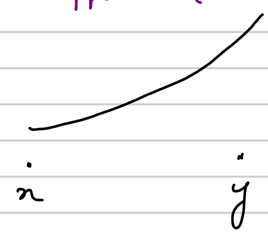
$$x_{t+1} = x_t - \eta \nabla f(x_t).$$



[DEFN.] <u>SMOOTHNESS</u> : A function $f(\cdot)$ is said to be L-Smooth if $\|\nabla f(x) - \nabla f(y)\| \le L \cdot \|x-y\|$. $\forall x, y$.

<u>LEMMA</u> : If $f(\cdot)$ is L-Smooth then $f(y) \le \underbrace{f(x) + \langle \nabla f(x), y-x \rangle}_{\text{First order Taylor approximation}} + \underbrace{\frac{L}{2}\|x-y\|^2}_{\text{quadratic upper bound}}$ $\forall x, y$.





<u>Proof</u>: Given x and y,

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

$$= \boxed{f(x) + \langle \nabla f(x), y-x \rangle} \longrightarrow \text{Linear Taylor expansion}$$

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

$$S \le \int_{\tau=0}^{1} \|\nabla f((1-\tau)x + \tau y) - \nabla f(x)\| \cdot \|y-x\| \, d\tau.$$

$$\overset{\text{(smoothness)}}{\leq} \int_{\tau=0}^{1} L \cdot \| (1-\tau)x + \tau y - x \| \cdot \|y-x\| \, d\tau$$

$$= L \, \|x-y\|^2 \int_{\tau=0}^{1} \tau \, d\tau = \frac{L}{2} \|x-y\|^2$$

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

<u>Thm</u>: If $f(\cdot)$ is L-Smooth and $\eta \leq \frac{1}{L}$ then

$$\min_{t=1,\dots T} \|\nabla f(x_t)\|^2 \leq \frac{2\,(f(x_0) - f^*)}{\eta T}. \qquad f^* \triangleq \min_x f(x).$$

<u>Proof</u>:

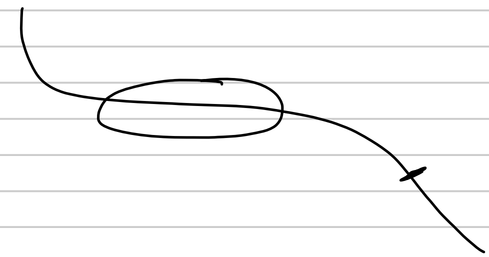
$$f(x_{t+1}) \overset{\text{(smoothness lemma)}}{\leq} f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_t - x_{t+1}\|^2$$

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

$$= f(x_t) - \eta \underbrace{\left(1 - \frac{\eta L}{2}\right)}_{\geq \frac{1}{2}} \|\nabla f(x_t)\|^2$$

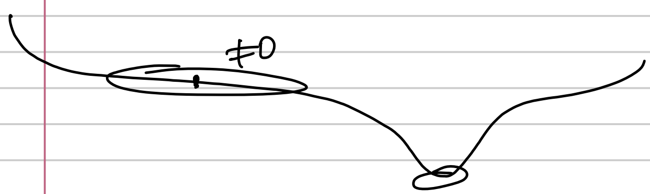
$$\leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2$$

$$f^* \le \underbrace{f(x_{T+1})} \le f(x_0) - \frac{\eta}{2} \sum_{t=0}^{T} \|\nabla f(x_t)\|^2$$



Reorganizing proves the theorem $\blacksquare$

Consider a special class of functions where $\nabla f(x)=0 \Rightarrow x$ is global optimum.



$\boxed{\mathcal{F}}$: $\left\{ \begin{array}{l} ① \ \underline{\nabla f(x)=0 \Rightarrow x \text{ is global opt. of } f(\cdot)} \\ ② \ \left[\begin{array}{l} \text{If } g(x) = \underbrace{f(x) + \langle w, x\rangle} \text{ then} \\ \nabla g(x)=0 \Rightarrow x \text{ is global opt. of } g(\cdot). \end{array} \right. \end{array} \right.$

__Lemma:__ $\boxed{\mathcal{F}} = \left\{ f : f(y) \ge f(x) + \langle \nabla f(x), y-x \rangle \ \forall x,y \right\}.$

__Proof:__ Any $f \in \mathcal{F}$ satisfies

Given x and y; $\underbrace{g(y)} = \underbrace{f(y)} + \underbrace{\langle -\nabla f(x), y-x\rangle}$

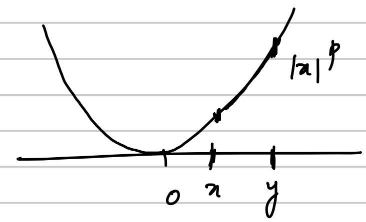
$\nabla g(y) = \nabla f(y) - \nabla f(x) \Rightarrow \nabla g(x)=0 \Rightarrow x$ is global opt. of $g(\cdot)$.

$\Rightarrow g(x) \le g(y) \quad \forall y$

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

[Defn.] <u>CONVEX FNS</u>: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall\, x, y$

<u>Examples</u>: ① $f(x) = |x|^p \qquad p > 1$



$$\frac{df}{dx} = p\, x^{p-1}.$$

$$f(y) = f(x) + \underbrace{(y-x)\int_{\alpha=0}^{1} \left.\frac{df}{dx}\right|_{(1-\alpha)x+\alpha y} \cdot d\alpha}$$

$$\left.\frac{df}{dx}\right|_{(1-\alpha)x+\alpha y} = p\left[(1-\alpha)x + \alpha y\right]^{p-1} \geq p\, x^{p-1}$$

$$f(y) \geq f(x) + \underbrace{p\, x^{p-1}} \cdot \underbrace{(y-x)}.$$

② If $f_1(\cdot)$ is convex and $f_2(\cdot)$ is convex then

$f(x) = f_1(x) + f_2(x)$ is also convex.

$$\begin{aligned} f(y) &= f_1(y) + f_2(y) \\ &\geq f_1(x) + \langle \nabla f_1(x), y - x \rangle + f_2(x) + \langle \nabla f_2(x), y - x \rangle \\ &= f(x) + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

(3) $x = (x_1, x_2)$ $\qquad f_1(x) = x_1^2$

$$f_2(x) = x_2^2$$

So, $f(x) = \|x\|^2 = x_1^2 + x_2^2$ is also convex.

$f(x) = \|x\|_p^p$ is also convex

(4) If $f(\cdot)$ is convex then $f(Ax+b)$ is also convex.

Let $g(x) \triangleq f(Ax+b)$

$$g(y) = f(\underbrace{Ay+b}_{w_y}) \overset{f(\cdot) \text{ is convex}}{\geq} f(\underbrace{Ax+b}_{w_x}) + \langle \nabla f\big|_{Ax+b}, (Ay+b) - (Ax+b) \rangle$$

$$= g(x) + \langle \nabla f\big|_{Ax+b}, A(y-x) \rangle$$

$$= g(x) + \langle \underbrace{A^T \nabla f\big|_{Ax+b}}, y-x \rangle$$

$$= g(x) + \langle \nabla g(x), y-x \rangle.$$

$\min_x \frac{1}{2}\|Ax-b\|^2.$

$f(x) = \|Ax-b\|^2$ is convex.

Find x s.t. $a^T \underbrace{x} \approx b$

Temp	Humidity	...	Measure Rainfall
a_1	a_2	. a_d	b

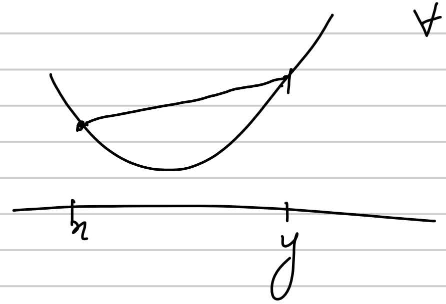
LEC - 2

$$\boxed{f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle}$$

Lemma: If $f(\cdot)$ is a Convex function then $f(\alpha x + (1-\alpha) y)$

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

$$\forall\ x, y,\ \alpha \in [0,1].$$



Proof: Exercise.

$\square$

$\widehat{\text{Convex}}$: $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$

Lemma: If f is $\widehat{\text{Convex}}$ then $\forall\ x$ in the domain, $\exists$ non-empty set $\boxed{G(x)}$ $\to$ Subgradients

s.t. $f(y) \geq f(x) + \langle z, y - x \rangle \quad \forall\ x, y$ and $z \in G(x)$.

MORALLY: CONVEX $=$ $\widehat{\text{CONVEX}}$.

CONVEX $\to$ First defn.

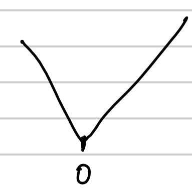
$\widehat{\text{CONVEX}}$ $\to$ Second defn.

Lasso: $\underline{\min_x \frac{1}{2} \|Ax - b\|^2 + \|x\|_1}$

$f(x) = |x|$

$$\frac{df}{dx} = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$

does not exist at $x = 0$.



$f(\cdot)$ is convex $\quad \Big| \quad f(y) \geq f(x) + \langle \underbrace{\nabla f(x)}_{\text{Subgradient}}, y - x \rangle.$

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

$\boxed{\eta_t = \frac{1}{\sqrt{t}}}$

Let $x^* = \arg\min_x f(x).$

Theorem: If $f(\cdot)$ is G-Lipschitz $[\|\nabla f(x)\| \leq G \;\; \forall x]$, then

$$\frac{1}{T} \sum_{t=1}^{T} \underbrace{\left[f(x_t) - f(x^*) \right]}_{\text{Suboptimality of } x_t} \leq \frac{\|x_0 - x^*\|^2}{2\eta T} + \frac{\eta G^2}{2}.$$

Remarks: $\eta^* = \dfrac{\|x_0 - x^*\|}{G\sqrt{T}}$; RHS $= \dfrac{\|x_0 - x^*\| \cdot G}{\sqrt{T}}.$

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

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

$y = x^*$, $x = x_t$ $\quad \left[f(x^*) \geq f(x_t) + \underbrace{\langle \nabla f(x_t), x^* - x_t \rangle} \right.$

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

Telescopic sum: $0 \leq \|x_{T+1} - x^*\|^2 \leq \|x_0 - x^*\|^2 - 2\eta \sum_{t=0}^{T} \left[f(x_t) - f(x^*) \right] + \eta^2 G^2 (T+1).$

Dividing by $2\eta(T+1)$ gives the theorem.

