南京大学
人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

Lecture 8. Online Mirror Descent

Advanced Optimization (Fall 2022)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- OGD and Hedge
- Online Mirror Descent: View I
- Online Mirror Descent: View II
- Follow-the-Regularized-Leader (FTRL)

PEA vs. OCO

Prediction with Expert Advice

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

(1) the player first picks a weight $\boldsymbol{p}_t$ from a simplex Δ_N;

(2) and simultaneously environments pick an loss vector $\boldsymbol{\ell}_t \in \mathbb{R}^N$;

(3) the player suffers loss $f_t(\boldsymbol{p}_t) \triangleq \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, observes $\boldsymbol{\ell}_t$ and updates the model.

require domain to be a simplex $\mathcal{X} = \Delta_N$ $\Uparrow$ linear loss $f_t(\mathbf{x}) \triangleq \langle \mathbf{x}, \boldsymbol{\ell}_t \rangle$

PEA is a ***special case*** of OCO!

Online Convex Optimization

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

(1) the player first picks a model $\mathbf{x}_t \in \mathcal{X}$;

(2) and simultaneously environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes f_t and updates the model.

Deploying OGD to PEA

- PEA is a special case of OCO:

Why not directly deploy OGD (proposed in last lecture) to address PEA?

> **Theorem 4** (Regret bound for OGD). *Under Assumption 1, 2 and 3, online gradient descent (OGD) with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ for $t \in [T]$ guarantees:*
>
> $$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2} GD\sqrt{T}.$$

Regret guarantee: $D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{2}$ $\quad G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}$

$$\Longrightarrow \text{Regret}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \min_{\boldsymbol{p}\in\Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t\rangle \leq \mathcal{O}(\sqrt{TN})$$

Deploying OGD to PEA

- OGD for PEA Problem:

$$D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{2} \qquad G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}$$

$$\Rightarrow \quad \text{Regret}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{p}\in\Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle \le \mathcal{O}(\sqrt{TN})$$

- A natural question: is the $\mathcal{O}(\sqrt{TN})$ regret bound tight enough?
  - recall that the lower bound of PEA is $\Omega(\sqrt{T \ln N})$
  - OGD is not optimal with respect to N (number of experts)

Deploying OGD to PEA

- PEA is a special case of OCO:

Why not directly deploy OGD (proposed in last lecture) to address PEA?

> **Theorem 4** (Regret bound for OGD). *Under Assumption 1, 2 and 3, online gradient descent (OGD) with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ for $t \in [T]$ guarantees:*
>
> $$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2} GD\sqrt{T}.$$

Regret guarantee: $D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{2}$ $\quad \boxed{G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}}$

$$\Longrightarrow \text{Regret}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \min_{\boldsymbol{p}\in\Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t\rangle \leq \mathcal{O}(\sqrt{TN})$$

Why OGD Fails for PEA

- PEA has a special structure whereas general OCO doesn't have.

Convex Problem	***PEA Problem***
Domain: convex set $\mathcal{X}$	Domain: simplex $\mathcal{X} = \Delta_N$
Online function: convex function f_t	Online function: linear $f_t(\boldsymbol{p}) \triangleq \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle$
Lower Bound: $\Omega(GD\sqrt{T})$	Lower Bound: $\Omega(\sqrt{T \ln N})$

Why OGD Fails for PEA

- Remember that for the general OCO, we linearized the function to analyze the first gradient descent lemma:

$$
\begin{aligned}
\left\|\mathbf{x}_{t+1}-\mathbf{x}^{\star}\right\|^{2} &= \left\|\Pi_{\mathcal{X}}\left[\mathbf{x}_{t}-\eta_{t} \nabla f\left(\mathbf{x}_{t}\right)\right]-\mathbf{x}^{\star}\right\|^{2} \quad \text{(GD)} \\
&\leq \left\|\mathbf{x}_{t}-\eta_{t} \nabla f\left(\mathbf{x}_{t}\right)-\mathbf{x}^{\star}\right\|^{2} \quad \text{(Pythagoras Theorem)} \\
&= \left\|\mathbf{x}_{t}-\mathbf{x}^{\star}\right\|^{2}-2 \eta_{t}\left\langle\nabla f\left(\mathbf{x}_{t}\right), \mathbf{x}_{t}-\mathbf{x}^{\star}\right\rangle+\eta_{t}^{2}\left\|\nabla f\left(\mathbf{x}_{t}\right)\right\|^{2} \\
&\leq \left\|\mathbf{x}_{t}-\mathbf{x}^{\star}\right\|^{2}-2 \eta_{t}\left(f\left(\mathbf{x}_{t}\right)-f^{\star}\right)+\eta_{t}^{2}\left\|\nabla f\left(\mathbf{x}_{t}\right)\right\|^{2} \\
&\quad \left(\text{convexity: } f\left(\mathbf{x}_{t}\right)-f^{\star}=f\left(\mathbf{x}_{t}\right)-f\left(\mathbf{x}^{\star}\right) \leq\left\langle\nabla f\left(\mathbf{x}_{t}\right), \mathbf{x}_{t}-\mathbf{x}^{\star}\right\rangle\right)
\end{aligned}
$$

