V. 2. 2
2022.04.06

1. Conjugate Duality.

A. Definition. Let $F: \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$. Define $F^*: \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ by

$$f^*(y) = \sup_{x \in \mathbb{R}^n} y^T x - f(x),$$

This is the convex conjugate of f.

B. Intuition:

① $\forall x \in \mathbb{R}^n$, 过点 $(x, f(x))$ 作以 y 为斜率的直线，那么

$\hat{x} = \nabla f^*(y) = \arg\max_{x \in \mathbb{R}^n} y^T x - f(x)$ 为让直线的纵截距最小的 x.

② $\hat{x}$ 为让直线 $g = y^T x$ 比 $f(x)$ 多最多的那个点 x.

C. Example

Ex. 1.2. Let $f(x) = Cx$ for some $C \in \mathbb{R}^d$. Then $f^* = \delta_{\{C\}}$, i.e.,

$$f^*(y) = \begin{cases} +\infty, & y \neq C \\ 0, & y = C. \end{cases}$$

Ex. 1.5 (Negative Entropy). Define $f: \mathbb{R}^n_+ \to \mathbb{R}$ by

$f(x) = \sum_{i=1}^{n} x_i \cdot \ln x_i$. Then $f^*(y) = \sum_{i=1}^{d} e^{y_i - 1}$ by Proposition 1.9.

Ex. 1.6 Let $\|\cdot\|$ be a norm on $\mathbb{R}$. and let $f(x) = \frac{1}{2}\|x\|^2$, Then $f^*(x) = \frac{1}{2}\|x\|_*^2$.

D. Properties.

Proposition 1.7 [Young-Fenchel Inequality]. $\forall x, y \in \mathbb{R}^n$.

$$f^*(y) + f(x) \geq y^T \cdot x.$$

epi(f) is closed

Proposition 1.8. Regardless of whether f is, f^* is closed and convex.

Proposition 1.9. (Conjugate of separable function).
Let $f: \mathbb{R}^a \times \mathbb{R}^b \to \mathbb{R}$ be defined by $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$.
Then $f^*(y_1, y_2) = f_1^*(y_1) + f_2^*(y_2)$.

Proposition 1.10. If f is closed and convex, then $f^{**} = f$.

Proposition 1.11. If f is closed and convex then the following are equivalent.

(a) $y \in \partial f(x)$.

(b) $f(x) + f^*(y) = y^T \cdot x$.

(c) $x \in \partial f^*(y)$.

- f 是关于 primal variable x 的函数, 而 f^* 是关于 dual variable y 的函数.

Proof. 1> (a) ⇒ (b).

$y^T x - f(x) \geq y^T u - f(u), \ \forall u \in \mathbb{R}^n$ (Definition of subgradient)

$\Rightarrow y^T x - f(x) \geq \sup_{u \in \mathbb{R}^n} y^T u - f(u) = f^*(y).$ ①

$f^*(y) \geq y^T x - f(x)$ (Proposition 1.7) ②

① ② $\Rightarrow y^T x = f^*(y) + f(x).$

2> (b) ⇒ (c).

Let $g = f^*$, then $g^* = f$.

$x \in \partial g(y) \sim \forall v \in \mathbb{R}^n. \ g(v) - x^T v \geq g(y) - x^T y.$

Since $x^T y - f^*(y)$

$= x^T y - g(y)$

$= f(x)$

$= g^*(x)$

$= \sup_{v \in \mathbb{R}^n} x^T v - g(v).$

$\Rightarrow x^T y - g(y) \geq x^T v - g(v). \quad \forall v \in \mathbb{R}^n.$

3> (c) ⇒ (a).

Let $g = f^*$.