Remarks: ① The Convergence rate is independent of dimension

② In general we do not know G or $\|x_0 - x^*\|$.
Need to learn them on the go.

LOWER BOUNDS

Gradient span Algorithms (GSA): $x_{t+1} \in \text{Span}\{x_0, \nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_t)\}$.

GD $\in$ Gradient span algorithms.

Theorem: For any value of G, R and T $\exists$ a function $f_{G,R,T}(x)$ s.t.

0) f is convex

i) $\|\nabla f_{G,R,T}(x)\| \leq G$

ii) $\|x_0 - x^*\| \leq R$ and

iii) for any GSA $f_{G,R,T}(\bar{x}) - f_{G,R,T}(x^*) \geq \dfrac{GR}{2(1+\sqrt{T+1})}$

$\bar{x}$: any vector in the span $\{x_0, \ldots x_T\}$

Remark: Matches upper bound upto Constant factors. [GD is optimal for this class of functions]

Proof:

$$f_{\gamma,\mu}(x) = \gamma \max_{1 \leq i \leq T+1} x(i) + \frac{\mu}{2}\|x\|^2.$$

x : dimension $= T+1$

Ex: Show that $f_{\gamma,\mu}(\cdot)$ is convex.

i) ONLY PROPERTY WE NEED: $f(y) \geq f(x) + \langle z, y-x \rangle \ \forall x, y$

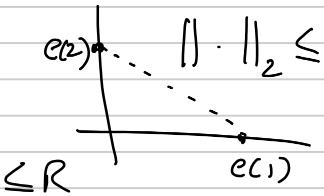
Any z that satisfies this is a Subgradient of f at x

$$f(x) = \max_{i} f_i(x)$$

Let $I(x) = \{ i : f_i(x) = f(x) \}$.

Any vector $z \in$ Convex hull of $\{\nabla f_i(x) : i \in I(x)\}$ is a Subgradient of $f(\cdot)$ at x.

$$\nabla f(x) = \gamma \text{ Conv hull}\left(e(i) : i \in I(x)\right) + \mu x.$$



$\|\cdot\|_2 \le 1$.

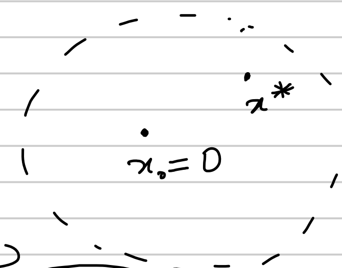
$$\|\nabla f(x)\| \le \| \quad \| + \mu \underbrace{\|x\|}.$$

$$\le \gamma \cdot 1 + \mu R$$

If $\|x\| \le R$

① $\gamma + \mu R \le G$

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



$\|x^*\| \le R$

$$x(i) = \alpha$$

$$f(x) = \gamma\alpha + \frac{\mu}{2}\alpha^2(T+1) \rightarrow \alpha\mu(T+1) = -\gamma$$

$$\underbrace{\alpha = \frac{-\gamma}{\mu(T+1)}}$$

① $\gamma + \mu R \le G$

$x_0 = 0$ ② $\|x_0 - x^*\|^2 = \|x^*\|^2$

$$= \frac{\gamma^2}{\mu^2(T+1)} \le R^2.$$

$$\boxed{f(x^*) = \frac{-\gamma^2}{2\mu(T+1)} \le 0}$$

$$x^* = -\frac{\gamma}{\mu(T+1)}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$$

Choice: $\gamma = \dfrac{\sqrt{T+1}\cdot G}{1+\sqrt{T+1}}$ and $\mu = \dfrac{G}{R(1+\sqrt{T+1})}$.

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

$x_0 = 0.$

$$\nabla f(x_0) = \gamma \text{ Conv. hull}(e_i : i \in \underbrace{I(0)}) + \mu\, 0$$

Resisting oracle : $\nabla f(x) = \gamma e_{i(x)} + \mu x$ where $i(x) = \min I(x).$

$x_0 = 0$

$$\nabla f(x_0) = \gamma e(1) + \mu \cdot 0 = \gamma e(1)$$

$$x_1 \in \text{Span}\{x_0, \nabla f(x_0)\} = \text{Span}(e(1))$$

$x_1 = \alpha e_1$, $x_0 = 0$, $\nabla f(x_0) = \text{span}(e(1))$

Let $x_1 = \alpha e_1$. If $\alpha \ge 0$ then $\nabla f(x_1) = \gamma e(1) + \mu x_1$

$\alpha < 0$ then $\nabla f(x_1) = \gamma e(2) + \mu x_1.$

$$\Longrightarrow \nabla f(x_1) \in \text{Span}\{e(1), e(2)\}.$$

$$f(x) = \gamma \underbrace{\max\{\underset{=\alpha}{x(1)}, \underset{=0}{x(2)}, \underset{x(3),\dots}{}\}} + \frac{\mu}{2}\|x\|^2.$$

$f_1(x) \Longrightarrow f_2(x)$

$$x_1 = \alpha e_1 = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}.$$

If $\alpha > 0$ then $I(x_1) = 1 \longrightarrow \gamma e_1$

If $\alpha = 0$ then $I(x_1) = 1, 2 \longrightarrow \gamma e_1$

If $\alpha < 0$ then $I(x_1) = 2 \longrightarrow \gamma \boxed{e_2}$

$$x_t \in \text{Span}\{e(1), \dots, e(t)\}.$$

$x_t(t+1) = 0$

$$f(x_t) = \gamma \max_{1 \le i \le T+1} \underbrace{x_t(i)} + \frac{\mu}{2} \|x_t\|^2$$

$x_t(T+1) = 0.$

So, $f(x_t) \ge 0.$

$$\gamma = \frac{\sqrt{T+1} \cdot G}{1+\sqrt{T+1}} \quad ; \quad \mu = \frac{G}{R(1+\sqrt{T+1})} \quad ; \quad f(x_T) \geq 0.$$

$$f(x^*) = \frac{-\gamma^2}{2\mu(T+1)}$$

$$f(x_T) - f(x^*) \geq \frac{\gamma^2}{2\mu(T+1)} = \frac{(T+1)\,G^2}{(1+\sqrt{T+1})^2} \cdot \frac{R(1+\sqrt{T+1})}{2G} \cdot \frac{1}{T+1}$$

$$= \frac{GR}{2(1+\sqrt{T+1})}. \qquad \blacksquare$$

$$f(x) = \max\{f_1(x), f_2(x)\}.$$

LECTURE - 3

Note Title 26-Jun-19

f is Convex

$\|\nabla f(x)\| \leq G$

$\|x_0 - x^*\| \leq R$

GD: $\frac{1}{T} \sum_{t=1}^{T} \left(f(x_t) - f(x^*)\right) \leq \frac{GR}{\sqrt{T}}$

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

$x_0 = 0 \xRightarrow{\text{Sub.Gr.De.}}$

$R \triangleq \|x_0 - x^*\| = \|x^*\|$

$G \triangleq \underbrace{\max_{x : \|x\| \leq R} A^T(Ax - b)}$ $\longrightarrow$ $G = \|A\|^2 \cdot R + \|A\| \cdot \|b\|$

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t)-f(x^*)\right) \le \frac{GR}{\sqrt{T}} = \frac{\|A\|^2\cdot R^2 + \|A\|\cdot\|b\|\cdot R}{\sqrt{T}}.$$

Qn.: Can we do better?

↳ Say for Smooth functions.

Smoothness: $\|\nabla f(x) - \nabla f(y)\| \le L\cdot\|x-y\|$.

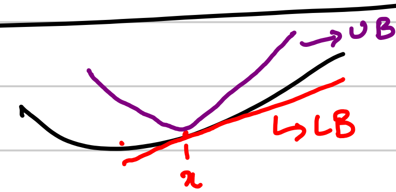
$$\downarrow$$

$$\|A^T(Ax-b) - A^T(Ay-b)\| = \|A^TA(x-y)\|$$

So, $L \triangleq \|A^TA\|$.

This lecture: ① Convergence rate of GD for smooth convex fns.

② Optimal method: Nesterov's accelerated gradient.



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

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

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

Thm: If f is an L-smooth convex fn. then GD with $\eta = \boxed{\frac{1}{2L}}$, we have $\underbrace{f(x_T)}_{\text{Final iterate}} - f(x^*) \leq \frac{C \cdot L \cdot \|x_0 - x^*\|^2}{T}$.

Lemma: If $f(\cdot)$ is L-smooth then $\underbrace{\|\nabla f(x)\|^2 \leq 2L \cdot [f(x) - f(x^*)]}$.

Proof: $f(x^*) \leq f(\underbrace{x - \frac{1}{L}\nabla f(x)})$

$$\leq f(x) + \langle \nabla f(x), -\frac{1}{L}\nabla f(x) \rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(x)\right\|^2$$

$$= f(x) - \frac{1}{2L}\|\nabla f(x)\|^2. \quad \blacksquare$$

Proof of thm: $\quad x_{t+1} = x_t - \eta \nabla f(x_t)$

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

$$= \|x_t - x^*\|^2 - 2\eta \langle \nabla f(x_t), x_t - x^* \rangle + \eta^2 \underbrace{\|\nabla f(x_t)\|^2}_{\text{Use Lemma above}}$$

$$\leq \|x_t - x^*\|^2 - 2\eta \underbrace{\langle \nabla f(x_t), x_t - x^* \rangle}_{\text{Convexity}} + \eta^2 \cdot 2L\,[f(x_t) - f(x^*)]$$

$$\leq \|x_t - x^*\|^2 - 2\eta\,[f(x_t) - f(x^*)] + 2\eta^2 L\,[f(x_t) - f(x^*)]$$

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

Take telescopic Sum,

$$\|x_{T+1}-x^*\|^2 \le \|x_0-x^*\|^2 - 2\eta(1-\eta L)\sum_{t=0}^{T}[f(x_t)-f(x^*)]$$

$$\frac{1}{T+1}\sum_{t=0}^{T}[f(x_t)-f(x^*)] \le \frac{\|x_0-x^*\|^2 - \underbrace{\|x_{T+1}-x^*\|^2}_{+ve}}{2\eta(1-\eta L)\cdot(T+1)}$$

Choose $\eta = \frac{1}{2L}$ $\therefore$ $\le \dfrac{2L\|x_0-x^*\|^2}{T+1}$.

