Lecture 7. Online Mirror Descent

Advanced Optimization (Fall 2025)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Online Mirror Descent
- General Regret Analysis
- Primal-Dual View
- Follow-the-Regularized Leader

Part 1. OMD Framework

- Motivation

- Algorithmic Framework

- Primal-Dual Interpretation

Recap: Prediction with Expert Advice

- The online learner (player) aims to make the prediction based by combining N experts' 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 a 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.

The feasible domain is the $(N-1)$-dim simplex $\Delta_N = \left\{\boldsymbol{p} \in \mathbb{R}^N \mid p_i \geq 0, \sum_{i=1}^N p_i = 1\right\}$.

We typically assume that $0 \leq \ell_{t,i} \leq 1$ holds for all $t \in [T]$ and $i \in [N]$.

Recap: Hedge Algorithm

- Hedge: replacing the "*max*" operation in FTL by "***softmax***".

```
At each round $t = 1, 2, \cdots$
    (1) compute $\boldsymbol{p}_t \in \Delta_N$ such that $\boldsymbol{p}_{t,i} \propto \exp(-\eta L_{t-1,i})$ for $i \in [N]$
    (2) the player submits $\boldsymbol{p}_t$, suffers loss $\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, and observes loss $\boldsymbol{\ell}_t \in \mathbb{R}^N$
    (3) update $\boldsymbol{L}_t = \boldsymbol{L}_{t-1} + \boldsymbol{\ell}_t$
```

FTL update

$$\boldsymbol{p}_t^{\text{FTL}} = \operatorname*{arg\,max}_{\boldsymbol{p} \in \Delta_N} \langle \boldsymbol{p}, -\boldsymbol{L}_{t-1} \rangle$$

Hedge update

$$p_{t,i} \propto \exp(-\eta L_{t-1,i}), \forall i \in [N]$$

Recap: Regret Bound

Theorem 2. *Suppose that* $\forall t \in [T]$ *and* $i \in [N], 0 \leq \boldsymbol{\ell}_{t,i} \leq 1$*, then Hedge with learning rate* η *guarantees*

$$\textsc{Reg}_T \leq \frac{\ln N}{\eta} + \eta T = \mathcal{O}\big(\sqrt{T \log N}\big),$$

where the last equality is by setting η *optimally as* $\sqrt{(\ln N)/T}$.

Theorem 3 (Lower Bound of PEA)**.** *For any algorithm* $\mathcal{A}$*, we have that*

$$\sup_{T,N} \max_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \frac{\textsc{Reg}_T}{\sqrt{T \ln N}} \geq \frac{1}{\sqrt{2}}.$$

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 $\mathcal{X} = \Delta_N$ ⇑ 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 2 (Regret bound for OGD). *Under Assumption 1 (G-Lipschitz) and Assumption 2 (D-bounded domain), online gradient descent (OGD) with step sizes $\eta_t = \frac{D}{G\sqrt{t}}$ for $t \in [T]$ guarantees:*

$$\mathrm{REG}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2} GD\sqrt{T} = \mathcal{O}(\sqrt{T}).$$

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

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

Deploying OGD to PEA

- OGD for PEA Problem:

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

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

- Notice: the T-dependence is good, but N-dependence is suboptimal
  - OGD (using gradient update) is good to some extent
  - but OGD is not optimal with respect to N (number of experts), recall that the lower bound of PEA is $\Omega(\sqrt{T\log N})$
  - moreover, Hedge algorithm (designed for PEA) can achieve $O(\sqrt{T\log N})$

Deploying OGD to PEA

- PEA is a special case of OCO:

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

Theorem 2 (Regret bound for OGD). *Under Assumption 1 (G-Lipschitz) and Assumption 2 (D-bounded domain), online gradient descent (OGD) with step sizes* $\eta_t = \frac{D}{G\sqrt{t}}$ *for* $t \in [T]$ *guarantees:*

$$\textsc{Reg}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2} GD\sqrt{T} = \mathcal{O}(\sqrt{T}).$$

