



Lecture 7. Online Mirror Descent

Advanced Optimization (Fall 2023)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Algorithmic Framework
- Regret Analysis
- Interpretation from Primal-Dual View
- Follow-the-Regularized Leader

Recap: Reinvent Hedge Algorithm

- Proximal update rule for OGD:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}$$

- Proximal update rule for Hedge:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \parallel \mathbf{x}_t) \right\}$$

- More possibility: changing the distance measure to a more general form using *Bregman divergence*

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

Bregman Divergence

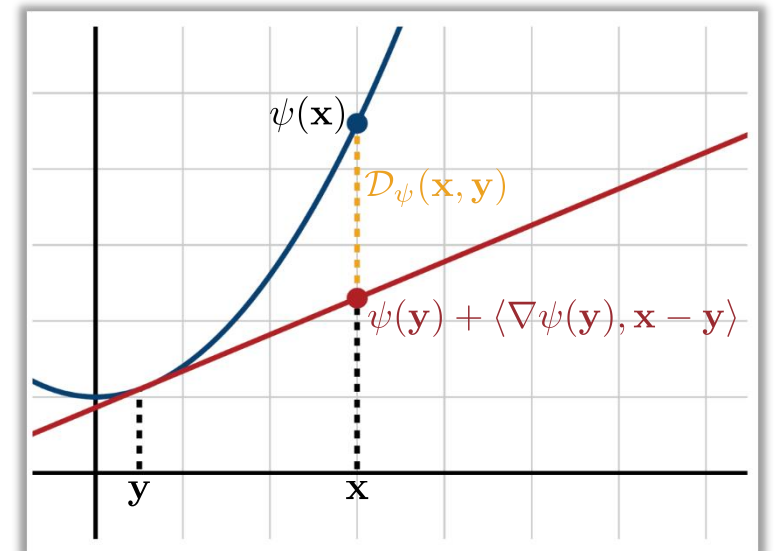
Definition 1 (Bregman Divergence). Let ψ be a **strongly convex** and differentiable function over a convex set \mathcal{X} , then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the bregman divergence \mathcal{D}_ψ associated to ψ is defined as

$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Bregman divergence measures the **difference** of a **function** and its **linear approximation**.
- Using second-order Taylor expansion, we know

$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 \nabla^2 \psi(\boldsymbol{\xi})$$

for some $\boldsymbol{\xi} \in [\mathbf{x}, \mathbf{y}]$.



Bregman Divergence

Definition 1 (Bregman Divergence). Let ψ be a **strongly convex** and differentiable function over a convex set \mathcal{X} , then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the bregman divergence \mathcal{D}_ψ associated to ψ is defined as

$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Table 1: Choice of $\psi(\cdot)$ and the corresponding Bregman divergence.

	$\psi(\mathbf{x})$	$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y})$
Squared L_2 -distance	$\ \mathbf{x}\ _2^2$	$\ \mathbf{x} - \mathbf{y}\ _2^2$
Mahalanobis distance	$\ \mathbf{x}\ _A^2$	$\ \mathbf{x} - \mathbf{y}\ _A^2$
Negative entropy	$\sum_i x_i \log x_i$	$\text{KL}(\mathbf{x} \parallel \mathbf{y})$

Online Mirror Descent

Online Mirror Descent

At each round $t = 1, 2, \dots$

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

where $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ is the Bregman divergence.

- $\psi(\cdot)$ is required to be **strongly convex** and differentiable over a convex set \mathcal{X} .
- Strong convexity of ψ will ensure the uniqueness of the minimization problem, and actually we further need some analytical assumptions (see later mirror map definition) to ensure the solutions' feasibility in domain \mathcal{X} .

Online Mirror Descent

- So we can unify OGD and Hedge from the same view of OMD.

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret _T
OGD	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2 \right\}$	$\ \mathbf{x}\ _2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
Hedge	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \Delta_N} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \parallel \mathbf{x}_t) \right\}$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T \log N})$

- We also learn ONS for exp-concave functions, can it be included?

Recap: ONS in a view of Proximal Gradient

Convex Problem

Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$

$$\text{OGD: } \mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \frac{1}{\sqrt{t}} \nabla f_t(\mathbf{x}_t) \right]$$

Proximal type update:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}_t\|_2^2$$

Exp-concave Problem

Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|_{\nabla f_t(\mathbf{y}) \nabla f_t(\mathbf{y})^\top}^2$

$$\text{ONS: } A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top$$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}^{A_t} \left[\mathbf{x}_t - \frac{1}{\gamma} A_t^{-1} \nabla f_t(\mathbf{x}_t) \right]$$

Proximal type update:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_t\|_{A_t}^2$$

Online Mirror Descent

- Our previous mentioned algorithms can **all be covered** by OMD.

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret _T
OGD for convex	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2$	$\ \mathbf{x}\ _2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2$	$\ \mathbf{x}\ _2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma} \log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _{A_t}^2$	$\ \mathbf{x}\ _{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma} \log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \Delta_N} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \ \mathbf{x}_t)$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T \log N})$

General Regret Analysis for OMD

OMD update:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

Lemma 1 (Mirror Descent Lemma). *Let \mathcal{D}_ψ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \rightarrow \mathbb{R}$ and assume ψ to be λ -strongly convex with respect to a norm $\|\cdot\|$. Then, $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds*

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\frac{1}{\eta_t} (\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}))}_{\text{bias term (range term)}} + \underbrace{\frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2}_{\text{variance term (stability term)}} - \underbrace{\frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)}_{\text{negative term}}$$

bias term (range term)

variance term (stability term)

negative term

Proof of Mirror Descent Lemma

Lemma 1 (Mirror Descent Lemma). Let \mathcal{D}_ψ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \rightarrow \mathbb{R}$ and assume ψ to be λ -strongly convex with respect to a norm $\|\cdot\|$. Then, $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{1}{\eta_t} (\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})) + \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

Proof.
$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{u}) &\leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle \\ &\leq \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}} \end{aligned}$$

In the following, we will use the *stability lemma* to analyze term (a), and use the *Bregman proximal inequality* to analyze term (b).