- So, linearized loss is not the essence, but the ***simplex domain*** of the PEA problem is worthy specifically considering.

Why OGD Fails for PEA

- Recall that for general OCO, we update the model as follows:

General Online Convex Optimization

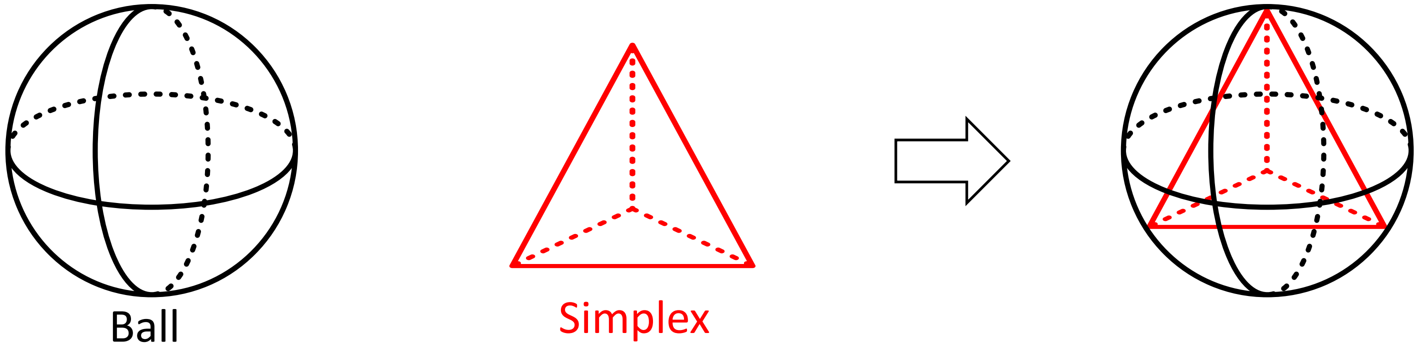
OGD:

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)\right]$$

Proximal type update:

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

- In PEA, is it proper to use 2-norm (ball) to measure distance?



A ball is too pessimistic (loose) to measure a simplex!

Why OGD Fails for PEA

- Recall that for general OCO, we update the model as follows:

General Online Convex Optimization

OGD:

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)\right]$$

Proximal type update:

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

- In PEA, is it proper to use 2-norm (ball) to measure distance?

$\Rightarrow$ We need to find an ***alternative distance measure*** for the ***special structure*** in PEA.

Reinvent Hedge Algorithm

$\Rightarrow$ We need to find an ***alternative distance measure*** for the ***special structure*** in PEA.

- Intuitively, for Euclidean space, 2-norm is the most natural measure:

$$\|\mathbf{x} - \mathbf{y}\|_2^2$$

- For PEA problem
  - the decision can be viewed as a distribution within the simplex
  - for two distributions P and Q, KL divergence is a natural measure:

$$\mathrm{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Reinvent Hedge Algorithm

Theorem 3. *Consider $f_t(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle$. An online learning algorithm that updates the model following*

$$\boldsymbol{p}_{t+1} = \arg\min_{\boldsymbol{p} \in \Delta_N} \eta \langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t) \rangle + \mathrm{KL}(\boldsymbol{p} \| \boldsymbol{p}_t)$$

is equal to Hedge update, i.e.,

$$\boldsymbol{p}_{t+1}(i) \propto \exp\left(-\eta L_t(i)\right) \text{ for all } i \in [N].$$

Proof.

$$\begin{aligned} \boldsymbol{p}_{t+1} &= \arg\min_{\boldsymbol{p} \in \Delta_N} \eta \langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t) \rangle + \mathrm{KL}(\boldsymbol{p} \| \boldsymbol{p}_t) \\ &= \arg\min_{\boldsymbol{p} \in \Delta_N} \underbrace{\eta \langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t) \rangle - \sum_{i=1}^N \boldsymbol{p}(i) \ln \left(\frac{\boldsymbol{p}_t(i)}{\boldsymbol{p}(i)} \right)}_{F(\boldsymbol{p})} \qquad \text{(definition of KL divergence)} \end{aligned}$$