Exercise: For L-smooth, convex $f(\cdot)$, show that GD with $\eta \leq \frac{1}{2L}$ satisfies $f(x_{t+1}) \leq f(x_t)$. $\forall t$.

Nesterov's accelerated gradient

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

$$x_{t+1} \triangleq \arg\min_x \underbrace{\Big[\underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}_{\text{First order Taylor exp.}} + \underbrace{\frac{1}{2\eta} \|x - x_t\|^2}_{\text{Quadratic term}} \Big]}_{\text{Local upper bound when } \eta \leq \frac{1}{L}.}$$

$$f(x) \leq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \underbrace{\frac{L}{2} \|x - x_t\|^2}_{}$$

If $\quad L \leq \frac{1}{\eta}$ then $\qquad \leq \frac{1}{2\eta} \|x - x_t\|^2.$

Estimate Sequences

$$\phi_0(x) = \underline{f(v_0)} + \frac{L}{2} \|x - v_0\|^2.$$

$$\phi_{t+1}(x) = (1 - \alpha_t)\, \phi_t(x) + \alpha_t \Big[\underbrace{f(y_t) + \langle \nabla f(y_t), x - y_t \rangle}_{\leq f(x)} \Big]$$

Lemma: If $\exists\, x_t$ s.t. $f(x_t) \leq \min_x \Phi_t(x) \triangleq \phi_t^*$ then

$$f(x_t) - f(x^*) \leq \underbrace{\left[\prod_{s=0}^{t-1}(1-\alpha_s)\right]}\left[\Phi_0(x^*) - f(x^*)\right].$$

Proof: $f(x_t) - f(x^*) \xrightarrow{\text{Hypothesis}} \leq \phi_t^* - f(x^*)$

$$\leq \underline{\Phi_t(x^*) - f(x^*)}.$$

$-ve$

$$= (1-\alpha_{t-1})\Phi_{t-1}(x^*) + \alpha_{t-1}\left[f(y_{t-1}) + \langle \nabla f(y_{t-1}), x^* - y_{t-1}\rangle\right]$$

$$- (1-\alpha_{t-1}) f(x^*) - \alpha_{t-1}\left[f(x^*)\right]$$

$$\phi_t(x^*) - f(x^*) \leq (1-\alpha_{t-1})\left[\phi_{t-1}(x^*) - f(x^*)\right]$$
$$\leq \prod_{s=0}^{t-1}(1-\alpha_s)\left[\phi_0(x^*) - f(x^*)\right]. \quad (11)$$

Given : x_t s.t. $f(x_t) \leq \min_x \phi_t(x)$.

Also given ϕ_t

Task : ① Choose y_t and query $\nabla f(y_t)$.

② choose $\alpha_t \longrightarrow \phi_{t+1}$

③ Find x_{t+1} s.t. $f(x_{t+1}) \leq \min_x \phi_{t+1}(x)$.

Lemma : If $\phi_{t+1}(x) = (1-\alpha_t)\,\phi_t(x) + \alpha_t\left[f(y_t) + \langle \nabla f(y_t), x - y_t \rangle\right]$

then $\phi^*_{t+1} \triangleq \min_x \phi_{t+1}(x) =$

and $v_{t+1} \triangleq \arg\min_x \phi_{t+1}(x) = v_t - \frac{\alpha_t}{\lambda_t (1-\alpha_t) L} \nabla f(y_t)$

Proof :

$\phi^*_t \triangleq \min_x \phi_t(x)$

$v_t \triangleq \arg\min_x \phi_t(x)$

$\lambda_t \triangleq \prod_{s=0}^{t-1} (1-\alpha_s)$

$\phi_0(x) = \phi^*_0 + \frac{L}{2}\|x - v_0\|^2$

$\phi_1(x) = (1-\alpha_0)\,\phi_0(x) + \alpha_0\left[a + \langle b, x \rangle\right]$

$$\phi_t(x) = \phi_t^* + \frac{\lambda_t L}{2} \| x - v_t \|^2 .$$

$$\phi_{t+1}(x) = (1-\alpha_t)\, \phi_t(x) + \alpha_t \left[f(y_t) + \langle \nabla f(y_t), x - y_t \rangle \right]$$

$$= (1-\alpha_t)\, \phi_t^* + \frac{(1-\alpha_t)\lambda_t L}{2} \| x - v_t \|^2$$

$$+ \alpha_t f(y_t) + \alpha_t \langle \nabla f(y_t), x - y_t \rangle .$$

$$= \frac{(1-\alpha_t)\lambda_t L}{2} \left[\|x\|^2 - 2 \langle v_t, x \rangle + \underline{\|v_t\|^2} \right]$$

Constant

Quadratic

Linear

$$+\ \alpha_t \underline{\langle \nabla f(y_t), x \rangle} \quad \boxed{\begin{aligned} &+ (1-\alpha_t)\phi_t^* + \alpha_t f(y_t) \\ &- \alpha_t \langle \nabla f(y_t), y_t \rangle \end{aligned}} \ \hat{\theta}$$

$$= \frac{(1-\alpha_t) d_t^2}{2} \Big[\|x\|^2 - 2 \Big\langle v_t - \frac{\alpha_t}{(1-\alpha_t) d_t^2} \nabla f(y_t), x \Big\rangle$$

$$+ \underline{\Big\| v_t - \frac{\alpha_t}{(1-\alpha_t) d_t^2} \nabla f(y_t) \Big\|^2}$$

Extra terms

$$\begin{cases} + \dfrac{2\alpha_t}{(1-\alpha_t) d_t^2} \langle \nabla f(y_t), v_t \rangle \\ - \left(\dfrac{\alpha_t}{(1-\alpha_t) d_t^2} \right)^2 \|\nabla f(y_t)\|^2 \end{cases} \Big] + \hat{\theta}$$

$$= \frac{(1-\alpha_t)\, d_t L}{2} \Big\| x - \underbrace{v_t + \frac{\alpha_t}{(1-\alpha_t) d_t L} \nabla f(y_t)}_{v_{t+1}} \Big\|^2 + \text{Extra terms} + \hat{\theta}$$

$$\phi^*_{t+1} = (1-\alpha_t)\phi^*_t + \alpha_t f(y_t) - \frac{\alpha_t^2}{2 d_t (1-\alpha_t) L} \|\nabla f(y_t)\|^2 + \alpha_t \langle \nabla f(y_t), v_t - y_t \rangle .$$

How to find x_{t+1} s.t. $f(x_{t+1}) \leq \phi^*_{t+1}$?

By smoothness: $f(\underbrace{y_t - \overset{\frac{1}{L}}{\eta} \nabla f(y_t)}_{x_{t+1}}) \le f(y_t) - \eta \|\nabla f(y_t)\|^2 + \frac{\eta^2 L}{2} \|\nabla f(y_t)\|^2$

$$\le f(y_t) - \frac{1}{2L} \|\nabla f(y_t)\|^2$$

$$f(x_{t+1}) - \phi_{t+1}^* \le f(y_t) - \frac{1}{2L} \|\nabla f(y_t)\|^2 - (1-\alpha_t)\phi_t^* - \alpha_t f(y_t) + \frac{\alpha_t^2}{2\alpha_t(1-\alpha_t)L} \|\nabla f(y_t)\|^2 - \alpha_t \langle \nabla f(y_t), v_t - y_t \rangle$$

$$\leq (1-\alpha_t) f(y_t) - (1-\alpha_t) f(x_t)$$

$$- \alpha_t \langle \nabla f(y_t), v_t - y_t \rangle$$

(Convexity: $x = y_t$, $y = x_t$)

$$\leq \langle \nabla f(y_t), (1-\alpha_t)(y_t - x_t) \rangle - \alpha_t \langle \nabla f(y_t), v_t - y_t \rangle$$

$$= \langle \nabla f(y_t), y_t - (1-\alpha_t) x_t - \alpha_t v_t \rangle$$

$$\leq 0$$

Want: $$\frac{\alpha_t^2}{d_t(1-\alpha_t)} \leq 1$$

Requirements

$$\frac{\alpha_t^2}{2 d_t (1-\alpha_t) L} \leq \frac{1}{2L}$$

Algorithm

$y_t = (1-\alpha_t) x_t + \alpha_t v_t$

$x_{t+1} = y_t - \frac{1}{L} \nabla f(y_t)$

$v_{t+1} =$ Update from Lemma.

and minimize $\lambda_t = \prod_{s=0}^{t-1} (1-\alpha_s)$.

Can choose α_t: $\boxed{\lambda_{t+1} \leq \frac{4}{t^2}.}$

$$f(x_T) - f(x^*) \leq \lambda_T \left[\phi_0(x^*) - f(x^*) \right]$$

$$\leq \frac{4}{(T-1)^2} \left[f(v_0) + \frac{L}{2} \| x^* - v_0 \|^2 - f(x^*) \right]$$

$$\leq \frac{4}{(T-1)^2} \left[L \| x^* - v_0 \|^2 \right].$$

Comparing with GD: $\frac{1}{T^2}$ instead of $\frac{1}{T}$.

LECTURE - 4

$$f(x) = \frac{1}{2}\|Ax-b\|^2 \; ; \quad L = \|A\|^2$$

$$\overset{AGD}{\Longrightarrow} \quad f(x_T) - f(x^*) \leq \frac{6\|A\|^2 \cdot \|x_0 - x^*\|^2}{T^2}$$

Can we do better?