Proof of Mirror Descent Lemma

Proof.
$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$$

We introduce the following stability lemma to analyze term (a):

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ -strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star.$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ -strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star.$$

Proof. For any convex function f , we have the **first-order optimality condition**:

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{X} \iff \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geq 0 \quad \forall \mathbf{y} \in \mathcal{X}$$

Therefore, for $\mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})\}$, we have

$$\langle \mathbf{g}' + \nabla \psi(\mathbf{x}') - \nabla \psi(\mathbf{c}), \mathbf{u} - \mathbf{x}' \rangle \geq 0 \text{ holds for } \forall \mathbf{u} \in \mathcal{X}.$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ -strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star.$$

Proof. $\langle \mathbf{g}' + \nabla\psi(\mathbf{x}') - \nabla\psi(\mathbf{c}), \mathbf{u} - \mathbf{x}' \rangle \geq 0$ holds for $\forall \mathbf{u} \in \mathcal{X}$.

By the first-order optimality conditions of \mathbf{x} and \mathbf{x}' ,

$$\langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{c}) + \mathbf{g}, \mathbf{x}' - \mathbf{x} \rangle \geq 0$$

$$\langle \nabla\psi(\mathbf{x}') - \nabla\psi(\mathbf{c}) + \mathbf{g}', \mathbf{x} - \mathbf{x}' \rangle \geq 0$$

$$\Rightarrow \langle \mathbf{x}' - \mathbf{x}, \mathbf{g} - \mathbf{g}' \rangle \geq \langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \quad (1)$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ -strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_*$$

Proof. Besides, by the **strong convexity** of ψ , we have

$$\begin{aligned} \langle \nabla \psi(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle &\geq \psi(\mathbf{x}) - \psi(\mathbf{x}') + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \\ \langle \nabla \psi(\mathbf{x}'), \mathbf{x}' - \mathbf{x} \rangle &\geq \psi(\mathbf{x}') - \psi(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \end{aligned}$$

Summing them up, we get

$$\langle \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq \lambda \|\mathbf{x} - \mathbf{x}'\|^2 \quad (2)$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg \min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ -strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star.$$

Proof.

$$\langle \mathbf{x}' - \mathbf{x}, \mathbf{g} - \mathbf{g}' \rangle \geq \langle \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \quad (1)$$

$$\langle \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq \lambda \|\mathbf{x} - \mathbf{x}'\|^2 \quad (2)$$

$$\begin{aligned} \Rightarrow \quad \lambda \|\mathbf{x} - \mathbf{x}'\|^2 &\stackrel{(2)}{\leq} \langle \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \stackrel{(1)}{\leq} \langle \mathbf{x}' - \mathbf{x}, \mathbf{g} - \mathbf{g}' \rangle \\ &\leq \|\mathbf{x} - \mathbf{x}'\| \|\mathbf{g} - \mathbf{g}'\|_\star \quad (\text{Hölder's inequality}) \end{aligned}$$

$$\Rightarrow \quad \lambda \|\mathbf{x} - \mathbf{x}'\| \stackrel{(2)}{\leq} \|\mathbf{g} - \mathbf{g}'\|_\star \quad \square$$

Proof of Mirror Descent Lemma

Proof.
$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$$

We further introduce following lemma to analyze term (b).

Lemma 3 (Bregman Proximal Inequality). *Let \mathcal{X} be a convex set in a Banach space \mathcal{B} . Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a closed proper convex function on \mathcal{X} . Given a convex regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$, we denote its induced Bregman divergence by $\mathcal{D}_\psi(\cdot, \cdot)$. Then, any update of the form*

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \{ \langle \mathbf{g}_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \}$$

satisfies the following inequality for any $\mathbf{u} \in \mathcal{X}$:

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Crucial for analysis of **first-order optimization methods** based on Bregman divergence.

Bregman Proximal Inequality

Lemma 3 (Bregman Proximal Inequality). *The Bregman proximal update in the form of $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \{ \langle \mathbf{g}_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \}$ satisfies*

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Proof. Recall that for any convex function f , we have the following **first-order optimality condition**:

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{X} \iff \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geq 0 \quad \forall \mathbf{y} \in \mathcal{X}$$

Therefore, for $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \{ \langle \mathbf{g}_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \}$, we have

$$\langle \mathbf{g}_t + \nabla \psi(\mathbf{x}_{t+1}) - \nabla \psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle \geq 0 \text{ holds for any } \mathbf{u} \in \mathcal{X}.$$

Bregman Proximal Inequality

Lemma 3 (Bregman Proximal Inequality). *The Bregman proximal update in the form of $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \{ \langle \mathbf{g}_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \}$ satisfies*

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Proof. $\langle \mathbf{g}_t + \nabla \psi(\mathbf{x}_{t+1}) - \nabla \psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle \geq 0$ holds for any $\mathbf{u} \in \mathcal{X}$.

On the other hand, the right side of Lemma 3 is:

$$\begin{aligned} & \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \\ &= \cancel{\psi(\mathbf{u})} - \cancel{\psi(\mathbf{x}_t)} - \langle \nabla \psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_t \rangle - \cancel{\psi(\mathbf{u})} + \cancel{\psi(\mathbf{x}_{t+1})} + \langle \nabla \psi(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1} \rangle \\ & \quad - \cancel{\psi(\mathbf{x}_{t+1})} + \cancel{\psi(\mathbf{x}_t)} + \langle \nabla \psi(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ &= \langle \nabla \psi(\mathbf{x}_{t+1}) - \nabla \psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1} \rangle. \end{aligned}$$

Rearranging the terms can finish the proof. □

Proof of Mirror Descent Lemma

Proof.
$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$$

Lemma 2 (Stability Lemma).

$$\lambda \|\mathbf{x}_1 - \mathbf{x}_2\| \leq \|\mathbf{g}_1 - \mathbf{g}_2\|_*$$

\Rightarrow term (a) = $\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \leq \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_*^2$ (think of two updates: one for \mathbf{x}_{t+1} with $\nabla f_t(\mathbf{x}_t)$ and another one for \mathbf{x}_t with 0)

Lemma 3 (Bregman Proximal Inequality).

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

\Rightarrow term (b) $\leq \frac{1}{\eta_t} \left(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \underbrace{\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)}_{\text{negative term, usually dropped; but sometimes highly useful}} \right)$

$\Rightarrow f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{1}{\eta_t} (\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})) + \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$ □

General Regret Analysis for OMD

Lemma 1 (Mirror Descent Lemma). *Let \mathcal{D}_ψ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \rightarrow \mathbb{R}$ and assume ψ to be λ -strongly convex with respect to a norm $\|\cdot\|$. Then, $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds*

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{1}{\eta_t} (\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})) + \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

Using Lemma 1, we can easily prove the following cumulative regret bound for OMD.