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

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

Why OGD Fails for PEA

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

Convex Problem

Domain: convex set $\mathcal{X}$

Online function: convex function f_t

Lower Bound: $\Omega(GD\sqrt{T})$

PEA Problem

Domain: simplex $\mathcal{X} = \Delta_N$

Online function: linear $f_t(\boldsymbol{p}) \triangleq \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle$

Lower Bound: $\Omega(\sqrt{T \log N})$

Why OGD Fails for PEA

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

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

(convexity: $f(\mathbf{x}_t) - f^\star = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle$)

- 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

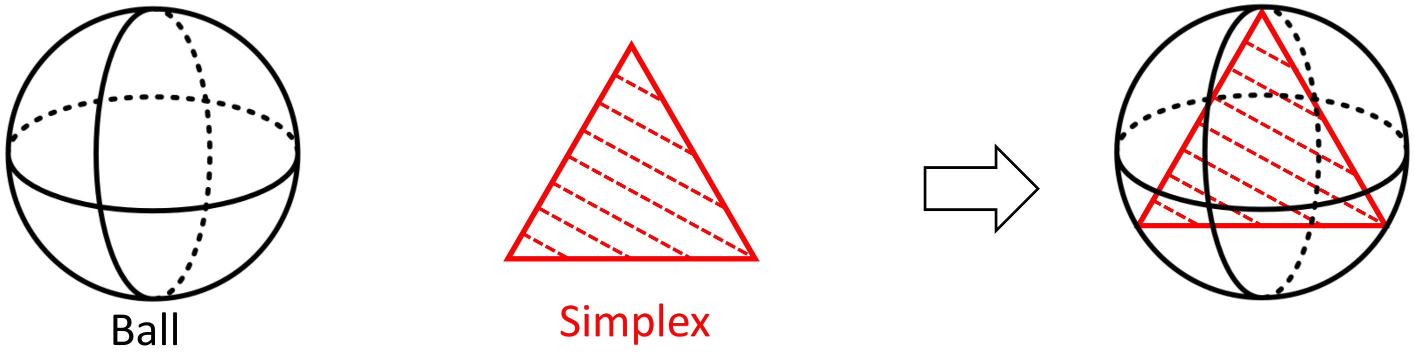
OGD:

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

Proximal type update:

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

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



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

Proximal View

- 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:

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

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

⇨ We need an alternative distance measure for the special ***geometry*** in PEA.

Geometry Matters

⇨ We need an alternative distance measure for the special ***geometry*** in PEA.

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

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

- For PEA problem: *geometry of the feasible domain* Δ_N
  - 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*

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

is equal to the Hedge algorithm (greedy update version), i.e.,

$$p_{t+1,i} \propto p_{t,i} \exp\left(-\eta \ell_{t,i}\right) \text{ for all } i \in [N].$$

Proof.

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

□

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:

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

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

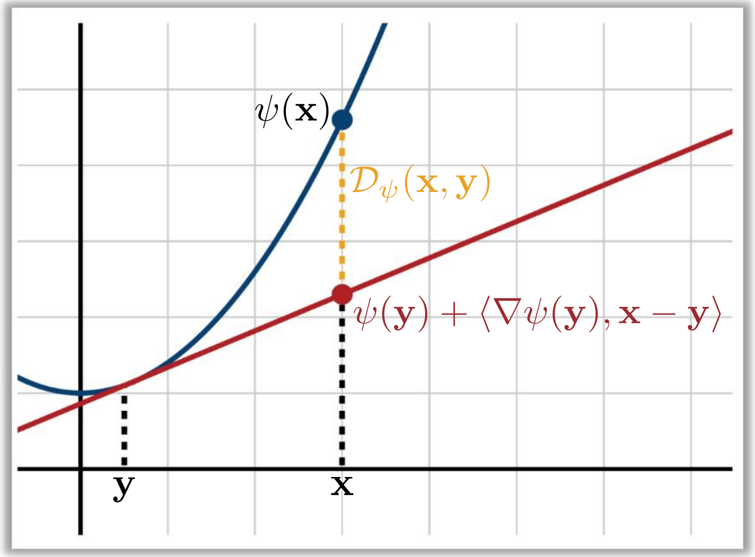
$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

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


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

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 $\mathcal{D}_\psi$ 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

Online Mirror Descent

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

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

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 a 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 defintion) to ensure the solutions' feasibility in domain $\mathcal{X}$.