Proof

Proof. $$\boldsymbol{p}_{t+1} = \arg\min_{\boldsymbol{p}\in\Delta_N} \underbrace{\eta\langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t)\rangle - \sum_{i=1}^{N} \boldsymbol{p}(i)\ln\left(\frac{\boldsymbol{p}_t(i)}{\boldsymbol{p}(i)}\right)}_{F(\boldsymbol{p})}$$ (definition of KL divergence)

$F(\boldsymbol{p})$ is convex, therefore we minimize $\boldsymbol{p}$ by taking $\nabla_{\boldsymbol{p}} F(\boldsymbol{p}) = 0$:

$$\forall i \in [N], \quad \eta(\nabla f_t(\boldsymbol{p}_t))_i - \ln(\boldsymbol{p}_t(i)) - 1 + \ln(\boldsymbol{p}_{t+1}(i)) = 0$$

$$\Rightarrow \quad \begin{aligned} \boldsymbol{p}_{t+1}(i) &= \exp\big(-\eta\nabla(f_t(\boldsymbol{p}_t))_i + \ln(\boldsymbol{p}_t(i)) + 1\big) \\ &= \boldsymbol{p}_t(i)\exp\big(-\eta\boldsymbol{\ell}_t(i) + 1\big) && (f_t(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{\ell}_t\rangle) \\ &= \boldsymbol{p}_{t-1}(i)\exp\big(-\eta(\boldsymbol{\ell}_t(i) + \boldsymbol{\ell}_{t-1}(i)) + 2\big) \\ &= \ldots \\ &= \boldsymbol{p}_0(i)\exp\big(-\eta L_t(i) + t\big) \end{aligned}$$

$\boldsymbol{p}_{t+1}(i) \propto \exp(-\eta L_t(i))$ for all $i \in [N]$

Reinvent Hedge Algorithm

- Proximal update rule for OGD:

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

- Proximal update rule for Hedge:

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

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

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

Online Mirror Descent Framework

Online Mirror Descent

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

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

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 a required to be strongly convex and differentiable function over a convex set $\mathcal{X}$.

Online Mirror Descent Framework

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 $\mathcal{D}_\psi$ associated to ψ is defined as

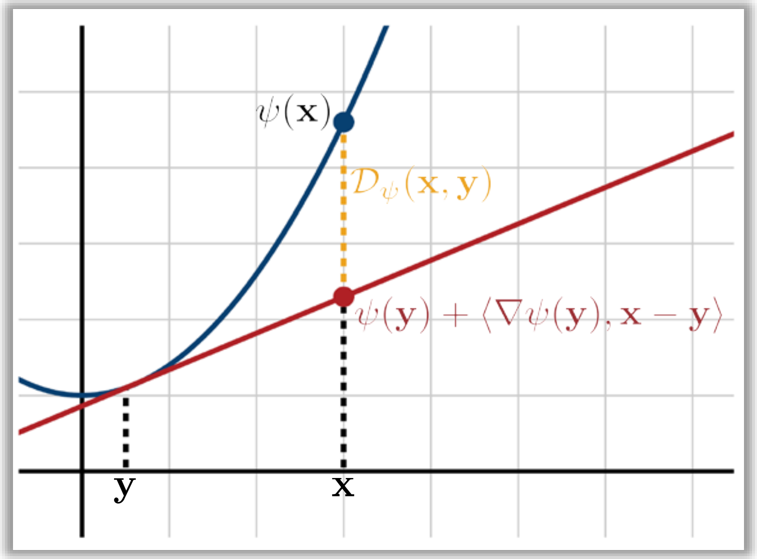
$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

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



Online Mirror Descent Framework

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 $\mathcal{D}_\psi$ 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	$\sum_i x_i \log x_i$	$\mathrm{KL}(\mathbf{x} \\| \mathbf{y})$

Online Mirror Descent Framework

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

	OMD form	Choice of $\mathcal{D}_\psi(\mathbf{x},\mathbf{y})$	η_t	Regret_T
OGD for convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\\|\mathbf{x}-\mathbf{x}_t\\|_2^2$	$\\|\mathbf{x}-\mathbf{y}\\|_2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\\|\mathbf{x}-\mathbf{x}_t\\|_2^2$	$\\|\mathbf{x}-\mathbf{y}\\|_2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{\log T}{\sigma})$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\\|\mathbf{x}-\mathbf{x}_t\\|_{A_t}^2$	$\\|\mathbf{x}-\mathbf{y}\\|_{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d\log T}{\alpha})$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \mathrm{KL}(\mathbf{x}\\|\mathbf{x}_t)$	$\mathrm{KL}(\mathbf{x}\\|\mathbf{y})$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T\ln N})$

Online Mirror Descent Framework

Online Mirror Descent

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

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

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 a required to be strongly convex and differentiable function over a convex set $\mathcal{X}$.