Theorem 4 (General Regret Bound for OMD). *Assume ψ is λ -strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds*

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^T \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

General Regret Analysis for OMD

Theorem 4 (General Regret Bound for OMD). Assume ψ is λ -strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^T \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

Proof.

$$\begin{aligned} \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) &\leq \sum_{t=1}^T \left(\frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) \right) + \sum_{t=1}^T \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \sum_{t=1}^T \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \\ &= \frac{1}{\eta_1} \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1) - \frac{1}{\eta_T} \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{T+1}) + \sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) + \sum_{t=1}^T \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \sum_{t=1}^T \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \\ &\leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^T \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \quad (\eta_t = \eta_{t-1}) \end{aligned}$$

□

General Regret Analysis for OMD

Theorem 4 (General Regret Bound for OMD). Assume ψ is λ -strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^T \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

With this general regret bound for OMD, it will become straightforward to analyze OGD/Hedge/ONS *in a unified way*, which we previously analyzed specifically for each algorithm.

OMD Implication: Recovering OGD

Algorithm. With Theorem 3, it is straightforward to recover OGD:

OGD for convex	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \left\langle \mathbf{x}, \frac{1}{\sqrt{t}} \nabla f_t(\mathbf{x}_t) \right\rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2 \right\}$
----------------	---

- $\psi(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$
- The dual norm of $\|\cdot\|_2$ is still $\|\cdot\|_2$

Regret Analysis.

$$\begin{aligned} \Rightarrow \quad & \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \sum_{t=1}^T \left(\frac{1}{\eta_t} \|\mathbf{u} - \mathbf{x}_t\|_2^2 - \frac{1}{\eta_t} \|\mathbf{u} - \mathbf{x}_{t+1}\|_2^2 \right) + \sum_{t=1}^T \eta_t \|\nabla f_t(\mathbf{x}_t)\|_2^2 \\ &= \frac{1}{\eta_1} \|\mathbf{u} - \mathbf{x}_1\|_2^2 - \frac{1}{\eta_T} \|\mathbf{u} - \mathbf{x}_{T+1}\|_2^2 + \sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \|\mathbf{u} - \mathbf{x}_t\|_2^2 + \sum_{t=1}^T \frac{\eta_t}{\lambda} \|f_t(\mathbf{x}_t)\|_2^2 \\ &\leq \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \sum_{t=1}^T \eta_t G^2 \leq 3DG\sqrt{T} \quad (\eta_t = \frac{D}{G\sqrt{t}} \text{ and } \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}) \quad \square \end{aligned}$$

OMD Implication: Recovering Hedge

Algorithm. With Theorem 3, it is straightforward to recover Hedge:

Hedge for PEA	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \eta \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \parallel \mathbf{x}_t) \right\}$
---------------	---

- Negative entropy is 1-strongly convex w.r.t. $\| \cdot \|_1$
- The dual norm of $\| \cdot \|_1$ is $\| \cdot \|_\infty$
- We initialize the initial prediction $\mathbf{x}_1 = \{\frac{1}{N}, \dots, \frac{1}{N}\}$

Regret Analysis.

$$\Rightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \frac{\text{KL}(\mathbf{u} \parallel \mathbf{x}_1)}{\eta} + \eta \sum_{t=1}^T \|\ell_t\|_\infty^2 \leq \frac{\ln N}{\eta} + \eta T$$

$(\text{KL}(\mathbf{u} \parallel \mathbf{x}_1) \leq \ln N, \forall \mathbf{u}) \quad (\ell_t(i) \leq 1, \forall i \in [N])$

□

OMD Implication: Recovering ONS

Algorithm. With Theorem 3, it is straightforward to recover ONS:

ONS for exp-concave	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \frac{1}{\gamma} \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _{A_t}^2 \right\}$
---------------------	--

- $\psi_t(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_{A_t}^2$ is 1-strongly convex w.r.t. $\|\cdot\|_{A_t}$ with $A_t = \varepsilon I + \sum_{s=1}^t \nabla f_s(\mathbf{x}_s) \nabla f_s(\mathbf{x}_s)^\top$
- The dual norm of $\|\cdot\|_{A_t}$ is $\|\cdot\|_{A_t^{-1}}$

Regret Analysis.

$$\begin{aligned}
 \Rightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) &\leq \frac{\gamma}{2} \sum_{t=1}^T \left(\|\mathbf{u} - \mathbf{x}_t\|_{A_t}^2 - \|\mathbf{u} - \mathbf{x}_{t+1}\|_{A_t}^2 - \underbrace{\|\mathbf{u} - \mathbf{x}_t\|_{\nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top}^2}_{\text{(exp-concavity)}} \right) + \frac{1}{2\gamma} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2 \\
 &= \frac{\gamma}{2} \sum_{t=1}^T \left(\underbrace{\|\mathbf{u} - \mathbf{x}_t\|_{A_{t-1}}^2}_{\text{(telescope)}} - \|\mathbf{u} - \mathbf{x}_{t+1}\|_{A_t}^2 \right) + \frac{1}{2\gamma} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2 \\
 &\leq \frac{\gamma}{2} \|\mathbf{u} - \mathbf{x}_1\|_{A_0}^2 + \frac{1}{2\gamma} \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2
 \end{aligned}$$

□

OMD Implication: Recovering OGD for S.C.

Algorithm. With Theorem 3, we can recover OGD for strongly convex function:

OGD for strongly convex	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \frac{1}{\sigma t} \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2 \right\}$
-------------------------	--

- $\psi_t(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$
- The dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$

Regret Analysis.

$$\begin{aligned}
 \Rightarrow \quad \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) &\leq \frac{1}{2} \sum_{t=1}^T \left(\frac{1}{\eta_t} \|\mathbf{u} - \mathbf{x}_t\|_2^2 - \frac{1}{\eta_t} \|\mathbf{u} - \mathbf{x}_{t+1}\|_2^2 - \underbrace{\sigma \|\mathbf{u} - \mathbf{x}_t\|_2^2}_{\text{strong convexity}} \right) + \frac{1}{2} \sum_{t=1}^T \eta_t \|\nabla f_t(\mathbf{x}_t)\|_2^2 \\
 &= \frac{1}{2} \sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma \right) \|\mathbf{u} - \mathbf{x}_t\|_2^2 + \frac{1}{2} \sum_{t=1}^T \eta_t G^2 \\
 &= 0 + \frac{1}{2} \sum_{t=1}^T \frac{G^2}{\sigma t}
 \end{aligned}$$

□

A Summary of OMD Deployment

- Our previous mentioned algorithms can **all be covered** by OMD.

