Announcement:

1. HW2 grades are out

   HW3 due on Apr. 15.

   Piazza: errata page for HW3.

2. mid project progress report due Apr. 6.

Problem 1: SC. First-order-pivoting for general norms

23: guarantee of OGD.

Cor: ① map $\|u-v\|_2 \le B$; ② $\ell(w,z)$ is ρ-Lip, cvx wrt w; OGD with $f_t(w) = \ell(w, z_t)$ for $z_1 \dots z_T \overset{iid}{\sim} D$, guarantees that: $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$,

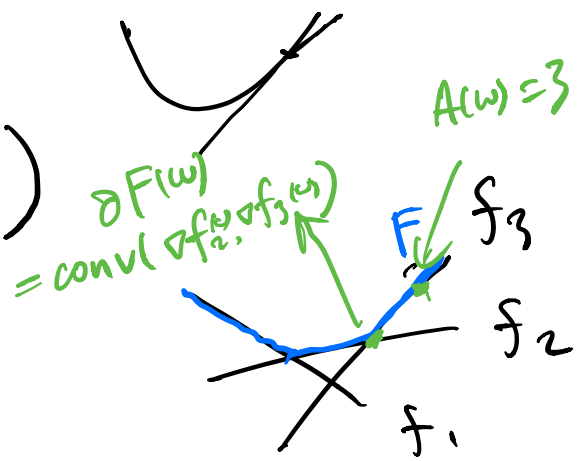
① $\eta = \frac{B}{\rho}\sqrt{\frac{1}{T}} \Rightarrow \mathbb{E}\, L_D(\bar{w}_T) \le \min_{w \in \Omega} L_D(w) + \frac{B\rho}{\sqrt{T}}$.

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

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

OGD + online to batch conversion competes with the guarantees of regularized loss minimization with only one pass over the data!

Calculating subgradients:



Fact: For $F(w) = \max_{i=1}^{n} (f_i(w))$

each f_i is cvx

$$\partial F(w) = \text{conv}\Big(\bigcup_{i \in A(w)} \partial f_i(w) \Big)$$

$$A(w) = \{ j : j \in \arg\max_i f_i(w) \}$$

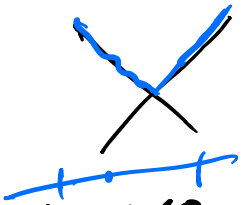
$\max(W, W)$

$\text{conv}(\{x_1, \dots x_n\}) = \{ \sum_{i=1}^{n} a_i x_i : a \in \Delta^{n-1} \}$

Ex: $F(w) = |w|$ $\quad \partial F(0) = \text{conv}(\{1, -1\}) = [-1, 1]$.

f_1 f_2

$F(w) = \max(0, 1 - y\langle w, x \rangle)$ $\quad \partial F(w) = \begin{cases} \{0\} & 1 - y\langle w, x \rangle < 0 \\ \{-yx\} & > 0 \\ \{-\alpha yx : \alpha \in [0,1]\} & = 0 \end{cases}$ cvx.



Def: if ψ is differentiable & strongly cvx.

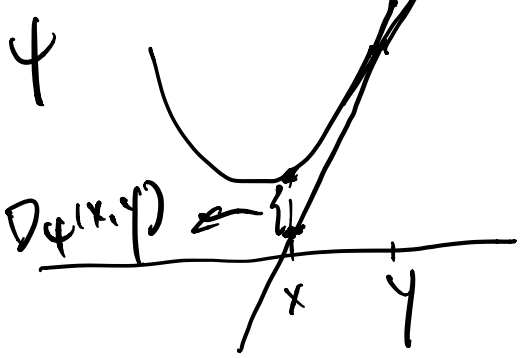
then $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$.

is called the Bregman divergence induced by ψ.

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

① ψ is λ-SC wrt $\|\cdot\|$. $x \neq y$.

$D_\psi(x, y) \geq \frac{\lambda}{2}\|x - y\|^2 > 0$.



② D_ψ can be asymmetric.

$D_\psi(x,y)$ not necessarily $= D_\psi(y,x)$.

Important examples of strongly cvx fns & Bregman divergences

① $\psi(w) = \frac{1}{2}\|w\|_A^2$ is 1-sc wrt $\|\cdot\|_A$.