General Analysis Framework for OMD

Lemma 1 (Mirror Descent Lemma). *Let $\mathcal{D}_\psi$ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \to \mathbb{R}$ 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*

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

For simplicity, consider the ***fixed step size version***, we then have the following regret.

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*

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

OMD Analysis

Lemma 1 (Mirror Descent Lemma). *Let $\mathcal{D}_\psi$ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \to \mathbb{R}$ 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*

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

Proof.

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

We will introduce two lemmas to bound term (a) and term (b), respectively.

OMD Analysis

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

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

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

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

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

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

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

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

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

By the first-order optimality conditions of $\mathbf{x}_1$ and $\mathbf{x}_2$,

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

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

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

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

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

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

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

Summing them up, we get

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

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.

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

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

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

$$\leq \|\mathbf{x} - \mathbf{x}'\| \|\mathbf{g} - \mathbf{g}'\|_\star \quad \text{(Hölder's inequality)}$$

$$\Rightarrow \quad \lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star \qquad \square$$

OMD Analysis

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*

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

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:

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

Rearranging the terms can finish the proof. $\square$

OMD Analysis

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

Lemma 2 (Stability Lemma)**.**

$$\lambda \|\mathbf{x}_1 - \mathbf{x}_2\| \le \|\mathbf{g}_1 - \mathbf{g}_2\|_\star$$

$$\Rightarrow \quad \text{term (a)} = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \le \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2$$

Lemma 3 (Bregman Proximal Inequality)**.**

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

$$\Rightarrow \quad \text{term (b)} \le \frac{1}{\eta_t} \Big(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \Big)$$

(negative term, usually dropped)

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

A Comparison of Different Methods

- Mirror Descent Lemma of OMD:

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

- First Gradient Lemma of OGD:

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \frac{1}{2\eta_t}\left(\|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 - \|\mathbf{x}_t - \mathbf{u}\|_2^2\right) + \frac{\eta_t}{2}\|\nabla f_t(\mathbf{x}_t)\|_2^2$$

- First Gradient Lemma of ONS:

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \frac{\gamma}{2}\left(\|\mathbf{x}_t - \mathbf{u}\|_{A_t}^2 - \|\mathbf{x}_{t+1} - \mathbf{u}\|_{A_t}^2\right) + \frac{1}{2\gamma}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$
$$-\frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{u}\|_{\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top}^2$$

General Analysis Framework for OMD

Lemma 1 (Mirror Descent Lemma). *Let $\mathcal{D}_\psi$ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \to \mathbb{R}$ 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*

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

Using Lemma 1, we can easily prove the following 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*

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

General Analysis Framework 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*

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

Proof.

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

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

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

General Analysis Framework 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*

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

With this generic theorem, it will become straight-forward to recover previous results.

OMD Implication: Recovering OGD

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

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

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

Regret Analysis.

$$\Longrightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \sum_{t=1}^T \left(\frac{1}{\eta_t}\|\mathbf{u}-\mathbf{x}_t\|_2^2 - \frac{1}{\eta_t}\|\mathbf{u}-\mathbf{x}_{t+1}\|_2^2\right) + \sum_{t=1}^T \eta_t \|\nabla f_t(\mathbf{x}_t)\|_2^2$$

$$= \frac{1}{\eta_1}\|\mathbf{u}-\mathbf{x}_1\|_2^2 - \frac{1}{\eta_T}\|\mathbf{u}-\mathbf{x}_{T+1}\|_2^2 + \sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\|\mathbf{u}-\mathbf{x}_t\|_2^2 + \sum_{t=1}^T \frac{\eta_t}{\lambda}\|f_t(\mathbf{x}_t)\|_2^2$$

$$\le \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \sum_{t=1}^T \eta_t G^2 \quad \le \quad 3DG\sqrt{T} \qquad \left(\eta_t = \frac{D}{G\sqrt{t}} \text{ and } \sum_{t=1}^T \frac{1}{\sqrt{t}} \le 2\sqrt{T}\right) \qquad \square$$