$\Rightarrow g^*(x) + g(y) = y^T x$ (by (a) ⇒ (b))

$\Rightarrow y \in \partial g^*(x) = \partial f(x).$ (by (b) ⇒ (c))

∎

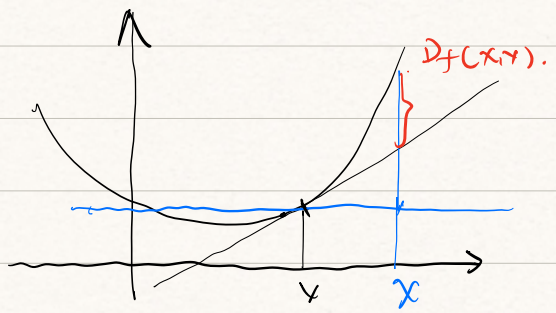
2. Bregman Divergence

A Definition 2.1. The Bregman divergence is defined to be

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

B. Intuition.

The difference between the value of function and the first order estimation.



C. Example.

Ex. 2.2. Define $f : \mathbb{R}^n \to \mathbb{R}$ by $f(x) = \|x\|_2^2$. Then

$$D_f(x,y) = \|x-y\|_2^2.$$

Ex. 2.3 (Negative Entropy).

Define $f : \mathbb{R}^n_+ \to \mathbb{R}$ by $f(x) = \sum_{i=1}^{d} x_i \ln x_i$. Then

$$D_f(x,y) = D_{KL}(x,y),$$

where $D_{KL}(x,y)$ is the generalized relative entropy.

Proof. $D_f(x,y) = f(x) - f(y) - \langle \nabla f(y),\ x-y \rangle$

$$= -H(x) + H(y) - \sum_{i=1}^{d} (1 + \ln y_i)(x_i - y_i)$$

$$= -H(X) + H(Y) - \sum_{i=1}^{d} x_i - \sum_{i=1}^{d} x_i \cdot \ln y_i + \sum_{i=1}^{d} y_i + \sum_{i=1}^{d} y_i \ln y_i.$$

$$= H(X,Y) - H(X) - \sum_{i=1}^{d} x_i + \sum_{i=1}^{d} y_i$$

$$= D_{KL}(X,Y).$$

∎

Proposition 2.4.

Negative entropy is 1-strongly-convex with respect to L_1-norm.

Proof.

Theorem. [Pinsker's Inequality]

For any distribution p, q, $D_{KL}(P,Q) \geq \frac{1}{2}\|p-q\|_1^2$.

$$D_f(P,Q) = f(P) - f(Q) - \langle \nabla f(Q), P-Q \rangle$$

$$= D_{KL}(P,Q)$$

$$\geq \frac{1}{2}\|p-q\|_1^2.$$

∎

D. Properties.

Proposition 2.8.

$D_f(x,y)$ is convex in x.

Proposition 2.9.

Let f be closed, convex and differential. Fix any $x, y \in \mathbb{X}$,

Define $\hat{x}=\nabla f(x)$, $\hat{y}=\nabla f(y)$. Then

$$\nabla f^*(\hat{x}) = x. \qquad (1)$$

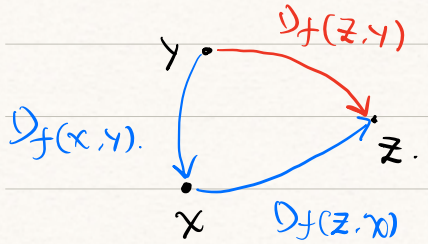
$$D_f(x,y) = D_{f^*}(\hat{y},\hat{x}). \qquad (2)$$

Proof of (2).

$$\begin{aligned} D_{f^*}(\hat{y},\hat{x}) &= f^*(\hat{y}) - f^*(\hat{x}) - \langle \nabla f^*(\hat{x}), \hat{y}-\hat{x}\rangle \\ &= \hat{y}^T y - f(y) - (\hat{x}^T x - f(x)) - \langle x, \hat{y}-\hat{x}\rangle \\ &= \hat{y}^T y - f(y) + f(x) - x^T\hat{y} \\ &= f(x) - f(y) - \langle \nabla f(y), x-y\rangle \\ &= D_f(x,y). \end{aligned}$$