$\hookrightarrow$ Strong convexity.

Strong Convexity : f is μ- Strongly convex

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

Linear regression : $\mu = \sigma_{min}(AA^T)$.

$\frac{1}{2}\|Ax-b\|^2$

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

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

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

$$+ \eta^2 \| \nabla f(x_t) \|^2$$

$\rightarrow$ Strong Convexity

$\rightarrow$ Smoothness

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

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

$$\le (1-\eta\mu)\|x_t-x^*\|^2 \underbrace{-2\eta(1-\eta L)\left[f(x_t)-f(x^*)\right]}$$

If we choose $\boxed{\eta\le\frac{1}{L}}$ then ≤ 0

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

$$f(x_t) - f(x^*) \le \frac{L}{2} \| x_t - x^* \|^2$$

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

$$\le \frac{L \cdot \| x_0 - x^* \|^2}{2} \cdot e^{-\eta \mu t} .$$

$\forall w \in \mathbb{R}$

FACT: $1 - w \le e^{-w}$

Choose η as large as possible

so, $\eta = \frac{1}{L}$.

$$f(x_t) - f(x^*) \le \frac{L \| x_0 - x^* \|^2}{2} \cdot e^{-\boxed{\frac{\mu}{L}} t} .$$

condition number $\kappa = \frac{L}{\mu} \ge 1$

$$f(x_t) - f(x^*) \leq \frac{L \|x_0 - x^*\|^2}{2} \cdot \underbrace{e^{-\frac{t}{\kappa}}}$$

Is this the best possible rate?

Acceleration for Smooth and Strongly Convex functions

Nesterov's AGD : $f(x_t) - f(x^*) \leq \frac{L \|x_0 - x^*\|^2}{2 t^2}$.

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

$f(x_0) \geq f(x^*)$

$+ \langle \cancel{\nabla f(x^*)}^{=0}, x_0 - x^* \rangle$

$+ \frac{\mu}{2} \|x_0 - x^*\|^2$

$$= \frac{K}{t^2}\left[f(x_0)-f(x^*)\right].$$

$$\boxed{\|x_0 - x^*\|^2 \le \frac{2}{\mu}\left[f(x_0) - f(x^*)\right]}$$

Choosing $t = \sqrt{2K}$ iterations,

$$f(x_{\sqrt{2K}}) - f(x^*) \le \frac{1}{2}\left[f(x_0) - f(x^*)\right]$$

$$f(x_{2\cdot\sqrt{2K}}) - f(x^*) \le \frac{1}{2}\left[f(x_{\sqrt{2K}}) - f(x^*)\right]$$

$$f(x_{i\cdot\sqrt{2K}}) - f(x^*) \le \frac{1}{2}\left[f(x_{(i-1)\cdot\sqrt{2K}}) - f(x^*)\right]$$

$$f(x_t) - f(x^*) \le \left(\frac{1}{2}\right)^{\left(t/\sqrt{2\kappa}\right)} [f(x_0) - f(x^*)].$$

GD: $e^{-t/\kappa}$; AGD: $e^{-t/\sqrt{\kappa}}$

$\|x_0 - x^*\| \le R$

CLASS	Algorithm	Guarantee
G-Lipschitz $\|\nabla f\| \le G$	SUBGRADIENT DESCENT	$\frac{GR}{\sqrt{T}}$
L-SMOOTH $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$	G.D.	$\frac{LR^2}{T}$
	A.G.D.	$\frac{LR^2}{T^2}$
L-SMOOTH μ- Str. Cvx.	G.D.	$LR^2 \cdot \exp\left(\frac{-\mu T}{L}\right)$
	A.G.D.	$LR^2 \cdot \exp\left(-\sqrt{\frac{\mu}{L}} \cdot T\right)$

G- Lipschitz μ- Str. cvx.	Sub. G.D	$\frac{G^2}{\mu T}$

CONSTRAINED OPTIM.

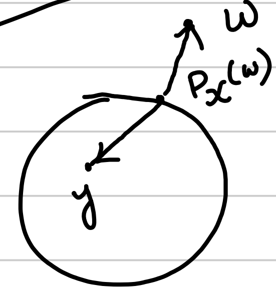
$$\min_{x \in \mathcal{X}} f(x)$$

$\mathcal{X}$ is a Simple closed Convex set.

We have access to projection operator

Given x, $\quad P_{\mathcal{X}}(x) = \underset{y \in \mathcal{X}}{\operatorname{argmin}} \|x-y\|^2.$

$\mathcal{X} = B_1(w)$ is a simple set



Pythagoras theorem: $\forall w$, we have

EXERCISE

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

ii) $\| P_{\mathcal{X}}(w) - y \|^2 \leq \| w - y \|^2$ $\forall y \in \mathcal{X}$

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

Equivalent

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \Big[\underbrace{\overbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}^{\text{Linear}} + \overbrace{\boxed{\tfrac{1}{2\eta} \|x - x_t\|^2}}^{\frac{1}{\eta}\text{-str. conv. quadratic}}}_{\hat{f}_\eta(x_t;\, x)} \Big].$$

Doing now
Conv. rate of PGD for L-Smooth fun.

Exercise: $\hat{f}_\eta(x_t; x)$ is $\frac{1}{\eta}$-Strongly Convex

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

$\boxed{+\langle\ \rangle}$ $\hookrightarrow$ Use Pythagoras thm.

For $\eta \leq \frac{1}{L}$; $f(x) \leq \hat{f}_\eta(x_t; x)$ [by L-smoothness of $f(\cdot)$]

$$f(x) \overset{\text{Conv.}}{\geq} f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$$

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

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

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

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

$x = x^*$ and iterate

$$\|x^* - x_{t+1}\|^2 \le \|x^* - x_0\|^2 - 2\eta \underbrace{\sum_{s=1}^{t+1} \left[f(x_s) - f(x^*) \right]}.$$

$$\frac{1}{t+1} \sum_{s=1}^{t+1} \left[f(x_s) - f(x^*) \right] \le \frac{\|x^* - x_0\|^2 - \|x^* - x_{t+1}\|^2}{2\eta(t+1)}.$$

$$\left[\eta = \frac{1}{L}\right] \qquad \le \frac{L\|x_0 - x^*\|^2}{2(t+1)}. \qquad \blacksquare$$

Ref: INTRODUCTORY LECTURES ON CONVEX OPT.

— YURII NESTEROV

LECTURE-5

Note Title 27-Jun-19

Faster algorithms (better than the black-box bounds) for structured non-smooth problems

$$f(x) = \underbrace{g(x)}_{\text{Convex \& smooth}} + \underbrace{h(x)}_{\text{Convex \& perhaps non-smooth}}$$

Access to : $\underset{x}{\arg\min} \; \langle w, x \rangle + h(x) + \frac{1}{2\eta}\|x-y\|^2$

$\forall \; w, \eta, y.$

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

GD with Prox

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

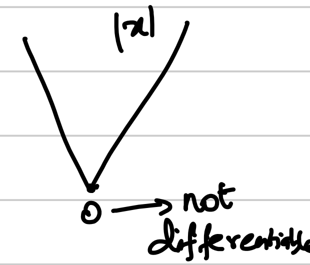
GD : $$x_{t+1} = \arg\min_{x} \; g(x_t) + \langle \nabla g(x_t), x - x_t \rangle + h(x_t) + \langle \nabla h(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|^2.$$

LASSO: $f(x) = \underbrace{\frac{1}{2}\|Ax-b\|^2}_{g(x)} + \underbrace{\lambda \|x\|_1}_{h(x)}$

(Compressed Sensing)

$g(x)$: Linear regression

$h(x)$: encourage sparsity of x.



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

$$= \arg\min_x \langle \nabla g(x_t), x \rangle + \underbrace{h(x)}_{\lambda \|x\|_1} + \frac{1}{2\eta}\|x - x_t\|^2$$

$$= \arg\min_{x} \sum_{i=1}^{d} x_i \nabla g(x_t)_i + |x_i| + \frac{1}{2\eta}\left(x_i - (x_t)_i\right)^2$$

$$(x_{t+1})_i = \arg\min_{x_i} \; x_i \nabla g(x_t)_i + |x_i| + \frac{1}{2\eta}\left(x_i - (x_t)_i\right)^2$$

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

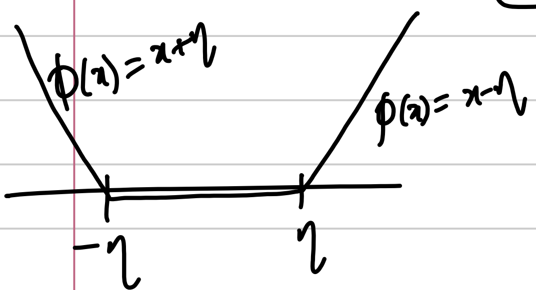
Exercise:

$$(x_{t+1})_i = (x_t)_i - \eta \nabla g(x_t)_i - \eta \quad \text{if } \boxed{(x_t)_i - \eta \nabla g(x_t)_i} \geq \eta$$

$$= \quad \text{''} \quad + \eta \quad \text{if} \quad \text{''} \quad \leq -\eta$$

$$= 0 \quad \text{Otherwise.}$$

$$= \phi\left(x_t - \eta \nabla g(x_t)\right),$$ where ϕ is applied to each coordinate.



$$x_{t+1} = \underbrace{x_t - \eta \nabla g(x_t)}_{\triangleq\, \hat{x}_{t+1}} + \eta\, \mathbb{1}\{\text{coordinates} < -\eta\}$$

$$- \eta\, \mathbb{1}\{\quad > \eta\}$$

$$- \hat{x}_{t+1}\, \mathbb{1}\{-\eta \le \text{coord.} \le \eta\}$$

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

$x^* = \underset{x}{\arg\min} f(x)$

$$x_{t+1} = \underset{x}{\arg\min}\ \hat{f}_\eta(x_t ; x).$$

g is L-Smooth
$\eta \le \frac{1}{L}.$

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

Conv. of $g(\cdot)$

$$\ge g(x_t) + \langle \nabla g(x_t), x^* - x_t \rangle + h(x^*) + \frac{1}{2\eta}\|x^* - x_t\|^2 - \frac{1}{2\eta}\|x^* - x_t\|^2$$