OMD Implication: Recovering Hedge

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

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

- 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 $\mathbf{x}_1 = \{\frac{1}{N}, \ldots, \frac{1}{N}\}$

Regret Analysis.

$$\Longrightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \frac{\mathrm{KL}(\mathbf{u}\|\mathbf{x}_1)}{\eta} + \eta \sum_{t=1}^T \|\boldsymbol{\ell}_t\|_\infty^2 \le \frac{\ln N}{\eta} + \eta T$$

$(\mathrm{KL}(\mathbf{u}\|\mathbf{x}_1) \le \ln N, \forall \mathbf{u})$

$(\boldsymbol{\ell}_t(i) \le 1, \forall i \in [N])$

$\square$

Another View for Mirror Descent

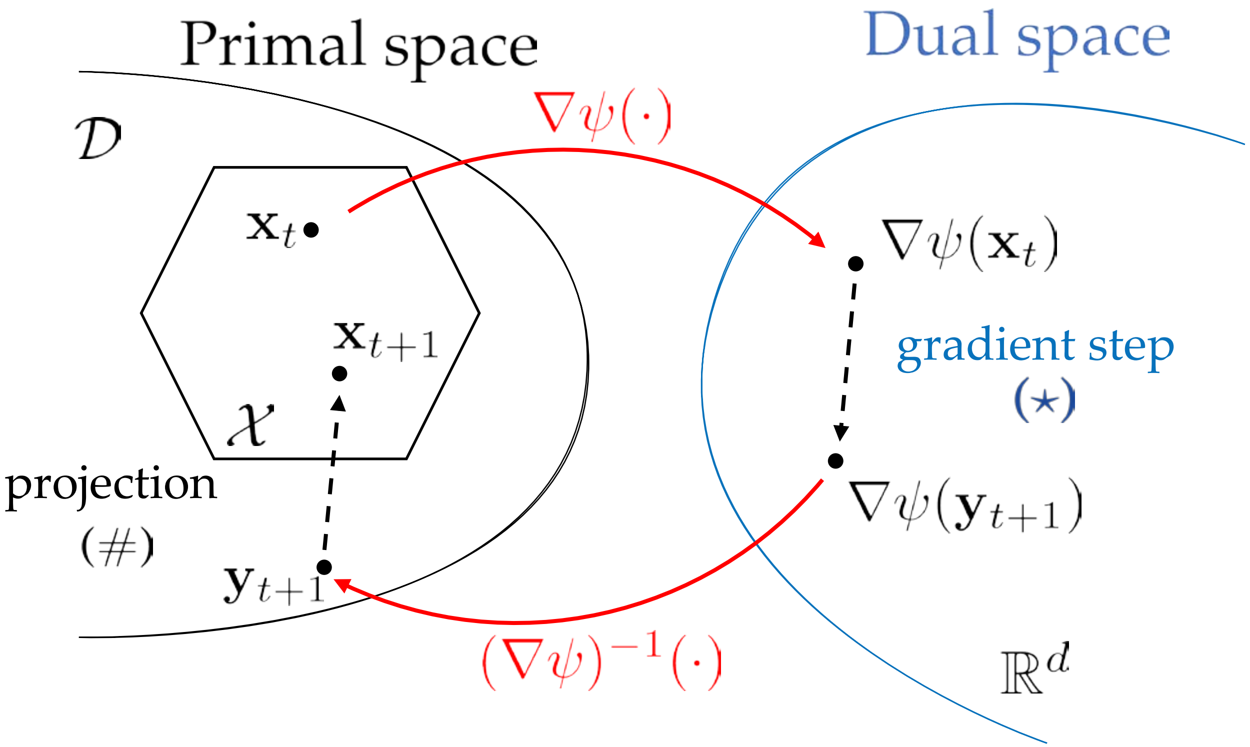
- Gradient Descent (GD) Method

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

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

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

Mirror Map



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

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

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

Mirror Map

Definition 2 (Mirror Maps). 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 $\psi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it safisfies the following properties 1 :

(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

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

Another View for Mirror Descent

Theorem 5. *The OMD update form*

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

is equal to the following two-step updates:

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

Proof.

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

Another View for Mirror Descent

Theorem 5. *The OMD update form*

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

is equal to the following two-step updates:

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

Proof.