□

Lemma 2.10. [Generalized Pythagoras Identity].



$$D_f(z,x) + D_f(x,y) - D_f(z,y) = (\nabla f(y) - \nabla f(x))^T (z-x)$$

Proof.

$$\begin{aligned} \text{LHS} &= f(z) - f(x) - \langle \nabla f(x), z-x\rangle + f(x) - f(y) - \langle \nabla f(y), x-y\rangle \\ &\quad - (f(z) - f(y) - \langle \nabla f(y), z-y\rangle) \\ &= \langle \nabla f(y), z-x\rangle - \langle \nabla f(x), z-x\rangle \\ &= (\nabla f(y) - \nabla f(x))^T (z-x). \end{aligned}$$

□

Ex. 2.11

Consider the case of $f(x) = \|x\|_2^2$. Let $\vec{a} = z - x$, $\vec{b} = y - x$, $\vec{c} = z - y$.

Then

$$\vec{a}^2 + \vec{b}^2 - \vec{c}^2 = D_f(z,x) + D_f(x,y) - D_f(z,y) = (\nabla f(y) - \nabla f(x))^T \cdot (z - x) = 2\vec{a} \cdot \vec{b},$$

which recovers the law of cosine.

Proposition. 2.12. $\nabla D_f(x,y) = \nabla f(x) - \nabla f(y)$.

3. Projections.

Let $\mathbb{X} \subseteq \mathbb{R}^n$ be a closed, convex set. Assume that $f: \mathbb{X} \to \mathbb{R}$ is a strictly convex function, which implies that $D_f(x,y)$ is strictly convex in x.

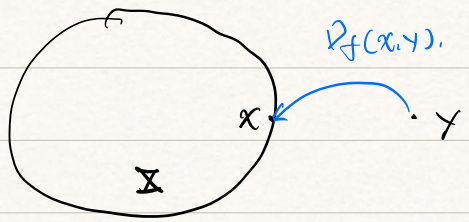
A. Definition 2.13.

The projection of y onto $\mathbb{X}$ under the Bregman divergence is

$$\Pi_{\mathbb{X}}^{f}(y) = \underset{x \in \mathbb{X}}{\operatorname{argmin}} D_f(x,y),$$

where the minimizer is uniquely determined since $D_f(x,y)$ is strictly convex in x.

B.



C. Properties.

Proposition 2.14.

Suppose that f is differentiable. Fix any $y \in \mathbb{R}^n$ and let $\pi = \Pi_{\mathbb{X}}^{f}(y)$.

$\forall\, x \in \mathbb{X}$,

$$(\nabla f(\pi) - \nabla f(y))^T \cdot (x - \pi) \geq 0.$$

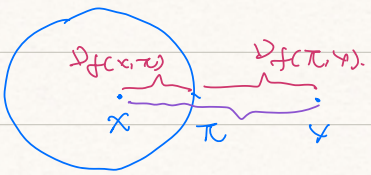
Proposition 2.15. [Anti-triangle Inequality].

Fix any $y \in \mathbb{R}^n$, and let $\pi = \Pi^{+}_{\mathcal{X}}(y)$. Then $\forall x \in \mathcal{X}$

$$D_f(x, \pi) + D_f(\pi, y) \leq D_f(x, y).$$

Proof.

$$\begin{aligned} & D_f(x, \pi) + D_f(\pi, y) - D_f(x, y) \\ &= (\nabla f(y) - \nabla f(\pi))^T \cdot (x - \pi) && \text{(Generalized Pythagoras Identity)} \\ &= -\nabla D_f(\pi, y)^T \cdot (x - \pi) \\ &\leq 0. && \text{(1st-order optimality condition)} \end{aligned}$$



Remark. [2.15.1]

In Proposition 2.15, the equality holds when $\mathcal{X}$ is a hyperplane.