$D_\psi(x,y) \overset{?}{=} \frac{1}{2}\|x-y\|_A^2$ ($\sqrt{w^T A w}$)

③ $\Omega = \Delta^{d-1}$, $\psi(w) = \sum_{i=1}^d w_i \ln w_i$ is 1-sc negative entropy wrt $\|\cdot\|_1$ (HW3).

$D_\psi(x,y) = \sum_{i=1}^d x_i \ln \frac{x_i}{y_i}$ (Relative entropy / K-L divergence / Kullback Leibler)

asymmetric.

④ $\psi(w) = \frac{1}{2(p-1)}\|w\|_p^2$ is 1-sc wrt $\|\cdot\|_p$.

For $p \in (1,2]$. (e.g. Shalev-Schwartz, 2007).

Online Mirror descent:

w_t .

$g_t \in \partial f_t(w_t)$

w_{t+1}

$$w_{t+1} = \operatorname*{argmin}_{w \in \Omega} \langle w, \eta g_t \rangle + D_\psi(w, w_t)$$

$D_\psi(w, w_t) = \psi(w) - \langle \nabla\psi(w_t), w - w_t \rangle - \psi(w_t)$

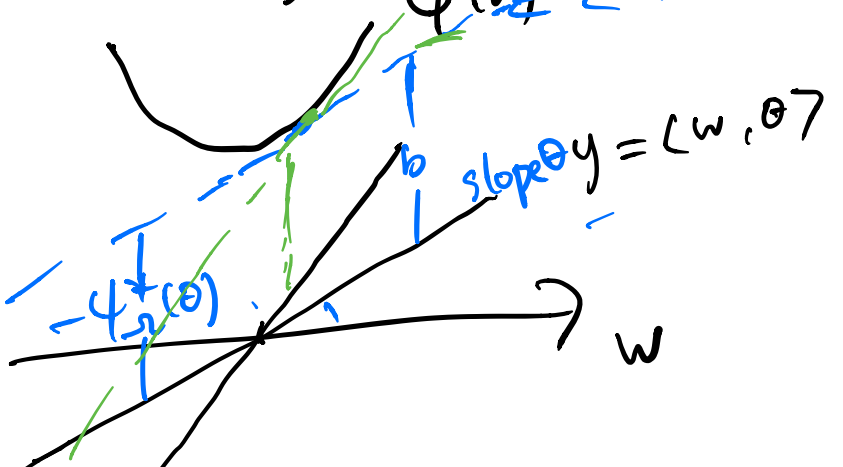
$$= \operatorname*{argmin}_{w \in \Omega} \langle w, \eta g_t - \nabla\psi(w_t) \rangle + \psi(w)$$

$$= \operatorname*{argmax}_{w \in \Omega} \langle w, \underbrace{\nabla\psi(w_t) - \eta g_t}_{\theta} \rangle - \psi(w)$$

Fact ψ is strongly cvx, Ω cvx set.

$\psi^*_\Omega(\theta) = \max_{w \in \Omega} \underbrace{\langle w, \theta \rangle - \psi(w)}_{f_w(\theta)}$ is ψ's Fenchel conjugate wrt Ω.

(this is cvx)