Online Mirror Descent

Mirror descent: simultaneously considering the "gradient update" (using $\nabla f_t(\mathbf{x}_t)$) and "geometry" (using ψ)

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

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.

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

- 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	$\psi(\cdot)$
OGD	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2}\Vert\mathbf{x} - \mathbf{x}_t\Vert_2^2 \right\}$	$\frac{1}{2}\Vert\mathbf{x}\Vert_2^2$
Hedge	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathrm{KL}(\mathbf{x}\Vert\mathbf{x}_t) \right\}$	$\sum_{i=1}^{N} x_i \log x_i$

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

Recap: ONS in a view of Proximal Gradient

Convex Problem

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

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

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

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

Proximal type update:

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

Online Mirror Descent

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

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	$\textsc{Reg}_T$
OGD for convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\lVert\mathbf{x}-\mathbf{x}_t\rVert_2^2$	$\frac{1}{2}\lVert\mathbf{x}\rVert_2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\lVert\mathbf{x}-\mathbf{x}_t\rVert_2^2$	$\frac{1}{2}\lVert\mathbf{x}\rVert_2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma}\log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\lVert\mathbf{x}-\mathbf{x}_t\rVert_{A_t}^2$	$\frac{1}{2}\lVert\mathbf{x}\rVert_{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma}\log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \mathrm{KL}(\mathbf{x}\Vert\mathbf{x}_t)$	$\sum_{i=1}^{N} x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T\log N})$

Online Mirror Descent

- Can their regret analysis also be unified by OMD?

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

Part 2. General Regret Analysis

- Mirror Descent Lemma

- Stability Lemma

- Bregman Proximal Inequality

General Regret Analysis for OMD

Online Mirror Descent

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

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*

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

bias term (range term)

variance term (stability term)

negative term

Proof of Mirror Descent Lemma

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*

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

Proof.

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

We use ***stability lemma*** to analyze term (a), and use ***Bregman proximal inequality*** to analyze term (b).

Proof of Mirror Descent Lemma

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

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

Lemma 2 (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 $\mathbf{u} \in \mathcal{X}$:

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

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

Bregman Proximal Inequality

Lemma 2 (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 2 (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*

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

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

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

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

Rearranging the terms can finish the proof. □

Three-points identity.
For all $\mathbf{x} \in X$ and $\mathbf{y}, \mathbf{z} \in \operatorname{int} X$,

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

Proof of Mirror Descent Lemma

Proof.

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

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

Lemma 2 (Bregman Proximal Inequality).

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

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

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

(negative term, usually dropped; but sometimes highly useful)

Stability Lemma

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

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

Lemma 3 (Stability Lemma). *Consider the following mirror-descent updates:*

$$\begin{cases} \mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_1, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}_2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_2, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

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

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

where $\|\cdot\|_\star$ is the dual norm.

Stability Lemma

Lemma 3 (Stability Lemma). *Consider the following mirror-descent updates:*

$$\begin{cases} \mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_1, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}_2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_2, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

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

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

where $\|\cdot\|_\star$ is the dual norm.

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 $\mathbf{x}_2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\langle \mathbf{g}_2, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})\}$, we have

$$\langle \mathbf{g}_2 + \nabla\psi(\mathbf{x}_2) - \nabla\psi(\mathbf{c}), \mathbf{u} - \mathbf{x}_2\rangle \geq 0 \text{ holds for } \forall \mathbf{u} \in \mathcal{X}.$$

Stability Lemma

Lemma 3 (Stability Lemma). *Consider the following mirror-descent updates:*

$$\begin{cases} \mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_1, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}_2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_2, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

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

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

where $\|\cdot\|_\star$ is the dual norm.

