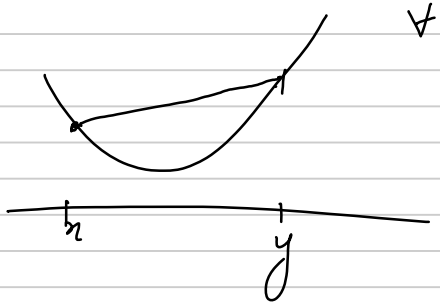
LEC - 2

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

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.



Ⓓ

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

Lemma: If f is convex then $\forall\ x$ in the domain, $\exists$ non-empty set $G(x)$ (Subgradients) s.t. $f(y) \geq f(x) + \langle z, y - x \rangle$ $\forall\ x, y$ and $z \in G(x)$.

MORALLY: CONVEX = CONVEX.

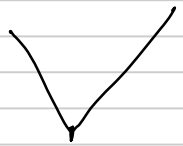
(First defn. = Second defn.)

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

$f(x) = |x|$

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

does not exist at $x = 0$.



$f(\cdot)$ is convex $\quad | \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

$\nabla f(x)$ ↓ Subgradient

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

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

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

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

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

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

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

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

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

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

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

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): $x_{t+1} \in \text{Span}\{x_0, \nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_t)\}$.

$GD \in$ Gradient span algorithms.

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

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

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

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

o) f is convex

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

Remark: Matches upper bound up to constant factors. [GD is optimal for this class of functions]

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

x $\begin{bmatrix} \\ \\ \\ \end{bmatrix}$ dimension $= T+1$

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

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

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

$$f(x) = \max_i f_i(x)$$

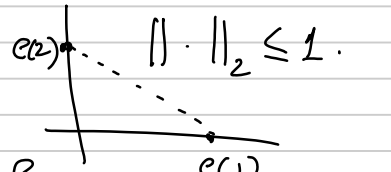
Let $I(x) = \{ i : f_i(x) = f(x) \}$.

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

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

$$\|\nabla f(x)\| \le \| \qquad \| + \mu \|x\|.$$

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



If $\|x\| \le R$

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

$x(i) = \alpha$

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

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

① $\sum + \mu R \le G$

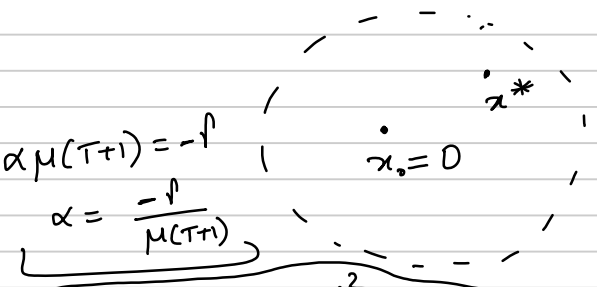
$x_0 = 0$

② $\|x_0 - x^*\|^2 = \|x^*\|^2$

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

① $\sum + \mu R \le G$

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



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

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

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

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

$x_0 = 0$.

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

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

$x_0 = 0$

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

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

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

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

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

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

$f_1(x) \Rightarrow f_t(x)$

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

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

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

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

If $\alpha < 0$ then $I(x_1) = 2 \longrightarrow e_2$

$x_t(t+1) = 0$

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

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

$x_t(T+1) = 0$.

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

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

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

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

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

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