Proof. Since $\mathcal{X}$ is a hyperplane, then there must exist a $v \in \mathbb{R}^d$ s.t. $v^T \cdot (x - \pi) = 0$ (v is actually the normal vector of $\mathcal{X}$), together with the 1st-order optimality condition in the proof of Prop. 2.15 concludes the proof.

4. The Mirror Descent Algorithm.

A. Motivations.

① Optimize the function f, which is Lipschitz with respect to a more general norm instead of $L2$-norm. Suppose f is L-Lipschitz w.r.t. $\|\cdot\|_\infty$, then f is $\sqrt{d}\cdot L$-Lipschitz w.r.t. L-2 norm., but the factor $\sqrt{d}$ might be undersiably large.

② In gradient descent, $x_{t+1} = x_t - \eta \cdot \nabla f(x_t)$. However, x_t and $\nabla f(x_t)$ lie in different vector spaces respectively.

B. Main Ideas.

① Mapping: Find a mirror map Φ. Use its gradient $\nabla\Phi$ which is bijection and $\nabla\Phi$'s inverse $\nabla\Phi^*$ to map back and forth between primal and dual points

② Ensuring Feasibility: If the goal is to optimize f over a constraint set $\mathcal{X}$ which is convex, project that point onto $\mathcal{X}$ under the Bregman divergence D_Φ.

C. The mirror map Φ.

Consider the mirror map $\Phi: D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$ is an open set.

A1: Φ is strictly convex and differential on all of D.

A2: The dual space of Φ is all of $\mathbb{R}^n$. That is, $\{\nabla\Phi(x): x \in D\} = \mathbb{R}^n$.

A3: The gradient of Φ diverges on the boundary of D. That is,

$$\lim_{x \to \partial D} \| \nabla F(x) \| = +\infty.$$

A function satisfying the above 3 assumptions is called Legendre function.

Proposition 3.2.

Suppose the assumptions A1 and A2 hold. Then $\nabla F: D \to \mathbb{R}^n$ is a bijection.

Proof.

1) surjection:

By Assumption A2, $\forall\, Y \in \mathbb{R}^n$, $\exists\, x \in D$ s.t. $\nabla F(x) = Y$.

2) injection:

Suppose $\exists\, x_1, x_2 \in D$ (where $x_1 \neq x_2$) s.t. $\nabla F(x_1) = \nabla F(x_2) = Y$. Then by Proposition 1.11, $Y \in \partial F(x_1) \cap \partial F(x_2)$, which contradicts with Proposition A.1.

∎

D. The constraint set X.

Typically we will optimize a constraint set X, which is supposed to have the following assumptions:

A.4. X is a closed convex set.

A.5. $X \subseteq \overline{D}$, where $\overline{D}$ is the closure of D.

A.6. $X \cap D \neq \phi$.

E. Projection in MD

A primal point $y \in D$ could be projected back onto the constraint set $\mathcal{X}$ as

$$\Pi^{\Phi}_{\mathcal{X}\cap D}(y) = \operatorname*{argmin}_{x \in \mathcal{X}\cap D} D_{\Phi}(x, y).$$

Proposition 3.3. Let $\Phi(x) = \sum_i x_i \ln x_i$. Suppose that $y \in D \setminus \mathcal{X}$, i.e., $y \in \mathbb{R}^n_{>0}$ but $\|y\|_1 \neq 1$. Then $\Pi^{\Phi}_{\mathcal{X}\cap D}(y) = \frac{y}{\|y\|_1}$.

Proof.

$$\operatorname*{argmin}_{x \in \mathcal{X}\cap D} D_{\Phi}(x,y) = \operatorname*{argmin}_{x \in \mathcal{X}\cap D} \sum_{i=1}^n (x_i \cdot \ln \frac{x_i}{y_i} - x_i + y_i) = \operatorname*{argmin}_{x \in \mathcal{X}\cap D} \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$$

↓

$D_{\Phi}(x,y)$ is the generalized KL divergence $D^G_{KL}(x \| y)$ when Φ is the negative entropy.