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret _T
OGD for convex	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2$	$\ \mathbf{x}\ _2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2$	$\ \mathbf{x}\ _2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma} \log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _{A_t}^2$	$\ \mathbf{x}\ _{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma} \log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \Delta_N} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \ \mathbf{x}_t)$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T \log N})$

Another View for Mirror Descent

Theorem 5. *The OMD update form*

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\} \quad (\star)$$

is equivalent to the following two-step updates:

$$\begin{cases} \nabla \psi(\mathbf{y}_{t+1}) = \nabla \psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{cases} \quad (\diamond)$$

Proof. $(\diamond) \mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$

$$= \arg \min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x}) - \psi(\mathbf{y}_{t+1}) - \langle \nabla \psi(\mathbf{y}_{t+1}), \mathbf{x} - \mathbf{y}_{t+1} \rangle$$

$$= \arg \min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x}) - \langle \nabla \psi(\mathbf{y}_{t+1}), \mathbf{x} \rangle \quad (\text{definition of Bregman divergence})$$

$$= \arg \min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x}) - \langle \nabla \psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle$$

Another View for Mirror Descent

Theorem 5. *The OMD update form*

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\} \quad (\star)$$

is equivalent to the following two-step updates:

$$\begin{cases} \nabla \psi(\mathbf{y}_{t+1}) = \nabla \psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{cases} \quad (\diamond)$$

Proof. $(\star) \quad \mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$

$$= \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \psi(\mathbf{x}) - \psi(\mathbf{x}_t) - \langle \nabla \psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle \right\}$$

(definition of Bregman divergence)

$$= \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \psi(\mathbf{x}) - \langle \nabla \psi(\mathbf{x}_t), \mathbf{x} \rangle \right\}$$

□

Another View for Mirror Descent

- A two-step update for mirror descent

$$\begin{cases} \nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{cases}$$

- The first step is somehow similar to a “*gradient descent*” step;
- The second step looks like a “*projection*” step.

⇒ Key role in **mirror** descent: the operator $\nabla\psi(\cdot)$

Primal-Dual View for Mirror Descent

- Recall the gradient descent update

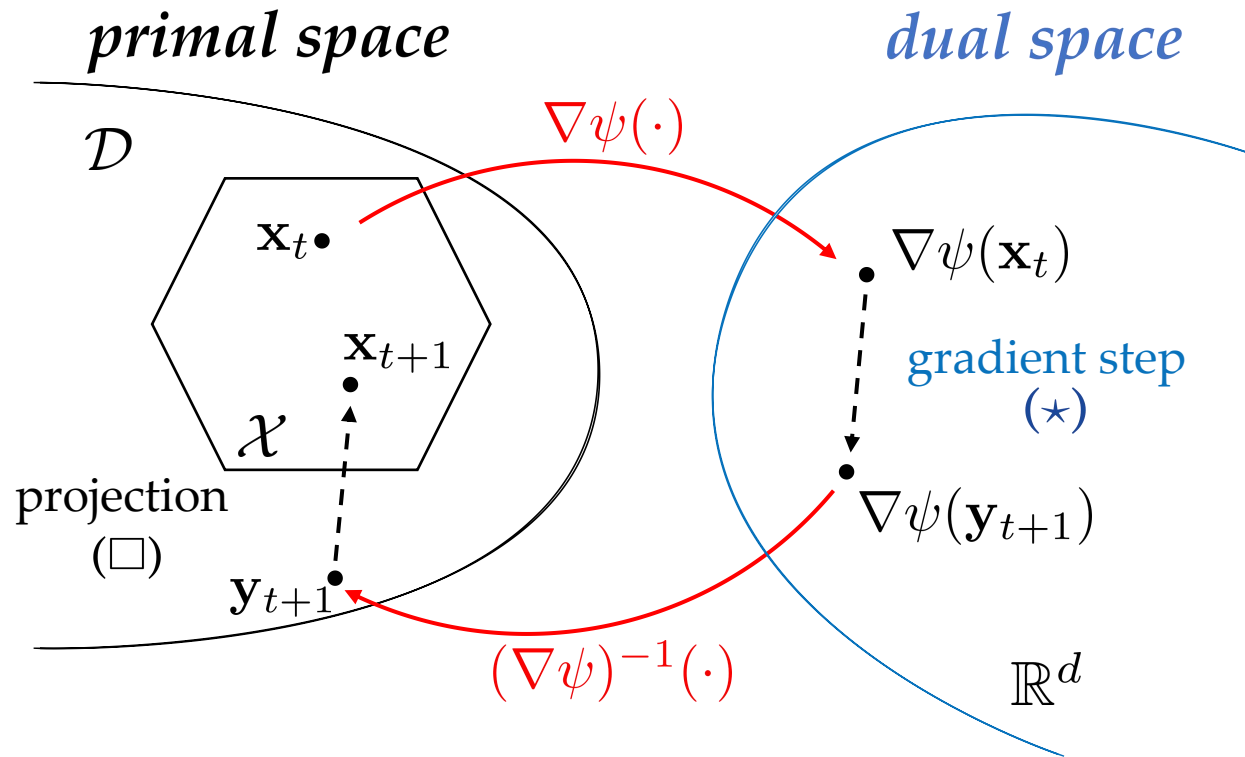
$$\mathbf{x} - \eta \nabla f(\mathbf{x})$$

but this simply *does not make sense* for general non-Euclidean space...

- Bits in convex analysis
 - consider a Banach space \mathcal{B} , whose dual space is denoted by \mathcal{B}^*
 - \mathbf{x} is in the primal space \mathcal{B} , and $\nabla f(\mathbf{x})$ is in the dual space \mathcal{B}^*

\Rightarrow a simple intuition: $f(\mathbf{x} + \Delta \mathbf{x}) \approx \langle \nabla f(\mathbf{x}), \Delta \mathbf{x} \rangle$

Primal-Dual View for Mirror Descent



$$(\star) \quad \nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t)$$

$$(\square) \quad \mathbf{x}_{t+1} \in \Pi_{\mathcal{X}}^{\psi}[\mathbf{y}_{t+1}]$$

$$(\Pi_{\mathcal{X}}^{\psi}[\mathbf{y}] = \arg \min_{\mathbf{x} \in \mathcal{X} \cap \mathcal{D}} \mathcal{D}_{\psi}(\mathbf{x}, \mathbf{y}))$$

