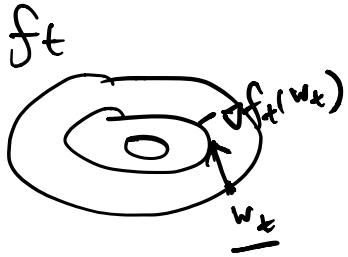
Online (sub) gradient descent algorithm:



Initialize $w_1 \in \Omega$. parameter η.

For $t = 1, 2, \cdots T$:

- choose w_t.
- Receive f_t. suffer loss $f_t(w_t)$
- set $g_t \in \partial f_t(w_t)$
- update: $w'_{t+1} \leftarrow w_t - \eta g_t$ $(\eta > 0)$

Ω $\quad w_{t+1} \leftarrow \Pi_\Omega(w'_{t+1})$

$= \operatorname{argmin}_{w \in \Omega} \| w - w'_{t+1} \|_2$.

$\|a+b\|_2^2 = \|a\|_2^2 + \|b\|^2 + 2\langle a, b \rangle$

Remark:

$$w_{t+1} = \operatorname*{argmin}_{w \in \Omega} \| w - w_t + \eta g_t \|_2^2$$

$D_\psi(w, w_t)$

$$= \operatorname*{argmin}_{w \in \Omega} \langle w, \eta g_t \rangle + \frac{1}{2} \| w - w_t \|_2^2$$

corrective-ness conservative-ness

(Kivinen & Warmuth '97).

OGD guarantees:

Thm: OGD w/ initializer w_1 & step size $\eta > 0$

guarantees: $\forall u \in \Omega$

$$R_T(u) \le \frac{\|u - w_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2 .$$

Moreover, if Ω has ℓ_2-diameter B ($\forall u, v \in \Omega$, $\|u - v\|_2 \le B$), and $\|g_t\|_2 \le \rho$. (which happens if all f_t's are ρ-Lipschitz), then

$$R_T(\Omega) \le \frac{B^2}{2\eta} + \frac{\eta}{2}\cdot T \cdot \rho^2$$

$\eta = \frac{B}{\rho}\cdot\sqrt{\frac{1}{T}}$ → $\sqrt{T}\cdot B^2\rho$ → $\sqrt{T}\cdot\rho$ → $\sqrt{T}\cdot\rho\cdot(B^2+1)$

$B \cdot \rho \cdot \sqrt{T}$

Cor: under the above setting, $\ell(w, z)$ is ρ-Lip wrt w; $0 \in D$, with $f_t(w) = \ell(w, z_t)$ for $z_1, \dots z_T \overset{iid}{\sim} D$, guarantees that: $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$,

① $\eta = \frac{B}{\rho}\sqrt{\frac{1}{T}} \Rightarrow \mathbb{E}\, L_D(\bar{w}_T) \le \min_{w \in \Omega} L_D(w) + \frac{B\rho}{\sqrt{T}}$.

② $\eta = \frac{1}{\rho}\sqrt{\frac{1}{T}}$, $\Omega = \mathbb{R}^d$. $w_1 = 0$. B^2

$\Rightarrow \mathbb{E}\, L_D(\bar{w}_T) \le L_D(w^*) + \frac{(\|w^*\|^2+1)\rho}{\sqrt{T}}$ $\forall w^* \in \mathbb{R}^d$.

Pf of corollary:

high prob. upper bound on $L_D(\bar{w}_T) \le L_D(w^*) + \frac{R_T(w^*)}{T}$ + concentration

$$\frac{1}{T}\sum_{t=1}^{T} l_t(w_t, z_t) - \frac{1}{T}\sum_{t=1}^{T} l_t(w^*, z_t) \le R_T(w^*).$$

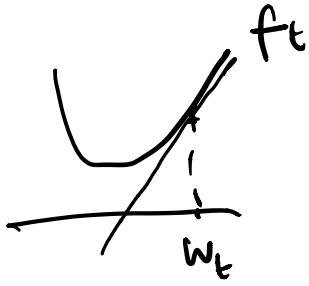
take expectation on LHS:

$$\mathbb{E}[\text{LHS}] = \mathbb{E}\left[\sum_{t=1}^{T} L_D(w_t)\right] - T \cdot L_D(w^*). \quad (*)$$

pf of OGD guarantees:

step 1: "linearization".

$$R_T(u) = \sum_{t=1}^{T}\left(f_t(w_t) - f_t(u)\right)$$



$f_t(u) \ge f_t(w_t) + \langle g_t, u - w_t \rangle$

$$\le \sum_{t=1}^{T} \langle g_t, w_t - u \rangle.$$

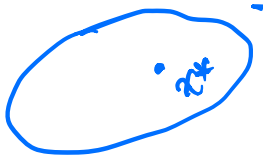
step 2: use optimality condition on w_{t+1}:

First order optimality condition:

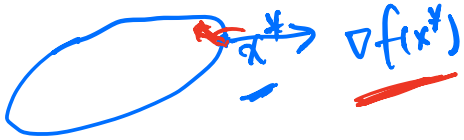
f is cvx. in cvx domain Ω, f differentiable.