Proof. $\langle \mathbf{g}_2 + \nabla\psi(\mathbf{x}_2) - \nabla\psi(\mathbf{c}), \mathbf{u} - \mathbf{x}_2\rangle \geq 0$ holds for $\forall \mathbf{u} \in \mathcal{X}$.

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

$$\langle \nabla\psi(\mathbf{x}_1) - \nabla\psi(\mathbf{c}) + \mathbf{g}_1, \mathbf{x}_2 - \mathbf{x}_1\rangle \geq 0$$

$$\langle \nabla\psi(\mathbf{x}_2) - \nabla\psi(\mathbf{c}) + \mathbf{g}_2, \mathbf{x}_1 - \mathbf{x}_2\rangle \geq 0$$

$$\Longrightarrow \quad \langle \mathbf{x}_2 - \mathbf{x}_1, \mathbf{g}_1 - \mathbf{g}_2\rangle \geq \langle \nabla\psi(\mathbf{x}_1) - \nabla\psi(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \quad (1)$$

Stability Lemma

Lemma 3 (Stability Lemma). *Consider the following mirror-descent updates:*

$$\begin{cases} \mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_1, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}_2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_2, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

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

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

where $\|\cdot\|_\star$ is the dual norm.

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

$$\langle \nabla\psi(\mathbf{x}_1), \mathbf{x}_1 - \mathbf{x}_2\rangle \geq \psi(\mathbf{x}_1) - \psi(\mathbf{x}_2) + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2$$

$$\langle \nabla\psi(\mathbf{x}_2), \mathbf{x}_2 - \mathbf{x}_1\rangle \geq \psi(\mathbf{x}_2) - \psi(\mathbf{x}_1) + \frac{\lambda}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2$$

Summing them up, we get

$$\langle \nabla\psi(\mathbf{x}_1) - \nabla\psi(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \geq \lambda\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \qquad (2)$$

Stability Lemma

Lemma 3 (Stability Lemma). *Consider the following mirror-descent updates:*

$$\begin{cases} \mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_1, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}_2 = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_2, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

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

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

where $\|\cdot\|_\star$ is the dual norm.

Proof.

$$\langle \mathbf{x}_2 - \mathbf{x}_1, \mathbf{g}_1 - \mathbf{g}_2\rangle \ge \langle \nabla\psi(\mathbf{x}_1) - \nabla\psi(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \quad (1)$$

$$\langle \nabla\psi(\mathbf{x}_1) - \nabla\psi(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \ge \lambda \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \quad (2)$$

⇨ $$\lambda \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \le \langle \nabla\psi(\mathbf{x}_1) - \nabla\psi(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \le \langle \mathbf{x}_2 - \mathbf{x}_1, \mathbf{g}_1 - \mathbf{g}_2\rangle$$

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

(Hölder's inequality)

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

Proof of Mirror Descent Lemma

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

We introduce the following lemma to analyze term (a).

Lemma 3 (Stability Lemma).

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

(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 $\mathbf{0}$)

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \eta_t \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\} \qquad \mathbf{x}_t = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \mathbf{0} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

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

Proof of Mirror Descent Lemma

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 (Bregman Proximal Inequality).

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

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

Lemma 3 (Stability Lemma).

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

⇨ $\text{term (a)} = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \le \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2$ (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 $\mathbf{0}$)

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

General Regret Analysis for OMD

Lemma 1 (Mirror Descent Lemma). *Let $\mathcal{D}_\psi$ 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*