$\nabla\psi(\cdot)$ is the **mirror map** to link two spaces

Mirror Map

Definition 2 (Mirror Map). Let $\mathcal{D} \subset \mathbb{R}^n$ be a convex open set such that \mathcal{X} is included in its closure, that is $\mathcal{X} \subset \overline{\mathcal{D}}$, and $\mathcal{X} \cap \mathcal{D} \neq \emptyset$. We say that $\psi : \mathcal{D} \rightarrow \mathbb{R}$ is a mirror map if it satisfies the following properties:

- (i) ψ is strictly convex and differentiable;
- (ii) The gradient of ψ takes all possible values, that is $\nabla\psi(\mathcal{D}) = \mathbb{R}^n$;
- (iii) The gradient of ψ diverges on the boundary of \mathcal{D} , that is

$$\lim_{\mathbf{x} \rightarrow \partial\mathcal{D}} \|\nabla\psi(\mathbf{x})\| = +\infty$$

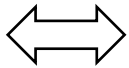
See [Chapter 4.1 of Bubeck's book](#) for rigorous discussions.

Mirror Map Calculation

$$\nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t)$$

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$$

equivalent



$$\mathbf{y}_{t+1} = \nabla\psi^*(\nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t))$$

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$$

- Here, $\nabla\psi^*(\cdot)$ is the *Fenchel Conjugate* of $\nabla\psi(\cdot)$.

Definition 3 (Fenchel Conjugate). For a function $f : \mathbb{R}^d \rightarrow [-\infty, \infty]$, we define its Fenchel conjugate $f^* : \mathbb{R}^d \rightarrow [-\infty, \infty]$ as

$$f^*(\mathbf{g}) = \sup_{\mathbf{y} \in \mathbb{R}^d} \{ \langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y}) \}.$$

Mirror Map Calculation

Proof. We first show for any convex and closed f , $\mathbf{g} = \nabla f(\mathbf{x}) \iff \mathbf{x} = \nabla f^*(\mathbf{g})$.

By the convexity of f ($f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{y}$):

$$\langle \mathbf{g}, \mathbf{x} \rangle - f(\mathbf{x}) \geq \langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y}), \forall \mathbf{y}$$

which means $\langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y})$ achieves its supremum in \mathbf{y} at $\mathbf{y} = \mathbf{x}$. Thus, by the definition of Fenchel Conjugate:

$$f^*(\mathbf{g}) = \sup_{\mathbf{y} \in \mathbb{R}^d} \langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y}) = \langle \mathbf{g}, \mathbf{x} \rangle - f(\mathbf{x})$$

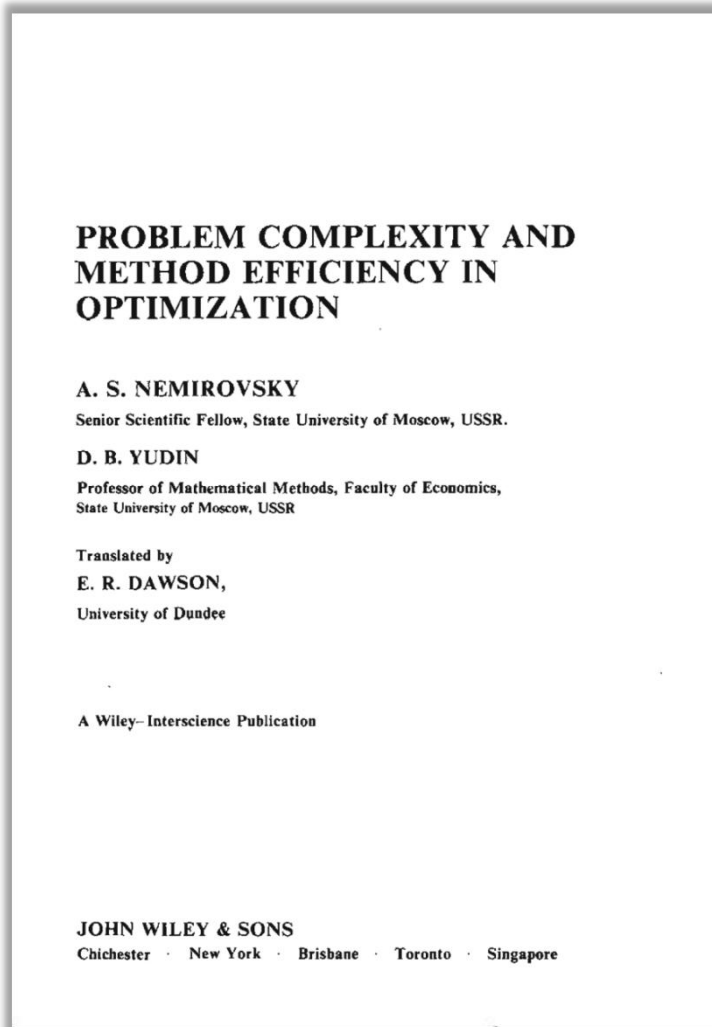
By taking the gradient w.r.t. \mathbf{g} at both sides:

$$\nabla f^*(\mathbf{g}) = \mathbf{x}$$

Therefore we have proved that $\mathbf{g} = \nabla f(\mathbf{x}) \iff \mathbf{x} = \nabla f^*(\mathbf{g})$.

By setting $f(\cdot) = \psi(\cdot)$ and $\mathbf{x} = \mathbf{y}_{t+1}$, we finish the proof. □

Mirror Descent: history bits



A. S. Nemirovski (1947 -



D. B. Yudin (1919 - 2006)

A.S. Nemirovski, D.B. Yudin, **Problem Complexity and Method Efficiency in Optimization**. Wiley-Interscience Series in Discrete Mathematics (A Wiley-Interscience Publication/Wiley, New York, 1983)

23. Nemirovskiy, A. S., and Yudin, D. B. (1979). Efficient methods of solving convex-programming problems of high dimensionality. *Ekonomika i matem. metody*, XV, No. 1. (In Russian.)