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x}) - \langle \nabla\psi\left(\mathbf{y}_{t+1}\right), \mathbf{x}\rangle \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x}) - \langle \nabla\psi\left(\mathbf{x}_t\right) - \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle \qquad \text{(update of } \hat{\mathbf{x}}_{t+1}\text{)} \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \psi(\mathbf{x}) - \langle \nabla\psi\left(\mathbf{x}_t\right), \mathbf{x}\rangle \end{aligned}$$

Another View for Mirror Descent

Theorem 5. *The OMD update form*

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

is equal to the following two-step updates:

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

Proof.

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \left\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\right\rangle + \psi(\mathbf{x}) - \left\langle \nabla\psi\left(\mathbf{x}_t\right), \mathbf{x}\right\rangle \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \left\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\right\rangle + \psi(\mathbf{x}) - \psi(\mathbf{x}_t) - \left\langle \nabla\psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\right\rangle \\ &\qquad\qquad\qquad\qquad \text{(definition of Bregman divergence)} \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \left\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\right\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \end{aligned}$$

$\square$

Mirror Map

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

equivalent

$$\Longleftrightarrow$$

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

where $\nabla\psi^\star(\cdot)$ is the ***Fenchel Conjugate*** of $\nabla\psi(\cdot)$.

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

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

Mirror Map

$$\boxed{\begin{aligned} \nabla\psi(\mathbf{y}_{t+1}) &= \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{aligned}} \quad \overset{\text{equivalent}}{\Longleftrightarrow} \quad \boxed{\begin{aligned} \mathbf{y}_{t+1} &= \nabla\psi^\star(\nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t)) \\ \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{aligned}}$$

where $\nabla\psi^\star(\cdot)$ is the ***Fenchel Conjugate*** of $\nabla\psi(\cdot)$.

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

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:

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

Mirror Map

$$\boxed{\begin{aligned}\nabla\psi\left(\mathbf{y}_{t+1}\right) &= \nabla\psi\left(\mathbf{x}_t\right) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi\left(\mathbf{x}, \mathbf{y}_{t+1}\right)\end{aligned}} \quad \overset{\text{equivalent}}{\Longleftrightarrow} \quad \boxed{\begin{aligned}\mathbf{y}_{t+1} &= \nabla\psi^\star(\nabla\psi\left(\mathbf{x}_t\right) - \eta_t \nabla f_t(\mathbf{x}_t)) \\ \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi\left(\mathbf{x}, \mathbf{y}_{t+1}\right)\end{aligned}}$$

where $\nabla\psi^\star(\cdot)$ is the ***Fenchel Conjugate*** of $\nabla\psi(\cdot)$.

Proof.

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

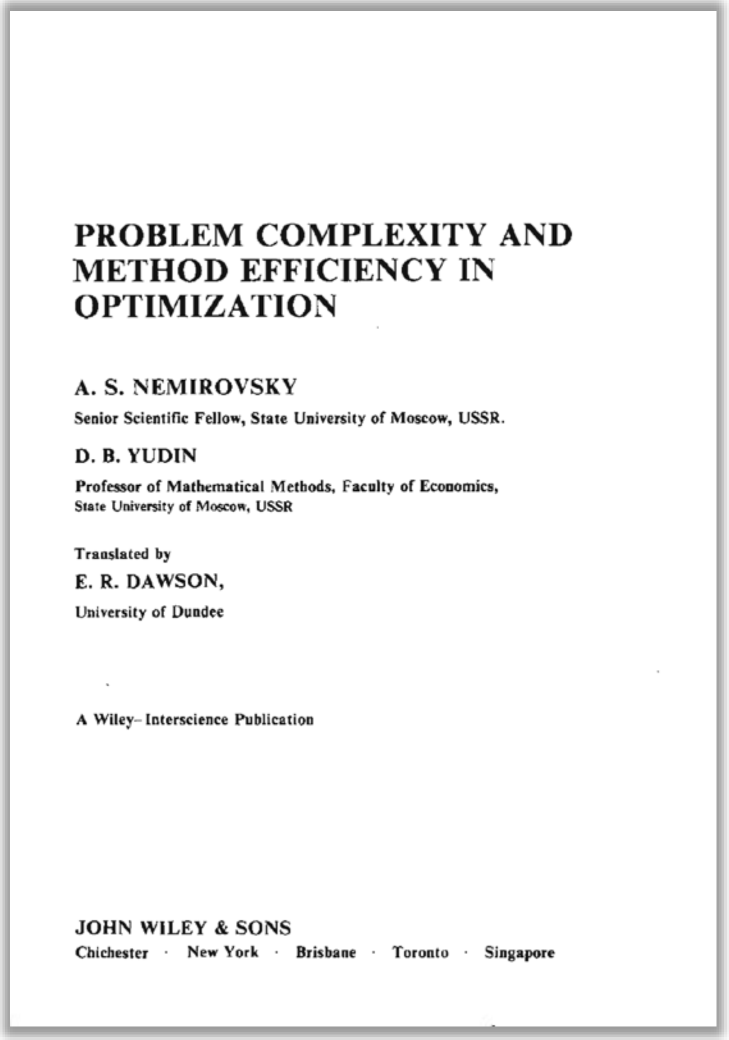
By taking the gradient w.r.t. $\mathbf{g}$ at both sides:

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

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

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

Mirror Descent: history bits



PROBLEM COMPLEXITY AND METHOD EFFICIENCY IN OPTIMIZATION

A. S. NEMIROVSKY
Senior Scientific Fellow, State University of Moscow, USSR.

D. B. YUDIN
Professor of Mathematical Methods, Faculty of Economics, State University of Moscow, USSR

Translated by
E. R. DAWSON,
University of Dundee

A Wiley–Interscience Publication

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

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)

FTL Idea: Select the expert that performs best so far, namely,

$$\boldsymbol{p}_t^{\mathrm{FTL}} = \operatorname*{arg\,min}_{\boldsymbol{p} \in \Delta_N} \langle \boldsymbol{p}, L_{t-1} \rangle$$

where $L_{t-1} \triangleq \sum_{s=1}^{t-1} \boldsymbol{\ell}_s \in \mathbb{R}^N$ is the cumulative loss vector.

- But, FTL is (highly) ***sub-optimal*** due to its ***unstable*** nature.

$\Longrightarrow$ a natural idea: adding regularizers to stabilize the algorithm.

Another OCO Framework: FTRL

$\Longrightarrow$ a natural idea: adding regularizers to stabilize the algorithm.

Follow The Regularized Leader (FTRL)

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

where ψ_t is the regularizer at time t.

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*

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) = \boxed{\psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x})} \quad \text{(range term)}$$

$$\boxed{+ \sum_{t=1}^T \left[F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)\right]} \quad \text{(stability term)}$$

$$\boxed{+ F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})} \quad (\mathbf{x}_{T+1} = \arg\min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \text{ thus} \leq 0)$$

General Analysis of FTRL

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

Lemma 4.

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

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

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

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

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

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

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

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

$$= \psi_{T+1}(\mathbf{u}) - F_{T+1}(\mathbf{u}),$$

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

General Analysis of FTRL

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

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*

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

- The stability term is crucial for regret analysis
- We will explain why it's called stability term 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, we have*

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

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

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

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

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

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, we have*

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

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

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

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

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

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 algorithm*

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

Proof.

$$\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) = \boxed{\psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x} \in \mathcal{X}} \psi_1(\mathbf{x})}$$

(range term)

$$\boxed{+ \sum_{t=1}^{T} [F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)]}$$

(stability term)

$$\boxed{+ F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})}$$

($\mathbf{x}_{T+1} = \arg\min_{\mathbf{x}} F_{T+1}(\mathbf{x})$, thus ≤ 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 algorithm*

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

Proof.

$$\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \leq \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^{T}\left[\overset{\text{(stability)}}{\frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t}} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1})\right]$$

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

Common Form for FTRL Regret

- FTRL updates by

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x}) = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{1}{\eta_{t-1}} \psi(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x}) \right\}$$

$$\boxed{F_t(\mathbf{x}) \triangleq \frac{1}{\eta_{t-1}} \psi(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})}$$

Theorem 6 (Regret Bound for FTRL). *Assume $\psi_t(\mathbf{x}) = \frac{1}{\eta_{t-1}} \psi(\mathbf{x})$, and $\psi(\mathbf{x})$ is 1-strongly convex on domain $\mathcal{X}$ w.r.t. $\|\cdot\|$. We further assume a decreasing step size sequence (i.e., $\eta_t \geq \eta_{t+1}$ for $t \in [T]$). Then, FTRL enjoys*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\psi(\mathbf{u})}{\eta_T} + \sum_{t=1}^{T} \eta_{t-1} \|\nabla f_t(\mathbf{x}_t)\|_\star^2$$

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

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

FTRL vs. OMD: Equal Condition

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

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

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:

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

FTRL as Dual Averaging

- Mirror Descent

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

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

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

averaging updates in dual space

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

this is the FTRL update (consider fixed step size for simplicity)

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 ite
gradient method. Computational experiments are p
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

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 framework can recover different OCO methods.