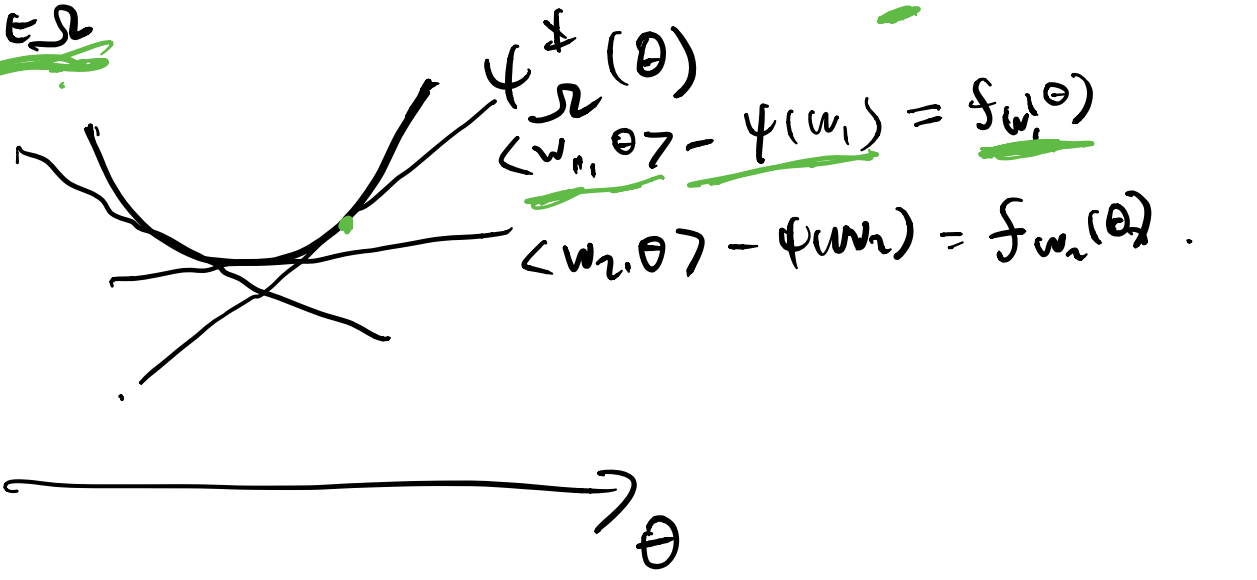
$\min_w \psi(w) - (\langle w, \theta \rangle + b)$

$= 0$.

$\nabla \psi_\Omega^*(\theta)$ exists. $\in \mathbb{R}^d$.

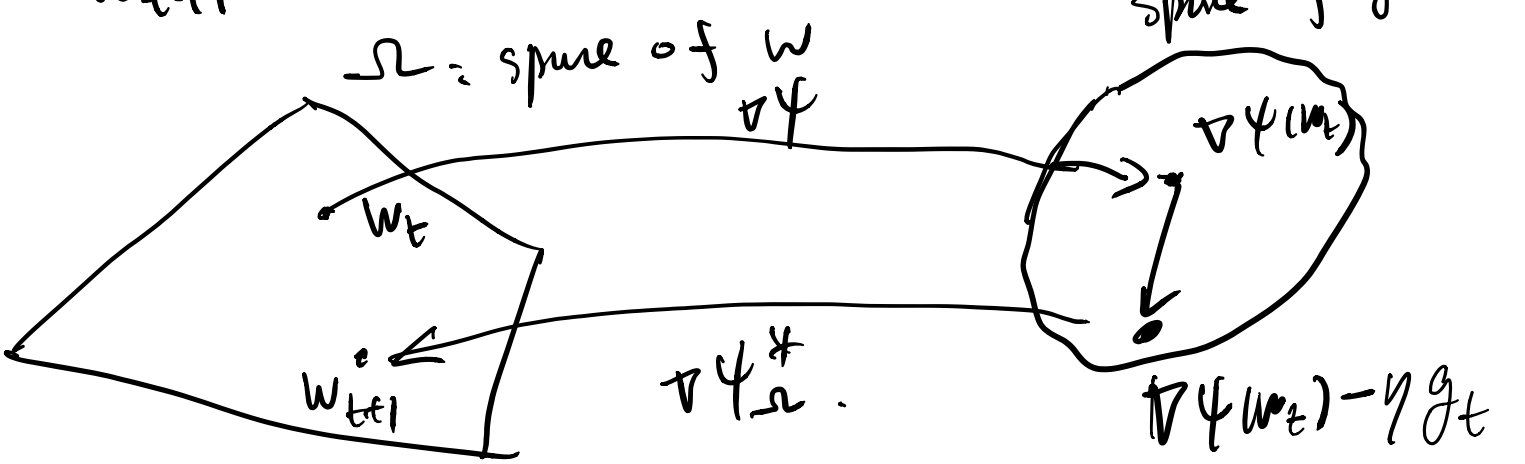
$= \arg\max_{w \in \Omega} \langle w, \theta \rangle - \psi(w)$.

$\nabla_\theta f_{w_1}(\theta) = w_1$.