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

$\left[x_{t+1} = \arg\min \hat{f}_\eta \;\&\; \hat{f}_\eta - \frac{1}{\eta} \text{ str. cvx.} \right]$

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

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

L-smoothness of $g(\cdot)$ & $\eta \le \frac{1}{L}$

$$\ge g(x) + h(x) = f(x). \quad \forall x.$$

x_t $\quad$ x

Smoothness Lemma: $g(y) \le g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$

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

Rearranging,

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

$$[\text{Telescope}] \quad \leq \|x_0 - x^*\|^2 - 2\eta \sum_{s=1}^{t+1} \left[f(x_s) - f(x^*)\right].$$

$$\frac{1}{t+1} \sum_{s=1}^{t+1} \left[f(x_s) - f(x^*)\right] \leq \frac{\|x_0 - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta(t+1)}.$$

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

When the non-smooth Component is Simple, we can get $O(\frac{1}{T})$ Convergence rate. [ISTA]
(Iterative Shrinkage Thresholding Algorithm

We can infact improve this to $O(\frac{1}{T^2})$ using Nesterov's AGD. [FISTA]

STOCHASTIC ALGORITHMS

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

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

Oracle : (a_i, b_i)

$$\hat{\nabla} f(x) \triangleq (a_i^T x - b_i) a_i \implies \mathbb{E}[\hat{\nabla} f(x)] = \underline{\nabla f(x)}.$$

② Empirical Risk Minimization (ERM)

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

If $a_i \in \mathbb{R}^d$

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

$O(d)$ time $\longleftarrow \hat{\nabla} f(x) \triangleq (a_i^T x - b_i) a_i$ where $i \sim \mathrm{Unif}[1, \ldots, n]$.

Setting: $x \longrightarrow \boxed{\text{oracle}} \longrightarrow \hat{\nabla} f(x)$

$\hat{\nabla} f(x)$ independent of everything else $\left(\mathbb{E}[\hat{\nabla} f(x)] = \nabla f(x) \right.$; $\mathbb{E}\left[\| \hat{\nabla} f(x) - \nabla f(x) \|^2 \right] \le \sigma^2$.

$\|\nabla f(x)\| \leq G$

f : Convex & G-Lipschitz
Random noise has bounded variance

SGD [Robbins & Monro]

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

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

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

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

1) $\mathbb{E}\left[\langle \hat{\nabla} f(x_t), x_t - x^* \rangle \middle| x_t\right] = \langle \mathbb{E}\left[\hat{\nabla} f(x_t) \middle| x_t\right] x_t - x^* \rangle$

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

[Convexity] $\geq f(x_t) - f(x^*)$.

2) $\mathbb{E}\left[\langle \hat{\nabla} f(x_t), \hat{\nabla} f(x_t) \rangle \middle| x_t\right] = \mathbb{E}\left[\langle \nabla f(x_t) + n_t, \nabla f(x_t) + n_t \rangle\right]$

$n_t \triangleq \hat{\nabla} f(x_t) - \nabla f(x_t)$

$$= \mathbb{E}\left[\underbrace{\|\nabla f(x_t)\|^2}\right] + \mathbb{E}\left[\|n_t\|^2\right]$$

$$\leq G^2 + \sigma^2$$

Since $f(\cdot)$ is G-Lipschitz

Since noise variance $\leq \sigma^2$

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

$$- 2\eta\,[\mathbb{E}[f(x_t)] - f(x^*)]$$

$$+ \eta^2 (G^2 + \sigma^2).$$

Using the above inequality iteratively

$$\leq \mathbb{E}[\|x_0 - x^*\|^2] - 2\eta \sum_{s=0}^{t} [\mathbb{E}[f(x_s)] - f(x^*)]$$

$$+ \eta^2 (t+1)(G^2 + \sigma^2).$$

$$\frac{1}{t+1} \sum_{s=0}^{t} \mathbb{E}[f(x_s)] - f(x^*) \leq \frac{1}{2\eta(t+1)} [\|x_0 - x^*\|^2 - \mathbb{E}[\|x_{t+1} - x^*\|^2]]$$

$$+ \frac{\eta}{2} (G^2+\sigma^2).$$

Choose $\eta = \sqrt{\frac{\|x_0 - x^*\|^2}{(t+1)(G^2+\sigma^2)}}$

function parameter

Stochastic oracle

Av. expected Suboptimality $\leq \frac{\sqrt{(G^2+\sigma^2)} \cdot \|x_0 - x^*\|}{\sqrt{t+1}}$.

For nonsmooth : this is the best possible rate $\left(\frac{1}{\sqrt{t}}\right)$

For Smooth : Cannot improve on $\frac{1}{\sqrt{t}}$; but can obtain $\frac{L\|x_0 - x^*\|^2}{t^2} + \frac{\sigma\|x_0 - x^*\|}{\sqrt{t}}$.

Exercise: SGD for L-Smooth $f(\cdot)$ gets rate of

$$O\left(\frac{L \|x_0 - x^*\|^2}{t} + \frac{\sigma \|x_0 - x^*\|}{\sqrt{t}} \right).$$

Exercise: If f is G-Lipschitz and μ-Strongly convex, then SGD [with diff. η] gets rate of

$$O\left(\frac{G^2 + \sigma^2}{\mu t} \right).$$

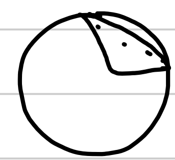
LECTURE - 6

Note Title 27-Jun-19

MIRROR DESCENT

$$\min_{x} \underbrace{\frac{1}{2}\|Ax - b\|^2}_{= f(x)}$$

$$\mathcal{X} \triangleq \boxed{\begin{array}{c} x_i \geq 0 \\ \sum_i x_i = 1 \end{array}}$$



PGD: $x_{t+1} = P_{\mathcal{X}}\left(x_t - \eta \nabla f(x_t)\right)$ | $O\left(\frac{\|A^TA\|\sqrt{2}}{t}\right)$

$L = \|A^TA\|$; $\|x_0 - x^*\| \leq \text{Diam}(\mathcal{X})$

$\leq \sqrt{2}$

Smoothness

each Column of A : $\|A_i\| \leq 1$ $\quad$ $[\,|\,|\,|\,|\,|\,]$

$\|A^T A\| = \|A\|^2$ Could be $\approx \underline{\underline{d}}$. $\quad$ unit norm columns.

When $\mathcal{X}$ and $f(\cdot)$ have good properties w.r.t. $\|\cdot\|$ which is not Euclidean, Can we do better?

$$x \longrightarrow \|\cdot\| \qquad\qquad \nabla f \longrightarrow \|\cdot\|_*$$

e.g., for probabilities $\|\cdot\|_2 \longmapsto \|\cdot\|_\infty$

<u>Defn. of dual norm</u>: Given a norm $\|\cdot\|$ on $\mathcal{X}$, the dual norm on the space of linear operators $\mathcal{L}(\mathcal{X})$ is defined as:

$$\|l\|_* = \sup_{\substack{\|x\| \leq 1 \\ x \in \mathcal{X}}} l(x) \quad \forall\, l \in \mathcal{L}(\mathcal{X}).$$

$$\|x\| \triangleq \|\underbrace{x}_{\in\mathbb{R}^d}\|_1 = \sum_i |x_i| \xrightarrow{\text{Dual norm}} \|\underbrace{y}_{\in\mathbb{R}^d}\|_* = \|y\|_\infty = \max_i |y_i|$$

$$\|y\|_* = \sup_{\|x\|_1 \le 1} \sum_i x_i y_i \qquad x_{i^*} = \text{Sign}(y_{i^*}) \text{ on } i^* = \arg\max_i |y_i|$$

$$= 0 \quad \text{o.w.}$$

$$= \|y\|_\infty .$$

$\|\cdot\|$ on x , $\|\cdot\|_*$ on $\nabla f(x)$

$\downarrow$ $\qquad\qquad$ $\downarrow$

$\|\cdot\|_1$ $\qquad\qquad$ $\|\cdot\|_\infty$

$$f(y) \ge f(x) + \langle \underbrace{\nabla f(x)}_{\nabla f(x)}, \underbrace{y-x}_{\in x} \rangle$$

$$\boxed{f\text{-Convex} \ \& \ \|\nabla f(x)\|_* \leq G.}$$ $\longrightarrow$ Lipschitz

$\|\cdot\|$

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

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

1. Not clear how to solve this step $\left(\sum_i |x_i - x_{t_i}|\right)^2$
2. Not clear what the potential function is.

Defn. Bregman divergence: Given a convex function $\phi : \mathcal{X} \to \mathbb{R}$,

$$D_\phi(x ; y) \triangleq \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle \geq 0$$

$\because \phi$ is convex.

Mirror descent:

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

Requirement: ϕ is 1-strongly convex w.r.t. $\|\cdot\|$ on $\mathcal{X}$.

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

ASSUMPTIONS

① f is Convex

② $\|\nabla f(x)\|_* \leq G$

③ $\phi(\cdot)$ is 1-Strongly Convex in $\|\cdot\|$ on $\mathcal{X}$.

$D_\phi(x^*; x_t)$

$x \in \mathcal{X}$ and $l \in \mathcal{L}(\mathcal{X})$
$= \nabla f(x)$.