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

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

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

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

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

□

regret bound for OMD w time-varying step size

General Regret Analysis for OMD

- OMD with Time-Varying Step Sizes

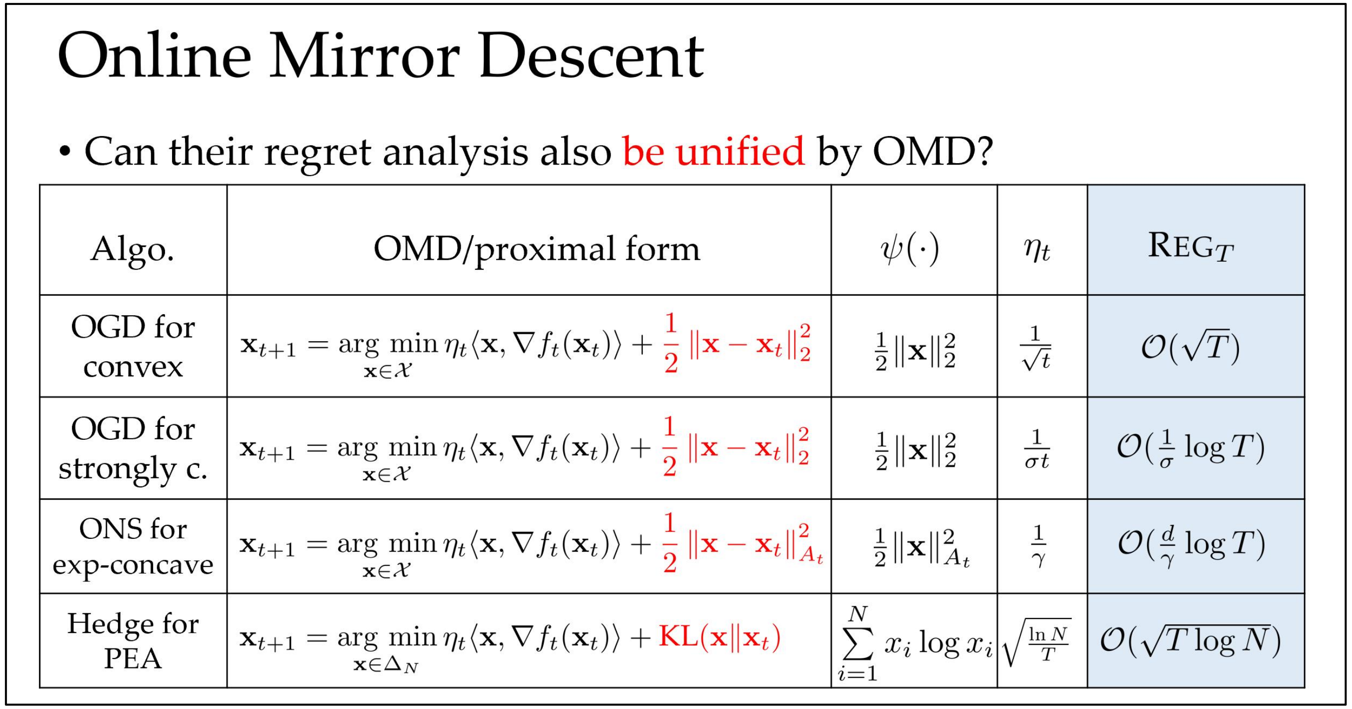
Theorem 5 (General Regret Bound for OMD with time-varying step sizes). *Assume ψ is λ-strongly convex w.r.t. $\|\cdot\|$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds*

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

Note: The time-varying step sizes may introduce some bothers to handle the bias term (especially for the third term on the right-hand side, but there are also many interesting (and tricky) usages, such as using increased step sizes to introduce negative regret…

General Regret Analysis for OMD

- With Theorem 4 and Theorem 5 (general regret bounds for OMD), it becomes more straightforward to **analyze** OGD/Hedge/ONS algorithms ***in a unified way***, which we previously analyzed the regret specifically for each algorithm.



Online Mirror Descent

- Can their regret analysis also be unified by OMD?