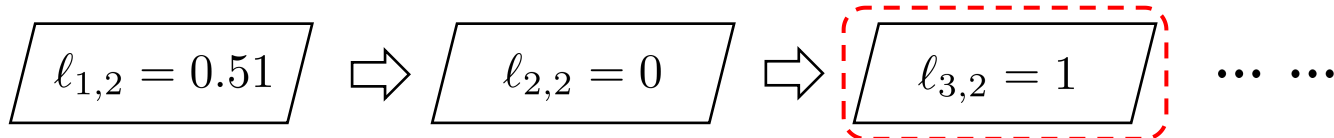
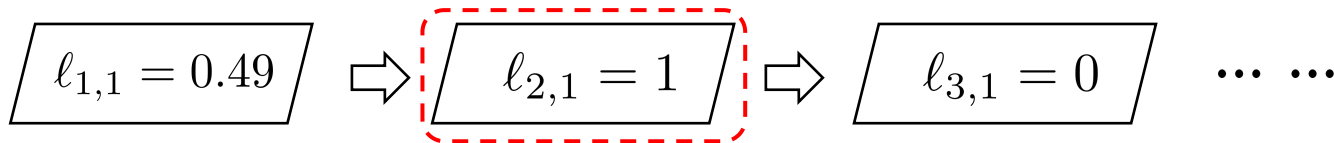
Another OCO Framework: FTRL

- Recall: Follow the Leader (FTL)

Select the expert that *performs best so far*, specifically,

$$\mathbf{p}_t^{\text{FTL}} = \arg \min_{\mathbf{p} \in \Delta_N} \langle \mathbf{p}, \mathbf{L}_{t-1} \rangle$$

where $\mathbf{L}_{t-1} \triangleq \sum_{s=1}^{t-1} \ell_s \in \mathbb{R}^N$ is the cumulative loss vector.



$$\begin{aligned} \text{Reg}_T &= \sum_{t=1}^T \langle \mathbf{p}_t, \ell_t \rangle - \min_{i \in [N]} \sum_{t=1}^T \ell_{t,i} \\ &= T - \frac{T}{2} = \mathcal{O}(T) \end{aligned}$$

FTL achieves *linear regret* in the worst case!

Another OCO Framework: FTRL

- Recall: Follow the Leader (FTL)

Select the expert that *performs best so far*, specifically,

$$\mathbf{p}_t^{\text{FTL}} = \arg \min_{\mathbf{p} \in \Delta_N} \langle \mathbf{p}, \mathbf{L}_{t-1} \rangle$$

where $\mathbf{L}_{t-1} \triangleq \sum_{s=1}^{t-1} \ell_s \in \mathbb{R}^N$ is the cumulative loss vector.

- As mentioned, FTL is *sub-optimal* due to its *unstable* nature.

⇒ a natural idea: *adding regularizers* to stabilize the algorithm.

Another OCO Framework: FTRL

Follow The Regularized Leader (FTRL)

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^t f_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\},$$

where $\psi_{t+1} : \mathcal{X} \mapsto \mathbb{R}$ is the regularizer at round $t + 1$ update.

FTRL: essentially adding regularizer to stabilize the FTL algorithm.

We use time-varying regularizer to encode the potentially changing step sizes.

FTRL vs. OMD: Update Styles

- OMD update style:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \eta_t \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

- FTRL update style:

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^t f_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\}$$

Comparison:

- in OMD, \mathbf{x}_{t+1} depends on \mathbf{x}_t and $f_t(\cdot)$;
- in FTRL, \mathbf{x}_{t+1} depends on entire history $\{f_s(\cdot)\}_{s=1}^t$ and regularizer ψ_{t+1} .

Linearization in FTRL

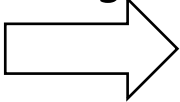
- FTRL update requires to store all the historical online functions.

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^t f_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\}$$

- *Surrogate optimization*: maintain regret while achieving one-pass update

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle \triangleq \ell_t(\mathbf{x}_t) - \ell_t(\mathbf{u})$$

where we define the linear surrogate loss as $\ell_t(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle$.

surrogate 

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^t \ell_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\} \\ &= \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^t \langle \nabla f_s(\mathbf{x}_s), \mathbf{x} \rangle + \psi_{t+1}(\mathbf{x}) \right\} \end{aligned}$$

It suffices to store gradient vectors only.

General Analysis of FTRL

Lemma 4 (FTRL Regret). We denote that $F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$. Thus, the FTRL algorithm runs $\mathbf{x}_t = \arg \min_{\mathbf{x} \in \mathcal{X}} F_t(\mathbf{x})$. Then, for any $\mathbf{u} \in X$, we have

$$\begin{aligned} \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) &= \underbrace{\psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x})}_{\text{(range term)}} \\ &\quad + \underbrace{\sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \right)}_{\text{(stability term)}} \\ &\quad + \underbrace{F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})}_{\substack{(\mathbf{x}_{T+1} = \arg \min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \\ \text{thus } \leq 0)}} \end{aligned}$$

General Analysis of FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Lemma 4.

$$\begin{aligned} \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) &= \underbrace{\psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x})}_{\text{(range term)}} \\ &\quad + \underbrace{\sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \right)}_{\text{(stability term)}} \\ &\quad + \underbrace{F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})}_{\text{(negative term)}} \end{aligned}$$

Proof. The term $\sum_{t=1}^T f_t(\mathbf{x}_t)$ appears at both side of the equality, thus we verify

$$-\sum_{t=1}^T f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \right) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}).$$

General Analysis of FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Proof. The term $\sum_{t=1}^T f_t(\mathbf{x}_t)$ appears at both side of the equality, thus we verify

$$-\sum_{t=1}^T f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \right) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}).$$

Recall that $F_1(\mathbf{x}_1) = \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x})$, telescoping over $\sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \right)$

$$\sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \right) = F_1(\mathbf{x}_1) - F_{T+1}(\mathbf{x}_{T+1})$$

$$\begin{aligned} \Rightarrow -\sum_{t=1}^T f_t(\mathbf{u}) &= \psi_{T+1}(\mathbf{u}) - F_1(\mathbf{x}_1) + F_1(\mathbf{x}_1) - F_{T+1}(\mathbf{x}_{T+1}) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}) \\ &= \psi_{T+1}(\mathbf{u}) - F_{T+1}(\mathbf{u}), \end{aligned}$$

which is true by the definition of $F_{T+1}(\mathbf{x}) \triangleq \psi_{T+1}(\mathbf{x}) + \sum_{s=1}^T f_s(\mathbf{x})$. \square

General Analysis of FTRL

Lemma 4 (FTRL Regret). We denote that $F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$. Thus, the FTRL algorithm runs $\mathbf{x}_t = \arg \min_{\mathbf{x} \in \mathcal{X}} F_t(\mathbf{x})$. Then, for any $\mathbf{u} \in X$, we have

$$\begin{aligned} \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) &= \underbrace{\psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x})}_{\text{(range term)}} \\ &\quad + \underbrace{\sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \right)}_{\text{(stability term)}} \\ &\quad + \underbrace{F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})}_{\substack{(\mathbf{x}_{T+1} = \arg \min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \\ \text{thus } \leq 0)}} \end{aligned}$$

- The first and third terms are similar to those in OMD regret analysis.
- The second term is the **stability term**, which is crucial for the regret analysis, and we will explain why it's called stability later.