Hölder's ineq.: $\langle l, x \rangle \leq \|l\|_* \cdot \|x\|$

$$x_{t+1} = \arg\min_x \underbrace{f(x_t)}_{\text{Const.}} + \underbrace{\langle \nabla f(x_t), x - x_t \rangle}_{\text{Linear in } x} + \underbrace{\frac{1}{2\eta} D_\phi(x; x_t)}_{\text{Convex in } x.}$$

$$\nabla f(x_t) + \frac{1}{2\eta} \nabla_x \underbrace{D_\phi(x ; x_t)}_{\phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x - x_t \rangle} = 0.$$

$$\nabla f(x_t) + \frac{1}{2\eta} \left[\nabla\phi(x_{t+1}) - 0 - \nabla\phi(x_t) \right] = 0.$$

$$x_{t+1} : \underbrace{\nabla\phi(x_{t+1})}_{= x_{t+1}} = \underbrace{\nabla\phi(x_t)}_{x_t} - 2\eta \nabla f(x_t).$$

] MD Step.

$\longrightarrow$ 1

By optimality of x_{t+1},

$$f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2\eta} D_\phi(x_{t+1}; x_t) \leq f(x_t) + \langle \nabla f(x_t), x_t - x_t \rangle + \frac{1}{2\eta} D_\phi(x_t; x_t)$$

$$= f(x_t)$$

$$\frac{1}{4\eta} \|x_t - x_{t+1}\|^2 \overset{\text{1- Str. Conv. of } \phi}{\leq} \frac{1}{2\eta} D_\phi(x_{t+1}; x_t) \leq - \langle \nabla f(x_t), x_{t+1} - x_t \rangle .$$

$$\overset{\text{Hölder's}}{\leq} \|\nabla f(x_t)\|_* \cdot \|x_{t+1} - x_t\| .$$

$$\|x_t - x_{t+1}\| \leq 4\eta \|\nabla f(x_t)\|_* \leq 4\eta G \longrightarrow 2$$

$$
\begin{aligned}
D_\phi(x; x_{t+1}) &= \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_{t+1}), x - x_{t+1} \rangle \\
&\overset{①}{=} \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_t) - 2\eta\nabla f(x_t), x - x_{t+1} \rangle \\
&\overset{\text{Conv. of } \phi}{\leq} \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x_{t+1} - x_t \rangle \\
&\qquad - \langle \nabla\phi(x_t) - 2\eta\nabla f(x_t), x - x_{t+1} \rangle \\
&= \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x - x_t \rangle \\
&\qquad - 2\eta \langle \nabla f(x_t), x_{t+1} - x \rangle \\
&= D_\phi(x; x_t) - 2\eta \langle \nabla f(x_t), x_t - x \rangle
\end{aligned}
$$

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

$$\overset{\text{Hölder's}}{\leq} D_\phi(x; x_t) - 2\eta \overbrace{\langle \nabla f(x_t), x_t - x \rangle}^{\text{Convexity}}$$

$$+ 2\eta \| \nabla f(x_t) \|_* \cdot \underbrace{\| x_{t+1} - x_t \|}_{2}$$

$$\leq D_\phi(x; x_t) - 2\eta \left[f(x_t) - f(x) \right]$$

$$+ 2\eta \underbrace{\| \nabla f(x_t) \|_*}_{\leq G} \cdot 4\eta G .$$

$$\leq D_\phi(x; x_t) - 2\eta \left[f(x_t) - f(x) \right] + 8\eta^2 G^2 .$$

$$D_\phi(x; x_{t+1}) \leq D_\phi(x; x_t) - 2\eta[f(x_t) - f(x)] + 8\eta^2 G^2.$$

Telescoping and averaging gives us

$$\frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x)] \leq \frac{D_\phi(x; x_0) - D_\phi(x; x_T)}{2\eta T} + 4\eta G^2.$$

$$\leq \frac{D_\phi(x; x_0)}{2\eta T} + 4\eta G^2.$$

$$\eta = \frac{1}{G}\sqrt{\frac{D_\phi(x; x_0)}{8T}}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[f(x_t)-f(x)\right] \leq C\cdot \frac{G\sqrt{D_\phi(x;x_0)}}{\sqrt{T}}.$$

$$\leq C \cdot \frac{G\,\|x-x_0\|}{\sqrt{T}}$$

(1) If $\|\cdot\| = \|\cdot\|_2$ ↳ Euclidean, then choose $\phi = \frac{1}{2}\|x\|_2^2$ → 1-str. cvx. w.r.t. $\|\cdot\|_2$

$$D_\phi(x;x_t) = \frac{1}{2}\|x-x_t\|^2.$$

In this case a) Mirror descent = GD

b) Obtained bounds are the same.

(2) $\|\cdot\| = \|\cdot\|_1$ and $\mathcal{X} = \{ x : x_i \geq 0 ; \sum x_i = 1 \}$.

$$\phi(x) = \sum_i x_i \log x_i$$

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

$$= \sum_i x_i \log x_i - \sum_i y_i \log y_i - \sum_i (1+\log y_i)(x_i - y_i)$$

$$= \sum_i x_i \log \frac{x_i}{y_i} + \underbrace{\sum_i (x_i - y_i)}_{=0}$$

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

ASIDE

$x_i \geq 0$: $\phi(x) = \sum_i x_i \log x_i - \sum_i x_i$ $\quad [\sum_i x_i \neq 1].$

$$\sum_i x_i = 1,\ x_i \geq 0. \longrightarrow \phi(x) = \sum_i x_i \log x_i$$

$$D_\phi(x; y) = KL(x \| y).$$

① ϕ [or eq. D_ϕ] is 1-strongly convex

<u>Pinsker's inequality</u>: $KL(x \| y) \geq 2 \|x-y\|_1^2$

② $$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \underbrace{f(x_t)}_{\text{Const.}} + \langle \nabla f(x_t), x - \underbrace{x_t}_{\text{Const.}} \rangle + \frac{1}{2\eta} D_\phi(x ; x_t)$$

$$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \langle \nabla f(x_t), x \rangle + \frac{1}{2\eta} \sum_i x_i \log \frac{x_i}{(x_t)_i}$$

$$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_i x_i \left[\nabla f(x_t)_i + \underbrace{\frac{1}{2\eta}} \log \frac{x_i}{(x_t)_i} \right].$$

$$= \operatorname*{argmin}_{x \in \mathcal{X}} \sum_i x_i \log \frac{x_i}{(x_t)_i \exp\left(-2\eta(\nabla f(x_t))_i\right)}$$

$$Z_i \triangleq \frac{(x_t)_i \exp\left(-2\eta\, (\nabla f(x_t))_i\right)}{\sum_{j=1}^{d} (x_t)_j \exp\left(-2\eta(\nabla f(x_t))_j\right)}$$

$$= \operatorname*{argmin}_{x \in \mathcal{X}} \sum_i x_i \log \frac{x_i}{Z_i} - \underbrace{\underbrace{\sum_i x_i}_{=1} \underbrace{\log\left(\sum_{j=1}^{d} (x_t)_j \exp\left(-2\eta(\nabla f(x_t))_j\right)\right)}_{\text{Constant}}}_{\text{Constant}}$$

$$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \; KL(x \| z) = z.$$

Ind. discovered

$$x_{t+1} = z = \frac{x_t \odot \exp(-2\eta \nabla f(x_t))}{\| x_t \odot \exp(-2\eta \nabla f(x_t)) \|_1}.$$

MULTIPLICATIVE WEIGHTS UPDATE

EXPONENTIATED GRADIENT ALGORITHM

$$\min_{x:\ x_i \geq 0,\ \sum x_i = 1} \frac{1}{2} \cdot \|Ax - b\|^2.$$

$\|\cdot\|_1$ on $\mathcal{X}$ $\longrightarrow$ $\|\cdot\|_\infty$ on $\nabla f(x)$.

$$\phi(x) = \sum_i x_i \log x_i$$

$$\text{Av. Suboptimality} \leq C \cdot \frac{\|\nabla f(x)\|_\infty \sqrt{D_\phi(x\,;\,x_0)}}{\sqrt{T}}$$

$$\nabla f(x) = A^T(Ax - b) \longrightarrow \|A^T(Ax-b)\|_\infty \leq \|A^T A x\|_\infty + \|A^T b\|_\infty$$

$$\leq \max_i \|(A^T A)_i\|_\infty + \underline{\|A^T b\|_\infty}$$

Hölder's inequality

i^{th} row of $A^T A$

Non- Euclidean		Euclidean
$\|A^TA\|_\infty$	$\leq$	$\|A^TA\|_2$ Could be d times larger.
$\downarrow$ largest element in A^TA	

$$G \doteq \|A^TA\|_\infty + \|A^Tb\|_\infty \leq \|A^TA\|_2 + \|A^Tb\|_2 .$$

$D_\phi(x; x_0)$: $x_0 = \frac{1}{d} \cdot \mathbb{1}$

$$= \sum_i x_i \log \frac{x_i}{(x_0)_i} = \underbrace{\sum_i x_i \log x_i}_{\leq 0} + \underbrace{\sum_i x_i \log d}_{= \log d}$$

$$\text{Av. Suboptimality} \leq C \cdot \frac{\left(\|A^TA\|_\infty + \|A^Tb\|_\infty\right) \cdot \sqrt{\log d}}{\sqrt{T}} \cdot$$

Could be $\Omega(d)$

$\sqrt{\log d} \cdot$

$$C \cdot \frac{\left(\|A^TA\|_2 + \|A^Tb\|_2\right) \cdot 1}{\sqrt{T}} \cdot$$

Note Title 28-Jun-19

LECTURE - 7

Stochastic Variance Reduced Gradient (SVRG)

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

E.g., $f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2$

Each f_i is L-Smooth & Convex $\quad L = \max_i \|a_i\|^2$

f is μ-Strongly convex. $\quad \mu = \sigma_{\min}\left(\frac{1}{n}\sum_i a_i a_i^T\right)$

GD : $O\Big(\underbrace{n \cdot d}_{\text{One } \nabla f \text{ Computation}} \cdot \underbrace{\frac{L}{\mu} \log\frac{1}{\epsilon}}_{\# \text{iterations}}\Big)$ Computations.

$$\text{SGD: } O\Big(\underbrace{d}_{\text{One } \nabla f_i \text{ computation.}} \cdot \underbrace{\frac{\sigma \cdot \|x_0 - x^*\|}{\epsilon^2}}_{\text{\# iterations}}\Big)$$

σ → Std. dev. in the Stochastic gradients

In most cases: $\sigma \ll n \cdot \frac{L}{\mu}$

	Dep. on prob. param.	Desired acc.
SGD	Usually better	
GD		Usually better

→ Best of both worlds?