OMD:

$$w_{t+1} = \nabla \psi_\Omega^* \left(\nabla \psi(w_t) - \eta \cdot g_t \right).$$



Two examples of OMD:

1. $\Omega = \mathbb{R}^d$. $\psi(w) = \frac{1}{2} \|w\|_p^2$ $p \in (1, 2]$

$$\psi_\Omega^*(\theta) = \frac{1}{2}\|\theta\|_q^2 \quad \text{where } q \text{ is s.t.}$$

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

(exercise)

$$\nabla \psi_\Omega^* = (\nabla \psi)^{-1}$$

OMD: $w_{t+1} = \nabla \psi_\Omega^*(\nabla \psi(w_t) - \eta g_t)$

$$\Rightarrow \quad \nabla \psi(w_{t+1}) = \nabla \psi(w_t) - \eta g_t.$$

$$\Rightarrow \quad \nabla \psi(w_t) = -\eta \sum_{s=1}^{t-1} g_s.$$

$$\Rightarrow \quad w_t = \nabla \psi_\Omega^*\left(-\eta \sum_{s=1}^{t-1} g_s\right).$$

p-norm algorithm.

2. exponential weight algorithm.

$\Omega = \Delta^{d-1}$. $\psi(w) = \sum_i w(i) \ln w(i)$

$=$ distance generating fn

$w_1 = \left(\frac{1}{d}, \cdots \frac{1}{d}\right)$.

$$\psi_\Omega^*(\theta) = \max_{w \in \Delta^{d-1}} \langle w, \theta \rangle - \sum_{i=1}^{d} w(i) \ln w(i)$$

$$= \max_{w(1) \cdots w(d-1) \geq 0,\ \sum_{i=1}^{d-1} w(i) \leq 1} \sum_{i=1}^{d-1} w(i)\theta(i) + \left(1 - \sum_{i=1}^{d-1} w(i)\right)\theta(d) - \sum_{i=1}^{d-1} w(i) \ln w(i) + \left(-\sum_{i=1}^{d-1} w(i)\right) \ln\left(1 - \sum_{i=1}^{d-1} w(i)\right)$$

$= \cdots$ (HW3 $\psi(w) \in [-\ln d, 0]$)

$$= \ln\left(\sum_{i=1}^{d} e^{\theta(i)}\right).$$

$$\nabla \psi_\Omega^*(\theta) = \left(\frac{e^{\theta(i)}}{\sum_{j=1}^{d} e^{\theta(j)}}\right)_{i=1}^{d}.$$

$$\nabla \psi(w) = \left(\ln w(i) + 1\right)_{i=1}^{d}$$

$$w_{t+1} = \nabla \psi_\Omega^*\left(\nabla \psi(w_t) - \eta g_t\right)$$

$$= \left(\frac{e^{\ln w_t(i) + 1 - \eta g_t(i)}}{\sum_{j=1}^{d} e^{\ln w_t(j) + 1 - \eta g_t(j)}} \right)_{i=1}^{d}$$

$$= \left(\frac{w_t(i)\, e^{-\eta g_t(i)}}{\sum_{j=1}^{d} w_t(j)\, e^{-\eta g_t(j)}} \right)_{i=1}^{d}$$

induction.

$$\Rightarrow \quad w_t(i) \propto \frac{w_1(i) \cdot e^{-\eta \sum_{s=1}^{t-1} g_s(i)}}{\frac{1}{d}}$$

$$\Rightarrow \quad w_t = \nabla \psi_R^* \left(-\eta \sum_{s=1}^{t-1} g_s \right).$$

This formula is not always true. Just coincidence.

Guarantees of OMD:

Thm: If ψ is 1-SC wrt $\|\cdot\|$, then OMD w/ ψ and learning rate η has regret guarantee:

$$\forall u \in \Omega: \quad R_T(u) \le \frac{D_\psi(u, w_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_*^2$$

specifically, if $D_\psi(u, w_1) \le H^2$ and $\forall t. \ \|g_t\|_* \le \rho$, then

$$\eta = \frac{H}{\rho\sqrt{T}} \Rightarrow R_T(u) \le H \cdot \rho \cdot \sqrt{T}.$$