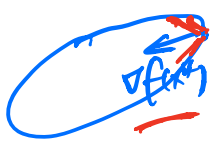
$x^* = \operatorname{argmin}_{x \in \Omega} f(x) \iff \forall y \in \Omega, \langle \nabla f(x^*), y - x^* \rangle \ge 0.$

① x^* in interior of Ω



$\nabla f(x^*) = 0.$

② 



we need $\forall y \in \Omega$

$\langle \nabla f(x^*), y - x^* \rangle \ge 0.$

Pf idea: ($\Rightarrow$) if $\exists y,\ \langle \nabla f(x^*), y-x^* \rangle < 0$.

$$f(x^* + \alpha(y-x^*)) = f(x^*) + \alpha\langle \nabla f(x^*), y-x^*\rangle + o(\alpha)$$

$$< f(x^*) \quad \text{for small } \alpha > 0.$$

($\Leftarrow$) $\forall y.\ f(y) \geq f(x^*) + \underbrace{\langle \nabla f(x^*), y-x^*\rangle}_{\geq 0}$.

$$w_{t+1} = \underset{w\in\Omega}{\operatorname{argmin}} \underbrace{\langle \eta g_t, w\rangle + \frac{1}{2}\|w-w_t\|^2}_{f(w)}$$

First order optimality:

$$\langle \underbrace{\eta g_t + w_{t+1} - w_t}_{a},\ \underbrace{u - w_{t+1}}_{b}\rangle \geq 0 \quad \forall u\in\Omega.$$

$$\Rightarrow \langle g_t, w_{t+1} - u\rangle \leq \frac{1}{\eta}\langle w_{t+1}-w_t,\ u-w_{t+1}\rangle.$$

$$\langle a, b\rangle = \frac{\|a+b\|^2 - \|a\|^2 - \|b\|^2}{2}$$

$$= \frac{1}{2\eta}\left(\|u-w_t\|^2 - \|u-w_{t+1}\|^2 - \|w_{t+1}-w_t\|^2\right)$$

step 3: bounding $\langle g_t, w_t - u\rangle$.

$$\langle g_t, w_t - u\rangle = \langle g_t, w_{t+1} - u\rangle + \underbrace{\langle g_t, w_t - w_{t+1}\rangle}_{\|g_t\|\,\|w_t - w_{t+1}\|}$$

Cauchy-Schwarz & AM-GM.

$$\leq \langle g_t, w_{t+1} - u\rangle + \frac{\eta}{2}\|g_t\|_2^2 + \frac{1}{2\eta}\|w_t - w_{t+1}\|^2$$

$$\leq \frac{\eta}{2}\|g_t\|_2^2 + \frac{1}{2\eta}\left(\|u-w_t\|^2 - \|u-w_{t+1}\|^2\right)$$

D_t $\quad$ D_{t+1}

Step 4: summing over t

$D_1 - D_2 + D_2 - D_3 + \dots - D_{T+1}$

$$\sum_{t=1}^{T} \langle g_t, w_t - u \rangle$$

$$\leq \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2 + \frac{1}{2\eta}\|u-w_1\|^2 . \quad \blacksquare$$

Online Mirror descent: exploiting different geometry in data.

— can we develop algs w/ regrets that scale w/ other geometric measures of data. (ℓ_∞ / $\ell_\star$, etc).

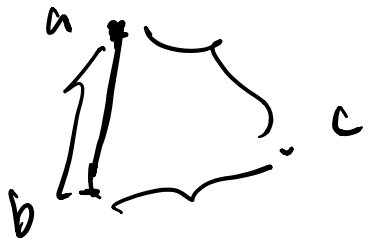
Background on norms:

Def: fn $\|\cdot\|: \mathbb{R}^d \to \mathbb{R}$ is said to be a norm, if

1. $\forall a \in \mathbb{R}, x \in \mathbb{R}^d$, $\|ax\| = |a| \cdot \|x\|$ homogeneity
2. $\forall x, y \in \mathbb{R}^d$, $\|x+y\| \leq \|x\| + \|y\|$.

triangle inequality

$d(a,b) = \|a-b\|$



3. $\|x\| = 0 \Rightarrow x = \vec{0}$.

point separation.

a, b. $\|a-b\| = 0 \Rightarrow a = b$.

Ex:

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_d^2}$$

$$\|x\|_\infty = \max_{i=1}^{d} |x_i|$$

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}} \cdot \quad p \in [1, \infty)$$

Mahalanobis norm:

A is a positive definite. $A = P \cdot P^T$ for invertible P.

$$\|x\|_A = \sqrt{x^T A x} = \sqrt{x^T P P^T x} = \|P^T x\|_2.$$

Dual norm: given $\|\cdot\|$ norm, define $\|\cdot\|_*$ as:

$$\|z\|_* = \sup_{x:\|x\| \le 1} \langle x, z \rangle.$$

$\|Px\|_2 \le 1$

$\|x\|_2$

$x = \frac{z}{\|z\|_2}$

C-S. $\le$ $\|z\|_2$

① $\|\cdot\|_*$ is also a norm.