The slow convergence rate of SGD $\left(\frac{1}{\sqrt{t}}\right)$ is due to large variance in the stochastic gradients even close to the optimal point.

$$x^* = \arg\min_x f(x) \quad \text{then} \quad \nabla f(x^*) = 0 \;\left(= \frac{1}{n}\sum_i \nabla f_i(x^*)\right)$$

<u>but</u> $\nabla f_i(x^*) \neq 0$ for most i.

Variance reduction step

Pick x_0

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

Do $t = 1, \dots T$

Pick 'i' uniformly at random from $1, \cdots n$

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

Suppose $x_0 = x^*$

$$x_1 = x_0 - \eta \left[\nabla f_i(x_0) - \nabla f_i(x_0) + \nabla f(x_0) \right] = x_0 = x^*$$

Contrast with SGD step: $x_t = x_{t-1} - \eta \nabla f_i(x_{t-1})$.

$$\mathbb{E}_i \left[\nabla f_i(x_0) - \nabla f(x_0) \right] = \mathbb{E}_i \left[\nabla f_i(x_0) \right] - \nabla f(x_0) = \nabla f(x_0) - \nabla f(x_0) = 0.$$

Recall that the convergence guarantee of SGD depends on the variance in stochastic gradients.

$$\sigma^2 \triangleq \mathbb{E}\Big[\big\|\underbrace{\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0)}_{\text{Stochastic gradient}} - \underbrace{\nabla f(x_t)}_{\substack{\text{True gradient} \\ = \mathbb{E}[\text{stoch. grad.}]}}\big\|^2\Big]$$

$$\leq 2\,\mathbb{E}\big[\|\nabla f_i(x_t) - \nabla f_i(x_0)\|^2\big] + 2\,\mathbb{E}\big[\|\nabla f(x_0) - \nabla f(x_t)\|^2\big] \quad \Big[\because (a+b)^2 \leq 2a^2 + 2b^2\Big]$$

$$\leq 4\,\mathbb{E}\big[\|\nabla f_i(x_t) - \nabla f_i(x_0)\|^2\big] \quad \text{Jensen's ineq.}$$

$$\Big\{\ \mathbb{E}\big[\|w\|^2\big] \geq \|\mathbb{E}[w]\|^2 \ \Big\}$$

Jensen's inequality

$x^* = \arg\min_x f(x)$

$\left((a+b)^2 \le 2a^2 + 2b^2\right)$

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

$$\mathbb{E}_i\left[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2\right] \le \underline{\frac{2L}{3}\left[f(x) - f(x^*)\right].}$$

<u>Goal</u>: Bound

<u>Approach</u>:
$$\underbrace{g_i(x)}_{\substack{\text{L-Smooth}\\ \text{Convex}}} = \underbrace{f_i(x)}_{\substack{\text{L-Smooth}\\ \text{Convex}}} - \underbrace{\langle \nabla f_i(x^*), x \rangle}_{\text{Linear}}$$

$$\nabla g_i(x^*) = \nabla f_i(x^*) - \nabla f_i(x^*) = 0$$

$$\Rightarrow \quad g_i(x^*) \overset{\text{Conv. of } g_i}{\leq} g_i\left(\tilde{x} - \eta \nabla g_i(\tilde{x})\right)$$

$$\overset{\text{L-smooth } g_i}{\leq} g_i(\tilde{x}) - \underbrace{\eta\left(2 - \frac{\eta L}{2}\right)}\underbrace{\|\nabla g_i(\tilde{x})\|^2}$$

$$\eta = \frac{1}{L} \longrightarrow \quad \leq g_i(\tilde{x}) - \frac{3}{2L}\|\nabla g_i(\tilde{x})\|^2$$

$$\|\nabla g_i(\tilde{x})\|^2 \leq \left[g_i(\tilde{x}) - g_i(x^*)\right]\frac{2L}{3}.$$

$$\|\nabla f_i(\tilde{x}) - \nabla f_i(x^*)\|^2 \leq \left[g_i(\tilde{x}) - g_i(x^*)\right]\frac{2L}{3}$$

$-\eta \nabla g_i(\tilde{x})$

$\tilde{x} - \eta \nabla g_i(\tilde{x}) \quad \tilde{x}$

x^*

$= \underset{x}{\operatorname{argmin}}\ g_i(x)$

Taking an average,

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\tilde{x}) - \nabla f_i(x^*)\|^2 \leq \frac{1}{n}\cdot\frac{2L}{3}\sum_{i=1}^{n}\Big[\underbrace{g_i(\tilde{x})}_{f_i(\tilde{x}) - \langle \nabla f_i(x^*), \tilde{x}\rangle} - g_i(x^*)\Big]$$

$$= \frac{2L}{3n}\sum_{i=1}^{n} f_i(\tilde{x}) - f_i(x^*) - \langle \nabla f_i(x^*), \tilde{x} - x^*\rangle$$

$$= \frac{2L}{3}\Big\{\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{x}) - \frac{1}{n}\sum_{i=1}^{n} f_i(x^*)$$

$$- \Big\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^*), \tilde{x} - x^*\Big\rangle$$

$$= \frac{2L}{3}\left\{ f(\tilde{x}) - f(x^*) - \langle \underbrace{\nabla f(x^*)}_{\substack{\| \\ 0}}, \tilde{x} - x^* \rangle \right\}$$

Lemma: $\mathbb{E}\left[\| \nabla f_i(\tilde{x}) - \nabla f_i(x^*) \|^2 \right] \leq \frac{2L}{3}\left[f(\tilde{x}) - f(x^*) \right].$

Furthermore,

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

[2nd & 3rd page computation]

$$\leq 8\left\{ \mathbb{E}\left[\| \nabla f_i(x_t) - \nabla f_i(x^*) \|^2 \right] + \mathbb{E}\left[\| \nabla f_i(x_0) - \nabla f_i(x^*) \|^2 \right] \right\}.$$

$$\leq \frac{16L}{3}\left[f(x_t) - f(x^*) + f(x_0) - f(x^*) \right].$$

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

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

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

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

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

$$+\eta^2 \cdot \frac{16L}{3} \underbrace{\left[f(x_t) - f(x^*) + f(x_0) - f(x^*) \right]}$$

$$\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] \le \mathbb{E}\left[\|x_t - x^*\|^2\right] - \eta\left(2 - \frac{19\eta L}{3}\right)\left(f(x_t) - f(x^*)\right)$$

$$+ \frac{16\eta^2 L}{3}\left(f(x_0) - f(x^*)\right).$$

Telescoping and averaging,

$$\frac{1}{T+1}\sum_{t=0}^{T}\left[\mathbb{E}[f(x_t)] - f(x^*)\right] \le \frac{\overbrace{\|x_0 - x^*\|^2}}{\eta\left(2 - \frac{19\eta L}{3}\right)(T+1)} + \frac{16\eta L}{3\left(2 - \frac{19\eta L}{3}\right)} \left[f(x_0) - f(x^*)\right].$$

f is μ- str. cvx. $\therefore$ $f(x_0) - f(x^*) \geq \frac{\mu}{2} \underbrace{\|x_0 - x^*\|^2}.$

$$\text{Av. exp. Suboptimality} \leq \left[\underbrace{\frac{2}{\mu\eta\underbrace{\left(2 - \frac{19\eta L}{3}\right)}(T+1)}}_{\frac{L}{\mu} = \kappa} + \underbrace{\frac{16\eta L}{3\left(2 - \frac{19\eta L}{3}\right)}}_{\leq \frac{1}{10}} \right] \underbrace{\{ f(x_0) - f(x^*) \}}_{\text{Initial Suboptimality}}$$

$$\eta = \frac{1}{100 \cdot L}$$

$$\approx \left[\frac{2\kappa}{T+1} + \frac{1}{10} \right] [\text{Initial Suboptimality}].$$

$$T > 6 \cdot \kappa \implies \underline{\text{Av. exp. Suboptimality} \leq \tfrac{1}{2}\ \text{Initial Subopt.}}$$

Pick x_0

For epochs $= 1, \dots \log(\frac{1}{\epsilon})$.

SVRG epoch

Compute $\nabla f(x_0) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_0) \leftarrow \boxed{O(nd)}$

For $t = 0, \dots T = 5K$ $\rightarrow O(Kd)$

— Pick i u.a.r. from $1, \dots n$.

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

Reset $x_0 = x_1, \dots x_{T+1}$ uniformly at random.

$$O\left(\underbrace{(nd + Kd)}\ \underbrace{\log\frac{1}{\epsilon}}\right)$$

$$f_i(x) = (a_i^T x - b_i)^2$$

$$O\Big((nd + \underset{\downarrow}{\kappa} d) \cdot \log \tfrac{1}{\epsilon} \Big)$$

$$\frac{\max_i \|a_i\|^2}{\mu}$$

Importance Sampling

i w.p. $p_i \propto \|a_i\|^2$.

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

$$p_i = \frac{\|a_i\|^2}{\sum_{j=1}^{n} \|a_j\|^2}$$

$$\text{Smoothness}\left(\frac{1}{np_i} (a_i^T x - b_i)^2 \right) = \frac{\sum_{j=1}^{n} \|a_j\|^2}{n}$$

Uniform sampling

$$\kappa_{us} = \frac{\max_i \|a_i\|^2}{\mu}$$

Weighted Sampling

$$\kappa_{ws} = \frac{\sum_{j=1}^{n} \|a_j\|^2}{n\mu}$$

$$\kappa_{ws} \leq \kappa_{us}$$

Note Title 28-Jun-19

LECTURE-8

CONVEX – CONCAVE MINIMAX OPTIMIZATION

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

$f(\cdot, y)$ is Convex

$f(x, \cdot)$ is Concave

g is Concave

$\triangleq -g$ is Convex

1) $g(\alpha y_1 + (1-\alpha) y_2) \geq \alpha g(y_1) + (1-\alpha) g(y_2)$

2) $g(z) \leq g(y) + \langle \nabla g(y), z - y \rangle$

$$\text{Find } x \ \text{ s.t. } \max_i f_i(x) \le 0$$

$\equiv$

1) Constrained Optimization : Find x

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

Each f_i is convex.