Let $f(x) = x \cdot \ln x$. It's clear that $f(x)$ is convex function since $f''(x) = \frac{1}{x} > 0$. Then (↗ Jensen's Ineq.)

$$\sum_i x_i \cdot \ln x_i = \sum_i y_i f(\frac{x_i}{y_i}) = \sum_i \|y\|_1 \cdot \frac{y_i}{\|y\|_1} \cdot f(\frac{x_i}{y_i}) \geq \|y\|_1 \cdot f(\sum_i \frac{y_i}{\|y\|_1} \cdot \frac{x_i}{y_i})$$

$$= \|y\|_1 \cdot f(\frac{1}{\|y\|_1}) = \ln \frac{1}{\|y\|_1},$$

where the equality holds iff $\forall i,j.\ \frac{x_i}{y_i} = \frac{x_j}{y_j}$. Therefore, $x = \operatorname*{argmin}_{x \in \mathcal{X}\cap D} D_{\Phi}(x,y)$ satisfies $x = \frac{y}{\|y\|_1}$.

∎

F. The MD Algorithm

- For $i = 1, 2, \ldots,$ do

- Incur cost $f_i(x_i)$, receive subgradient $g_i \in \partial f_i(x_i)$
  - $\hat{x}_i \in \nabla \Phi(x_i)$ (map primal point to dual)
  - $\hat{y}_{i+1} = \hat{x}_i - \eta \cdot g_i$. (take a gradient descent step in the dual)
  - $y_{i+1} = \nabla \Phi^*(\hat{y}_{i+1})$ (map new dual point back to a primal point in $\mathcal{D}$)
  - $x_{i+1} \in \Pi^{\Phi}_{\mathcal{X} \cap \mathcal{D}}(y_{i+1})$ (project new primal point onto $\mathcal{X}$)

$$y_{i+1} = \operatorname{argmin}_{y \in \mathcal{D}} \eta \cdot f_i(x_i) + D_\Phi(y, x_i)$$

$$x_{i+1} = \operatorname{argmin}_{x \in \mathcal{X} \cap \mathcal{D}} D_\Phi(x, y_{i+1})$$

Theorem 3.5

Given any $\eta > 0$ and

- A mirror map $\Phi: \mathcal{D} \to \mathbb{R}$ satisfies A_1, A_2, A_3 where $\mathcal{D}$ is an open subset of $\mathbb{R}^n$.
- A feasible set $\mathcal{X} \subseteq \mathbb{R}^n$ satisfying A_4, A_5, A_6.
- Convex functions $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$.

Furthermore, we assume Φ is ρ-strongly w.r.t. norm $\|\cdot\|$. Then the MD algorithm satisfies

$$\sum_{i=1}^{t} (f_i(x_i) - f_i(x^*)) \le \frac{D_\Phi(x^*, x_1)}{\eta} + \sum_{i=1}^{t} (\langle g_i, x_i - y_{i+1} \rangle - \frac{\rho}{2\eta} \|x_i - y_{i+1}\|^2)$$

$$\le \frac{D_\Phi(x^*, x_1)}{\eta} + \frac{\eta}{2\rho} \cdot \sum_{i=1}^{t} \|g_i\|_*^2.$$