- They share many similarities in both algorithm and regret, but they are *fundamentally different* in essence.

FTRL vs. OMD: Update Styles

- OMD update style:

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

→ OMD only depends on the last iteration (only keep the last result)

- FTRL update style:

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

→ FTRL is more informative (keep all the history data) and more sensitive

FTRL vs. OMD: Regret Bound

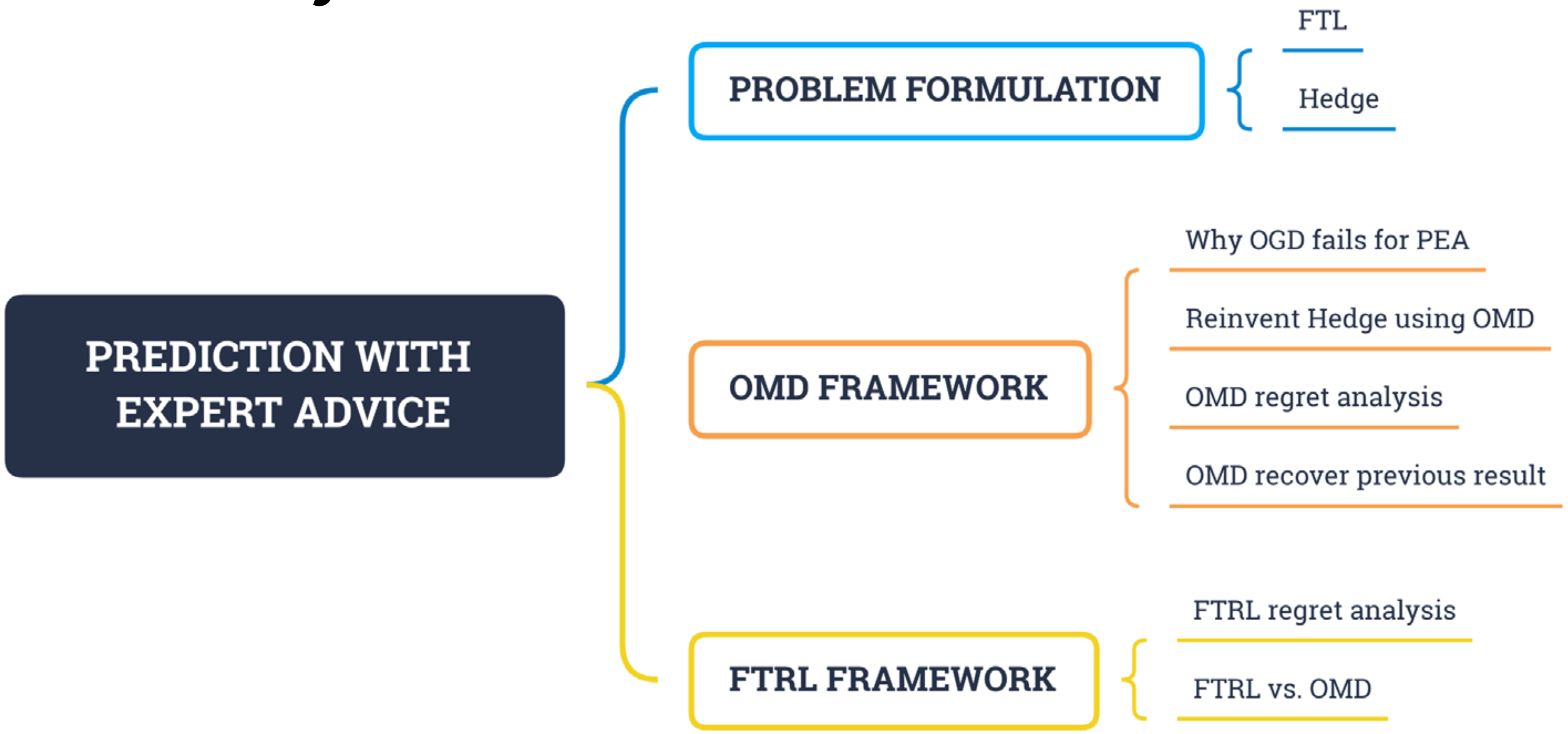
- OMD Regret:

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

- FTRL Regret:

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\psi(\mathbf{u})}{\eta_T} + \sum_{t=1}^{T} \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right) \psi(\mathbf{x}_{t+1}) + \sum_{t=1}^{T} \eta_{t-1} \left\| \nabla f_t(\mathbf{x}_t) \right\|_\star^2$$

Summary



Q & A

Thanks!