FTRL Stability

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Lemma 5 (FTRL Stability). Assume that ψ_t is λ_t -strongly convex w.r.t. $\|\cdot\|$. Then, the FTRL update satisfies

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \leq \frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}).$$

Proof. $F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)$

$$= F_t(\mathbf{x}_t) + f_t(\mathbf{x}_t) - (\textcolor{red}{F}_t(\mathbf{x}_{t+1}) + \textcolor{red}{f}_t(\mathbf{x}_{t+1})) + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1})$$

$$\leq \langle \nabla F_t(\mathbf{x}_t) + \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad (\text{strong convexity})$$

$$\leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad (\mathbf{x}_t = \arg \min_{\mathbf{x} \in \mathcal{X}} F_t(\mathbf{x}))$$

FTRL Stability

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Lemma 5 (FTRL Stability). Assume that ψ_t is λ_t -strongly convex w.r.t. $\|\cdot\|$. Then, the FTRL update satisfies

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \leq \frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}).$$

Proof. $F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)$

$$\leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1})$$

$$\leq \|\nabla f_t(\mathbf{x}_t)\|_* \cdot \|\mathbf{x}_t - \mathbf{x}_{t+1}\| - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad (\text{Hölder's inequality})$$

$$\leq \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad \square \quad (ab \leq \frac{a^2}{\lambda} + \frac{\lambda}{4} b^2)$$

Regret Bound for FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Theorem 6 (Regret Bound for FTRL). Assume $\psi_t(\mathbf{x})$ is λ_t -strongly convex on domain \mathcal{X} w.r.t. $\|\cdot\|$. We further assume that $\psi_t(\mathbf{x}) \leq \psi_{t+1}(\mathbf{x})$ for $t \in [T]$. Then, for FTRL satisfies

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^T \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \sum_{t=2}^T \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

Proof.
$$\sum_{t=1}^T (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) = \underbrace{\psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x})}_{\text{(range term)}} + \underbrace{\sum_{t=1}^T \left(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \right)}_{\text{(stability term)}} + \underbrace{F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})}_{\substack{(\mathbf{x}_{T+1} = \arg \min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \\ \text{thus } \leq 0)}}$$

Regret Bound for FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Theorem 6 (Regret Bound for FTRL). Assume $\psi_t(\mathbf{x})$ is λ_t -strongly convex on domain \mathcal{X} w.r.t. $\|\cdot\|$. We further assume that $\psi_t(\mathbf{x}) \leq \psi_{t+1}(\mathbf{x})$ for $t \in [T]$. Then, for FTRL satisfies

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \leq \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^T \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \sum_{t=2}^T \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

Proof.

$$\begin{aligned} \sum_{t=1}^T (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) &\stackrel{\text{(stability)}}{\leq} \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^T \left(\frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \right) \\ &\leq \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^T \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \sum_{t=2}^T \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \quad \square \end{aligned}$$

FTRL can be equivalent to OMD

Claim 1. Under online linear optimization (OLO) setting, with the same constant step size $\eta > 0$ and the same regularizer ψ (which is required to be *strongly convex* and a *barrier* function over \mathcal{X}), the OMD and FTRL algorithms **share the same output**:

$$\mathbf{x}_t = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^{t-1} \langle \eta \mathbf{g}_s, \mathbf{x} \rangle + \psi(\mathbf{x}) \right\},$$

and

$$\mathbf{x}_t = \arg \min_{\mathbf{x} \in \mathcal{X}} \{ \langle \eta \mathbf{g}_{t-1}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_{t-1}) \}.$$

FTRL vs. OMD: Equivalence Condition

Proof. For OMD, taking the gradient and setting it to 0 will lead to:

$$\eta \mathbf{g}_{t-1} + \nabla \psi(\mathbf{x}_t) - \nabla \psi(\mathbf{x}_{t-1}) = 0 \quad (\text{due to the barrier property of } \psi)$$

Telescoping from 1 to $t - 1$, and define $\mathbf{x}_0 \triangleq \arg \min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x})$,

$$\nabla \psi(\mathbf{x}_t) = -\eta \sum_{s=1}^{t-1} \mathbf{g}_s$$

On the other hand, for FTRL, setting the gradient to zero will lead to:

$$\nabla \psi(\mathbf{x}_t) = -\eta \sum_{s=1}^{t-1} \mathbf{g}_s \quad \square$$

FTRL as Dual Averaging

- Mirror Descent

$$\begin{aligned}\nabla\psi_t(\mathbf{y}_{t+1}) &= \nabla\psi_t(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})\end{aligned}$$

- Dual Averaging (**lazy** mirror descent)

$$\begin{aligned}\nabla\psi_t(\mathbf{y}_{t+1}) &= \nabla\psi_t(\mathbf{y}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \quad \text{averaging updates in dual space} \\ \mathbf{x}_{t+1} &= \arg \min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})\end{aligned}$$

$$\Rightarrow \mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta \sum_{s=1}^{t-1} \langle \nabla f_s(\mathbf{x}_s), \mathbf{x} \rangle + \psi(\mathbf{x}) \right\} \quad \begin{array}{l} \text{this is FTRL update} \\ \text{(consider fixed step size} \\ \text{for simplicity)} \end{array}$$

FTRL as Dual Averaging

Dual Averaging Method for Regularized Stochastic Learning and Online Optimization

Part of [Advances in Neural Information Processing Systems 22 \(NIPS 2009\)](#)

Bibtex

Metadata

Paper

Authors

Lin Xiao

Abstract

We consider regularized stochastic learning and online optimization problems, where the objective function is the sum of two convex terms: one is the loss function of the learning task, and the other is a simple regularization term such as L1-norm for sparsity. We develop a new online algorithm, the regularized dual averaging method, that can explicitly exploit the regularization structure in an online setting. In particular, at each iteration, the learning variables are adjusted by solving a simple optimization problem that involves the running average of all past subgradients of the loss functions and the whole regularization term, not just its subgradient. This method achieves the optimal convergence rate and often enjoys a low complexity per iteration. Computational experiments are provided for learning using L1-regularization.