② by defn of dual norm:

$$\langle x, z\rangle = \|x\| \left\langle \frac{x}{\|x\|}, z \right\rangle$$

$$\leq \|x\| \cdot \|z\|_* \quad \text{(generalizes Cauchy-Schwarz)}$$

Ex: $\|\cdot\|$ $\quad$ $\|\cdot\|_*$

$\|\cdot\|_2$ $\longrightarrow$ $\|\cdot\|_2$

$\|\cdot\|_1$ $\quad$ $\|\cdot\|_\infty$ $\quad \langle x, y\rangle \leq \|x\|_p \|y\|_q$ (Hölder's ineq)

$\|\cdot\|_p$, $p \in [1, \infty]$ $\quad$ $\|\cdot\|_q$.

q is p's conjugate exponent.

$$\frac{1}{p} + \frac{1}{q} = 1.$$

$\|\cdot\|_A$. $\quad$ $\|\cdot\|_{A^{-1}}$ (exercise)

General Lipschitz property (wrt arbitrary norms)

Recall: $f: \mathbb{R}^d \to \mathbb{R}$ is L-Lip wrt $\|\cdot\|$,

if, $\forall x, y$, $|f(x) - f(y)| \leq L\|x - y\|$

Relating Lipschitzness to gradient norm:

Lemma: ① $f: \mathbb{R}^d \to \mathbb{R}$ is differentiable, then

f L-Lip wrt $\|\cdot\|$ $\iff$ $\forall x$, $\|\nabla f(x)\|_* \le L$.

② f $\mathbb{R}^d \to \mathbb{R}$ is cvx, then

f L-Lip wrt $\|\cdot\|$ $\iff$ $\forall x$, $\forall g \in \partial f(x)$, $\|g\|_* \le L$.

Pf: ① left as exercise

② ($\Rightarrow$) $\forall x$, $\forall g \in \partial f(x)$. we have

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

pick g^* be $\|g^*\| \le 1$, $\langle g, g^* \rangle = \|g\|_*$

$y = x + g^*$.

$\|g\|_* = \langle g, y-x \rangle \le f(y) - f(x) \le L\|y-x\| \le L$.

($\Leftarrow$) $\forall x, y$ $\partial f(x)$

$\in \partial f(y)$ $\qquad\qquad$ $\in$

$\langle g_y, x-y \rangle \le f(x) - f(y) \le \langle g_x, x-y \rangle$

$\ge -L\|x-y\|$. $\qquad\qquad$ $\le \|g_x\|_* \cdot \|x-y\|$

$\le L\|x-y\|$

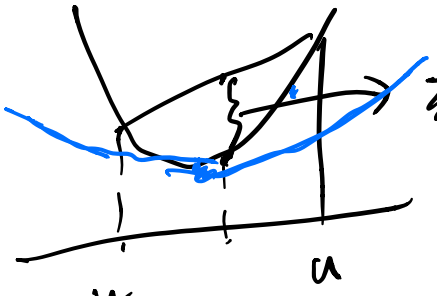
Strong convexity (for general norms):

Recall f is λ-SC wrt $\|\cdot\|$ (general) if. $\forall u, w$.

$\alpha \in (0,1)$,

$$f(\alpha w + (1-\alpha)u) \le \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2$$

general norms

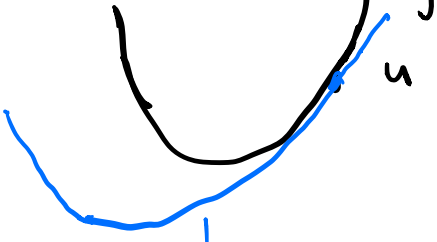


Lemma: If f is λ-SC. then $\forall u, w$.

$\forall g \in \partial f(u)$,

$$f(w) - f(u) \ge \langle g, w-u \rangle + \frac{\lambda}{2}\|w-u\|^2$$

, f



quadratic approximation at u $\geq \langle -g, \alpha(w-u) \rangle$

Pf: $\dfrac{f(\alpha w + (1-\alpha)u) - f(u)}{\alpha} \leq f(w) - f(u) - \dfrac{\lambda(1-\alpha)}{2}\|w-u\|^2$

$\downarrow$

$\langle g, w-u \rangle$

pick $\alpha = 0$.

$\Rightarrow$ Lemma statement ☒.

Def: if ψ is differentiable & strongly cvx.

then $D_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y \rangle$.

is called the Bregman divergence induced by ψ.

$\psi(w) = \frac{1}{2}\|w\|_2^2 \Rightarrow D_\psi(x,y) = \frac{1}{2}\|x-y\|_2^2$.

Next lecture:

Online Mirror descent:

$$w_{t+1} = \underset{w \in \Omega}{\arg\min} \; \langle \eta g_t, w \rangle + D_\varphi(w, w_t)$$