Proof. $\forall z \in \mathcal{X}$,

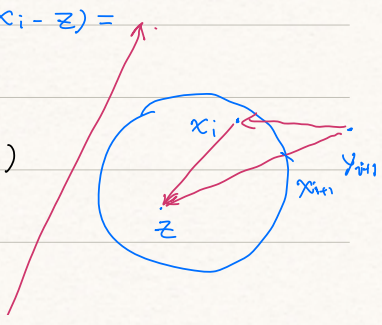
$$f_i(x_i) - f_i(z) \le g_i^T (x_i - z) \quad (x_i \to \hat{x}_i, \; y_{i+1} \to \hat{y}_{i+1})$$

$$= \frac{1}{\eta} \cdot (\nabla \Phi(x_i) - \nabla \Phi(y_{i+1}))^T (x_i - z)$$

$$= \frac{1}{\eta} \cdot (\nabla \Phi(y_{i+1}) - \nabla \Phi(x_i))^T \cdot (z - x_i)$$

Remark. In linear opt. setting

$$y_i^T \cdot (x_i - z) =$$



$$= \frac{1}{\eta}\cdot\left(D_\Phi(z,x_i)+D_\Phi(x_i,y_{i+1})-D_\Phi(z,y_{i+1})\right). \quad \text{(GPJ)}$$

$$\le \frac{1}{\eta}\left(D_\Phi(z,x_i)+D_\Phi(x_i,y_{i+1})-D_\Phi(z,x_{i+1})\right). \quad \text{(Prop. 2.15)}$$

$$\sum_{i=1}^{t}\left(f_i(x_i)-f_i(z)\right) \le \frac{1}{\eta}\cdot\left[D_\Phi(z,x_1)+\sum_{i=1}^{t} D_\Phi(x_i,y_{i+1})\right].$$

$$\begin{aligned} D_\Phi(x_i,y_{i+1}) &= \Phi(x_i)-\Phi(y_{i+1})-\nabla\Phi(y_{i+1})^T\cdot(x_i-y_{i+1}) \\ &= \Phi(x_i)-\Phi(y_{i+1})+\nabla\Phi(x_i)^T\cdot(y_{i+1}-x_i)+(\nabla\Phi(y_{i+1})-\nabla\Phi(x_i))^T(y_{i+1}-x_i). \\ &\le -\frac{\rho}{2}\|y_{i+1}-x_i\|_2^2 - \eta\cdot g_i^T\cdot(y_{i+1}-x_i) \\ &\le -\frac{\rho}{2}\|y_{i+1}-x_i\|_2^2 - \eta\cdot\|g_i\|_*\cdot\|y_{i+1}-x_i\| \\ &\le \frac{\eta^2\cdot\|g\|_*^2}{2\rho} \end{aligned}$$

∎

Corollary. 3.6

Consider the offline setting where $f_i = f$, and f is L-Lipschitz w.r.t. $\|\cdot\|$.

Let $R^2 = \sup_{x\in X\cap D} D_\Phi(x,x_1)$. Then

$$f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - f(x^*) \le \underbrace{\frac{D_\Phi(x^*,x_1)}{\eta t}}_{①} + \underbrace{\frac{\eta\cdot L^2}{2\rho}}_{②} \le RL\cdot\sqrt{\frac{2}{\rho t}}.$$

Proof. ① Theorem A.2. + Theorem 3.5

② Minimize over η (choose $\eta = \frac{R}{L}\cdot\sqrt{\frac{2\rho}{t}}$)

∎

Corollary 3.7 [Quadratic Form Regret Bound]

The regret of OMD satisfies

$$R_n \le \frac{\mathrm{diam}(\mathcal{D}, \mathcal{A})}{\eta} + \frac{\eta}{2} \cdot \sum_{t=1}^{n} \cdot \|g_t\|^2_{\nabla^{-2} \cdot \Phi(z_t)},$$

where $z_t = \alpha_t x_t + (1-\alpha) \cdot \tilde{x}_{t+1}$, $\alpha_t \in [0,1]$ ($\tilde{x}_{t+1} := y_{t+1}$ in the above).

proof. The analysis of Theorem 3.5 shows that

$$R_n \le \frac{\mathrm{diam}(F, \mathcal{A})}{\eta} + \frac{1}{\eta} \cdot \sum_{t=1}^{n} \cdot D_F(x_t, \tilde{x}_{t+1}).$$