**NIPS 2019 ten-year
Test of Time Award!**

Lin Xiao. Dual Averaging Method for Regularized Stochastic Learning and Online Optimization. NIPS 2009.

Math. Program., Ser. B (2009) 120:221–259
DOI 10.1007/s10107-007-0149-x

FULL LENGTH PAPER

Primal-dual subgradient methods for convex problems

Yurii Nesterov

Received: 29 September 2005 / Accepted: 13 January 2007 / Published online: 19 June 2007
© Springer-Verlag 2007

Abstract In this paper we present a new approach for constructing subgradient schemes for different types of nonsmooth problems with convex structure. Our methods are primal-dual since they are always able to generate a feasible approximation to the optimum of an appropriately formulated dual problem. Besides other advantages, this useful feature provides the methods with a reliable stopping criterion. The proposed schemes differ from the classical approaches (divergent series methods, mirror descent methods) by presence of two control sequences. The first sequence is responsible for aggregating the support functions in the dual space, and the second one establishes a dynamically updated scale between the primal and dual spaces. This additional flexibility allows to guarantee a boundedness of the sequence of primal test points even in the case of unbounded feasible set (however, we always assume the uniform boundedness of subgradients). We present the variants of subgradient schemes for nonsmooth convex minimization, minimax problems, saddle point problems, variational inequalities, and stochastic optimization. In all situations our methods are proved to be optimal from the view point of worst-case black-box lower complexity bounds.

Dedicated to B. T. Polyak on the occasion of his 70th birthday

Y. Nesterov. Primal-dual subgradient methods for convex problems, 2005.

1 Introduction

1.1 Prehistory

The results presented in this paper are not very new. Most of them were obtained by the author in 2001–2002. However, a further purification of the developed framework led to rather surprising results related to the smoothing technique. Namely, in [11] it was shown that many nonsmooth convex minimization problems with an appropriate

At that moment of time, the author got an illusion that the importance of black-box approach in Convex Optimization will be irreversibly vanishing, and, finally, this approach will be completely replaced by other ones based on a clever use of problem's structure (interior-point methods, smoothing, etc.). This explains why the results included in this paper were not published at time. However, the developments of the last years clearly demonstrated that in some situations the black-box methods are irreplaceable. Indeed, the structure of a convex problem may be too complex for constructing a good self-concordant barrier or for applying a smoothing technique. Note also, that optimization schemes sometimes are employed for modelling certain *adjustment processes* in real-life systems. In this situation, we are not free in selecting the type of optimization scheme and in the choice of its parameters. However, the results on convergence and the rate of convergence of corresponding methods remain interesting.



Yurii Nesterov
1956 –
UCLouvain, Belgium

FTRL vs. OMD

- FTRL and OMD frameworks can recover different OCO methods.
- They share many similarities in both algorithm and regret, but they are *fundamentally different* in essence, especially when the step size scheduling is time-varying.
- The dynamics of FTRL and OMD also exhibits great difference when considering beyond static regret minimization, such as in dynamic regret minimization, or repeated game convergence.

Congrats to Nemirovski and Nesterov



2023年世界顶尖科学家协会奖“智能科学或数学奖”获奖者
THE 2023 WLA PRIZE LAUREATES IN COMPUTER SCIENCE OR MATHEMATICS



阿尔卡迪·涅米罗夫斯基

Arkadi NEMIROVSKI

美国佐治亚理工学院



尤里·涅斯捷罗夫

Yurii NESTEROV

比利时法语鲁汶大学

表彰他们在凸优化理论方面的一系列开创性工作, 包括自协调函数和
内点法的理论、优化的复杂性理论、加速梯度算法设计以及在鲁棒优化方面的方法论进展等。

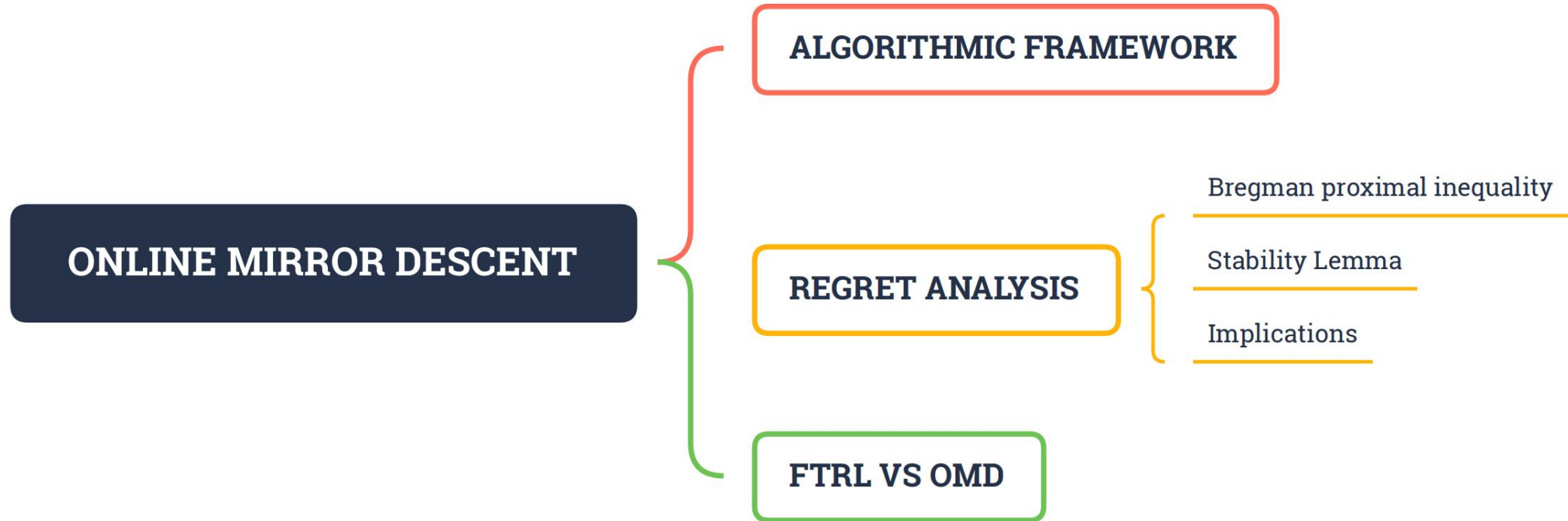
Congrats to WLA Prize (actually)



2023年顶科协奖“智能科学或数学奖”

2023年11月6日

Summary



Q & A

Thanks!