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

$$|||$$

$$\min_x \max_{\substack{y_i \ge 0 \\ \sum_{i=1}^m y_i = 1}} \sum_{i=1}^m y_i f_i(x)$$

$$f(x,y) = \sum_{i=1}^m y_i f_i(x)$$

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

$f(x, \cdot)$ is linear in $\cdot$ $\forall x$. $\longrightarrow$ concave

2) Many non-smooth functions can be written as smooth minimax problems.

a) $\|x\|_1 = \max_{-1 \le y_i \le 1} \sum_i y_i x_i$

$$\min_x \frac{1}{2}\|Ax-b\|^2 + \lambda \|x\|_1 = \min_x \max_{-1 \le y_i \le 1} \frac{1}{2}\|Ax-b\|^2 + \lambda \sum_i y_i x_i$$

b) Given $a_1, \dots a_m$ find the smallest ball covering $a_1, \dots a_m$.

$\longrightarrow$ Again can be written as smooth minimax problem.

3) Minimax $\equiv$ Zero sum games.

$f(\cdot, y)$ is convex $\forall y$
$f(x, \cdot)$ is concave $\forall x$.

$$\min_x \underbrace{\max_y f(x,y)}_{\triangleq g(x)}$$

Exercise: $g(x)$ is convex

How to evaluate $\nabla g(x)$?

(Informal) Danskin's thm: Under smoothness assumptions on $f(\cdot)$,

$$\nabla g(x) = \text{Conv. hull}\left\{ \nabla_x f(x,y) : y \in \arg\max_z f(x,z) \right\}.$$

Convex opt.

1) Find $y_t \in \arg\max_y f(x_t, y)$

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

Subgradient descent on $g(x)$.

Issues: ① Finding y_t takes time

② g could be non smooth $\Rightarrow$ slow convergence rate.

Non smooth : Gradient descent ascent $\longrightarrow O\left(\frac{1}{\sqrt{T}}\right)$

$g(x)$ could still be nonsmooth $\leftarrow$ Smooth : Mirror-Prox $\longrightarrow O\left(\frac{1}{T}\right)$.

$$\underline{\text{maximin} \leq \text{minimax}}$$

[f Could be arbitrary]

<u>Weak duality</u>: $\underbrace{\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x,y)}_{LHS} \leq \underbrace{\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x,y)}_{RHS}$.

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

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

$$LHS = \min_{x \in \mathcal{X}} f(x, y^*) \leq f(x^*, y^*) \leq \max_{y \in \mathcal{Y}} f(x^*, y) = RHS.$$

[Sion's minimax thm.]

Strong duality: If $f(x,y)$ is Convex-Concave and $\mathcal{X}$ and $\mathcal{Y}$ are Compact then $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x,y)$.

Defn. ϵ-primal dual pair: $(\bar{x},\bar{y})$ is Said to be an ϵ-primal dual pair if $\max_{y} f(\bar{x},y) - \min_{x} f(x,\bar{y}) \leq \epsilon$.

Note: $\max_{y} f(\bar{x},y) \geq f(\bar{x},\bar{y}) \geq \min_{x} f(x,\bar{y})$.

Let $(\bar{x}, \bar{y})$ be an ϵ-primal dual pair

Exercise: Show that $\bar{x}$ is an ϵ-optimal soln. for

$$\min_x g(x) \triangleq \max_y f(x,y)$$

Exercise: $\bar{y}$ is an ϵ-optimal soln. for $\max_y h(y) \triangleq \min_x f(x,y)$.

$\mathrm{diam}(\mathcal{X}), \mathrm{diam}(\mathcal{Y}) \leq R$

$f(\cdot, y)$ is convex

$f(x, \cdot)$ is concave

$\|\nabla_x f(x,y)\| \leq G$

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

<u>Gradient descent ascent</u>:

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t) \to \text{GD for min}$$

$$y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t) \to \text{GA for max}$$

<u>Thm</u>: GDA has conv. rate $O\left(\frac{GR}{\sqrt{T}}\right)$.

<u>Proof</u>:

x is arb.

$$\|x_{t+1} - x\|^2 = \|x_t - x\|^2 - 2\eta \underbrace{\langle \nabla_x f(x_t, y_t), x_t - x \rangle}_{\text{Conv. of } f(\cdot, y_t)} + \eta^2 \underbrace{\|\nabla_x f(x_t, y_t)\|^2}_{G\text{-Lipschitz}}$$

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

$\boxed{y \text{ is arb.}}$ $$\|y_{t+1}-y\|^2 = \|y_t - y\|^2 - 2\eta \underbrace{\langle \nabla_y f(x_t, y_t), y - y_t \rangle}_{\text{Concavity of } f(x_t,\cdot)} + \eta^2 \underbrace{\|\nabla_y f(x_t,y_t)\|^2}_{G\text{-Lipschitz}}$$

$$\leq \|y_t - y\|^2 - 2\eta\left[f(x_t, y) - f(x_t, y_t)\right] + \eta^2 G^2$$

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

$$\frac{1}{T+1}\sum_{t=0}^{T}\left[f(x_t,y) - f(x,y_t)\right] \leq \frac{\|x_0 - x\|^2 + \|y_0 - y\|^2}{2\eta(T+1)} + \eta G^2$$

$$f\Big(\underbrace{\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T} x_t}_{\bar{x}_T}, y\Big) \leq \frac{1}{T+1}\sum_{t=0}^{T} f(x_t, y) \quad \text{and} \quad f\Big(x, \underbrace{\tfrac{1}{T+1}\textstyle\sum_{t=0}^{T} y_t}_{\bar{y}_T}\Big) \geq \frac{1}{T+1}\sum_{t=0}^{T} f(x, y_t)$$

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

$$\max_y f(\bar{x}_T, y) - \min_x f(x, \bar{y}_T) \leq \frac{\max_x \|x_0 - x\|^2 + \max_y \|y_0 - y\|^2}{2\eta(T+1)} + \eta G^2 .$$

$$\leq \frac{R^2}{\eta(T+1)} + \eta G^2$$

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

$(\bar{x}_T, \bar{y}_T)$ is an $\frac{2GR}{\sqrt{T+1}}$ - primal dual pair. ∎

[Prox] Proximal algorithm : $x_{t+1} = \underset{x \in \mathcal{X}}{\arg\min} \left[f(x) + \frac{1}{2\eta} \|x - x_t\|^2 \right]$

GD step : $x_{t+1} = \underset{x \in \mathcal{X}}{\arg\min} \left[f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|^2 \right]$

Exercise: Prox has convergence rate $O\left(\frac{1}{T}\right)$ for [possibly non-smooth] convex functions.

Exercise: Prox step is efficiently implementable for

L-Smooth functions and $\eta \leq \frac{1}{2L}$.

Efficient : Can find ϵ- approximate Soln. to the prox step in $O(\log \frac{1}{\epsilon})$ iterations.

Prox step : Given x, find w^* s.t.

$$w^* = \arg\min_{z \in \mathcal{X}} f(z) + \frac{1}{2\eta} \|x - z\|^2$$

$$\equiv \quad w^* = x - \eta \nabla f(w^*) \longrightarrow \text{Prox.}$$

$\nabla f(x) \longrightarrow$ GD

Algorithm to find Prox Step]. Pick $w_0 = x$

$$w_{t+1} = x - \eta \nabla f(w_t) \longrightarrow (1)$$

(A) Claim: w^* is the unique fixed point of (1)

(B) Claim: (1) is a $\frac{1}{2}$ Contraction.

$$\left.\begin{array}{l} w \longrightarrow w^+ = x - \eta \nabla f(w) \\ \tilde{w} \longrightarrow \tilde{w}^+ = x - \eta \nabla f(\tilde{w}) \end{array}\right] \quad \begin{aligned} \|w^+ - \tilde{w}^+\| &= \eta \, \|\nabla f(w) - \nabla f(\tilde{w})\| \\ &\leq \eta L \, \|w - \tilde{w}\| \\ \text{If } \eta \leq \tfrac{1}{2L} \quad &\leq \tfrac{1}{2} \|w - \tilde{w}\|. \end{aligned}$$

(A) + (B) $\Longrightarrow$ (1) finds ϵ- approximate w^* in $\log \frac{1}{\epsilon}$ iterations.

Conceptual Mirror-Prox :

$$x_{t+1} = \underset{x \in \mathcal{X}}{\arg\min} \; f(x, y_{t+1}) + \frac{1}{2\eta} \|x - x_t\|^2$$

$$y_{t+1} = \underset{y \in \mathcal{Y}}{\arg\max} \; f(x_{t+1}, y) - \frac{1}{2\eta} \|y - y_t\|^2$$

Implementation of each step

$x_t^0 = x_t$

$y_t^0 = y_t$

$i = 1, \dots \underbrace{\log\frac{1}{\varepsilon}}$,

'2' in the actual mirrorProx alg.

$$x_t^i = \underset{x \in \mathcal{X}}{\arg\min} \; f(x, y_t^{i-1}) + \frac{1}{2\eta} \|x - x_t\|^2$$

$$y_t^i = \underset{y \in \mathcal{Y}}{\arg\max} \; f(x_t^{i-1}, y) - \frac{1}{2\eta} \|y - y_t\|^2.$$

Exercise: For Conceptual Mirror-Prox, obtain $O(\frac{1}{T})$ Convergence rate for L-Smooth minimax opt.

Exercise: Show that the algorithm for implementing the Prox step Converges in $O(\log\frac{1}{\epsilon})$ iterations.

Exercise: Careful accouting of all the terms to show that 2 inner steps suffice for $O(\frac{1}{T})$ Convergence rate.