For $D_F(x_t, \tilde{x}_{t+1})$, we have

$$\begin{aligned} D_F(x_t, \tilde{x}_{t+1}) &= -D_F(\tilde{x}_{t+1}, x_t) + \langle \nabla F(x_t) - \nabla F(\tilde{x}_{t+1}), x_t - \tilde{x}_{t+1} \rangle \\ &= -D_F(\tilde{x}_{t+1}, x_t) + \langle -\eta \cdot g_t, x_t - \tilde{x}_{t+1} \rangle \\ &\le -D_F(\tilde{x}_{t+1}, x_t) + \frac{1}{2} \|\eta \cdot g_t\|^2_{\nabla^{-2} \cdot \Phi(z_t)} + \frac{1}{2} \|x_t - \tilde{x}_{t+1}\|^2_{\nabla^{-2} \cdot \Phi(z_t)} \\ &= \frac{\eta^2}{2} \cdot \|g_t\|^2_{\nabla^{-2} \cdot \Phi(z_t)}, \end{aligned}$$

where the inequality follows from ① Hölder ineq. + AM-GM ineq / ② Young-Fenchel ineq + Lemma A.5, and the last equality from choosing z_t satisfies Lemma A.4.

∎

Remark 3.8 [Advantage of choosing $z_t = \alpha \cdot x_t + (1-\alpha) \cdot \tilde{x}_{t+1}$ instead of $z_t = \alpha \cdot x_t + (1-\alpha) \cdot x_{t+1}$]

Example 4.1 [Online convex optimization over the simplex]

Suppose that $f_1, f_2, \ldots : \mathcal{X} \to \mathbb{R}$ are convex functions, each of which is 1-Lipschitz w.r.t. $\|\cdot\|_1$. Choose Φ as the negative entropy function. $x_1 = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ and $\eta = \sqrt{\frac{2 \cdot \ln(n)}{t}}$. Then

$$\sum_{i=1}^{t} \cdot (f_i(x_i) - f_i(x^*)) \le \sqrt{2t \cdot \ln n}.$$

Proof. Thm 3.5 (2) shows that

$$\sum_{i=1}^{t} \cdot (f_i(x_i) - f_i(x^*)) \le \frac{D_{KL}(x^*, x_1)}{\eta} + \frac{\eta}{2} \cdot \sum_{i=1}^{t} \|g_i\|_\infty^2 .$$

$$\le \frac{\ln n}{\eta} + \frac{\eta}{2} t$$

$$\overset{\text{minimize over } \eta}{=} \sqrt{2t \cdot \ln n},$$

where the second inequality is due to Lemma A.2. and Theorem A.2.

Proposition 4.1.

Suppose $\mathcal{A} = \Delta^K$. Denote $p^t = \operatorname{argmin}_{p \in \Delta^K} D_\psi(p, p^{t-1}) + \eta_t \langle \hat{\ell}^{t-1}, p \rangle$.

$$\nabla \psi(p^{t+1}) = \nabla \psi(p^t) - \eta_{t+1} \cdot \hat{\ell}^t - \lambda \cdot \mathbb{1},$$

where $\mathbb{1}$ is the one-vector. (The same also holds for FTRL)

Proof. $p^{t+1} = \operatorname{argmin}_{p \in \Delta^K} D_\psi(p, p^t) + \eta_{t+1} \langle \hat{\ell}^t, p \rangle$ satisfying

λ: Lagrangian Multiplier.

$$\nabla_p \left(D_\psi(p, p^t) + \eta_{t+1} \langle \hat{\ell}^t, p \rangle + \lambda \cdot (\langle p, \mathbb{1} \rangle - 1) \right)\Big|_{p^{t+1}} = 0.$$

Appendix. A. Technical Stuff

Proposition A.1 Let $f: \mathbb{X} \to \mathbb{R}$ be strictly convex function, and $x, x' \in \mathbb{X}$ be distinct. Then $\partial f(x) \cap \partial f'(x) = \phi$.

Theorem. A.2.

Let $\mathbb{X}$ be convex and open, and $f: \mathbb{X} \to \mathbb{R}$ be convex. If f is L-Lipschitz w.r.t. $\|\cdot\|$, then

$$\|g\|_* \leq L, \quad \forall x \in \mathbb{X}, \quad \forall g \in \partial f(x).$$

Lemma. A.3

Suppose $p \in \mathbb{R}^n_{>0}$ is a distribution. Let $q = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) \in \mathbb{R}^n_{>0}$ be the uniform distribution. Then $D_{KL}(p \| q) \leq \ln n$.

Lemma A.4 [Mean Value of Bregman Divergence]

If Φ is twice differentiable, then

$$D_\Phi(x, y) = \frac{1}{2} \cdot \|y - x\|^2_{\nabla^2 \Phi(z)}$$

for some $z = \alpha x + (1-\alpha) y$, $\alpha \in [0, 1]$.

Proof. Taylor Theorem. (TBC).

□

Lemma A.5 [Convex conjugate of squared norms]

Let $f(x) = \frac{1}{2}\|x\|^2$ w.r.t. some norm $\|\cdot\|$. Then the convex conjugate of $f(x)$ is $f(y) = \frac{1}{2}\|y\|_*^2$.

Proof. (1) We show that $f(y) \leq \frac{1}{2}\|y\|_*^2$:

$$\text{LHS} = \sup_x \; x^T \cdot y - \frac{1}{2}\|x\|^2 \leq \sup_x \; \|x\| \cdot \|y\|_* - \frac{1}{2}\|x\|^2 \leq \frac{1}{2}\|y\|_*^2.$$

(x s.t. $\|x\| = \|y\|_*$)

(2) We show that $f(y) \geq \frac{1}{2}\|y\|_*^2$:

$\text{LHS} = \sup_x \; x^T \cdot y - \frac{1}{2}\|x\|^2$. Choosing x s.t. (a) $x^T \cdot y = \|x\| \cdot \|y\|_*$; and (b) $\|x\| = \|y\|_*$ proves the desired result.

∎

Appendix. B. Tiny Stuff

Proposition B.1 (Negative Entropy and Unnormalized Negative Entropy).
If $F_1 = NE$. $F_2 = UNE$., then $D_{F_1}(\cdot,\cdot) = D_{F_2}(\cdot,\cdot)$.

Proof. $D_{F_1}(x,y) = F_1(x) - F_1(y) - \langle \nabla F_1(y), x-y \rangle$

$D_{F_2}(x,y) = F_2(x) - F_2(y) - \langle \nabla F_2(y), x-y \rangle$

① If $x, y \in \Delta_{k-1}$.

$$\nabla F_1(y)_i = \ln y_i + 1$$

$$\nabla F_2(y)_i = \ln y_i$$

but $$\langle \nabla F_1(y), x-y \rangle = \sum_i (\ln y_i + 1)\cdot(x_i - y_i)$$

$$= \sum_i (\ln y_i)\cdot(x_i - y_i) + \sum_i (x_i - y_i)$$

$$= \langle \nabla F_2(y), x-y \rangle$$

② $x, y \in \Delta_{k-1}$ 不成立.

$$D_{F_2}(x,y) = \sum_i (x_i \cdot \ln x_i - x_i) - \sum_i (y_i \cdot \ln y_i - y_i) - \sum_i (x_i - y_i)$$

$$- \sum_i (\ln y_i)\cdot(x_i - y_i)$$

$$D_{F_1}(x,y) = \sum_i (x_i \cdot \ln x_i) - \sum_i y_i \cdot \ln y_i$$

$$- \sum_i (\ln x_i + 1)\cdot(x_i - y_i) \quad - \sum_i (x_i - y_i)$$

Same again!