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

Implication: OGD for convex functions

Algorithm. It is straightforward to instantiate OMD to recover OGD:

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

- $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$ (dual norm still $\|\cdot\|_2$)
- step size is $\eta_t = \frac{D}{G\sqrt{t}}$

Regret Analysis. With Theorem 5 (dropping negative terms), we have

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

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

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

Implication: OGD for strongly convex functions

Algorithm. It is straightforward to instantiate OMD to recover OGD:

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

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

Regret Analysis. With Theorem 5 (dropping negative terms), we have

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

curvature-induced negative term due to strong convexity

$$= \frac{1}{2}\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma \right) \|\mathbf{u}-\mathbf{x}_t\|_2^2 + \frac{1}{2}\sum_{t=1}^T \eta_t G^2$$

$$= 0 + \frac{1}{2}\sum_{t=1}^T \frac{G^2}{\sigma t}$$

□ □

Implication: Hedge for PEA

Algorithm. With Theorem 4, it is straightforward to recover Hegde:

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

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

Regret Analysis.

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

$(\mathrm{KL}(\mathbf{u}\|\mathbf{x}_1) \le \ln N, \forall \mathbf{u})$ $\quad (\boldsymbol{\ell}_t(i) \le 1, \forall i \in [N])$

□

Implication: ONS for exp-concave functions

Algorithm. With Theorem 4, 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 $A_t = \varepsilon I + \sum_{s=1}^{t} \nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top$
- The dual norm of $\|\cdot\|_{A_t}$ is $\|\cdot\|_{A_t^{-1}}$

Regret Analysis.

$$\Longrightarrow \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \frac{\gamma}{2}\sum_{t=1}^{T}\left(\|\mathbf{u}-\mathbf{x}_t\|_{A_t}^2 - \|\mathbf{u}-\mathbf{x}_{t+1}\|_{A_t}^2 \underline{- \|\mathbf{u}-\mathbf{x}_t\|_{\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top}^2}\right) + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2 \quad \text{(exp-concavity)}$$

$$= \frac{\gamma}{2}\sum_{t=1}^{T}\left(\underline{\|\mathbf{u}-\mathbf{x}_t\|_{A_{t-1}}^2 - \|\mathbf{u}-\mathbf{x}_{t+1}\|_{A_t}^2}\right) + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2 \quad \text{(telescope)}$$

$$\le \frac{\gamma}{2}\|\mathbf{u}-\mathbf{x}_1\|_{A_0}^2 + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2 \qquad \square$$

A Summary of OMD Deployment

- Our previous algorithms and regret can all be covered by OMD.

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

A Case Study of PEA

- **OMD (with negative-entropy regularizer) ← Hedge**

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

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

$$\textsc{Reg}_T \le \frac{\mathrm{KL}(\mathbf{u}\|\mathbf{x}_1)}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}_t\|_\infty^2 \approx \frac{\ln N}{\eta} + \eta T \cdot 1 = \mathcal{O}(\sqrt{T\log N})$$

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

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

- **OMD (with Euclidean regularizer) ← OGD**

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

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

$$\textsc{Reg}_T \le \frac{\|\mathbf{u} - \mathbf{x}_1\|_2^2}{\eta} + \eta \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2^2 \approx \frac{1}{\eta} + \eta T N = \mathcal{O}(\sqrt{TN})$$

$$D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{2}$$

$$G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}$$

OMD with a suitable regularizer (better exploiting geometry) can achieve a better bias-variance trade off.

Part 3. Primal-Dual View

- Primal-Dual Interpretation

- Mirror Map

- History Bits

Online Mirror Descent

Mirror descent: simultaneously considering the "gradient update" (using $\nabla f_t(\mathbf{x}_t)$) and "geometry" (using ψ)

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

But what does "**mirror**" mean?

Another View for Mirror Descent

Theorem 5. *The OMD update form*

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

is equivalent to the following two-step updates:

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

Proof. $(\diamond)$

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

Another View for Mirror Descent

Theorem 5. *The OMD update form*

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

is equivalent to the following two-step updates:

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

Proof. $(\star)$

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

□

Another View for Mirror Descent

- Gradient Descent

$$\begin{cases} \mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \ \|\mathbf{x} - \mathbf{y}_{t+1}\|_2^2 \end{cases}$$

- A two-step update for mirror descent

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

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

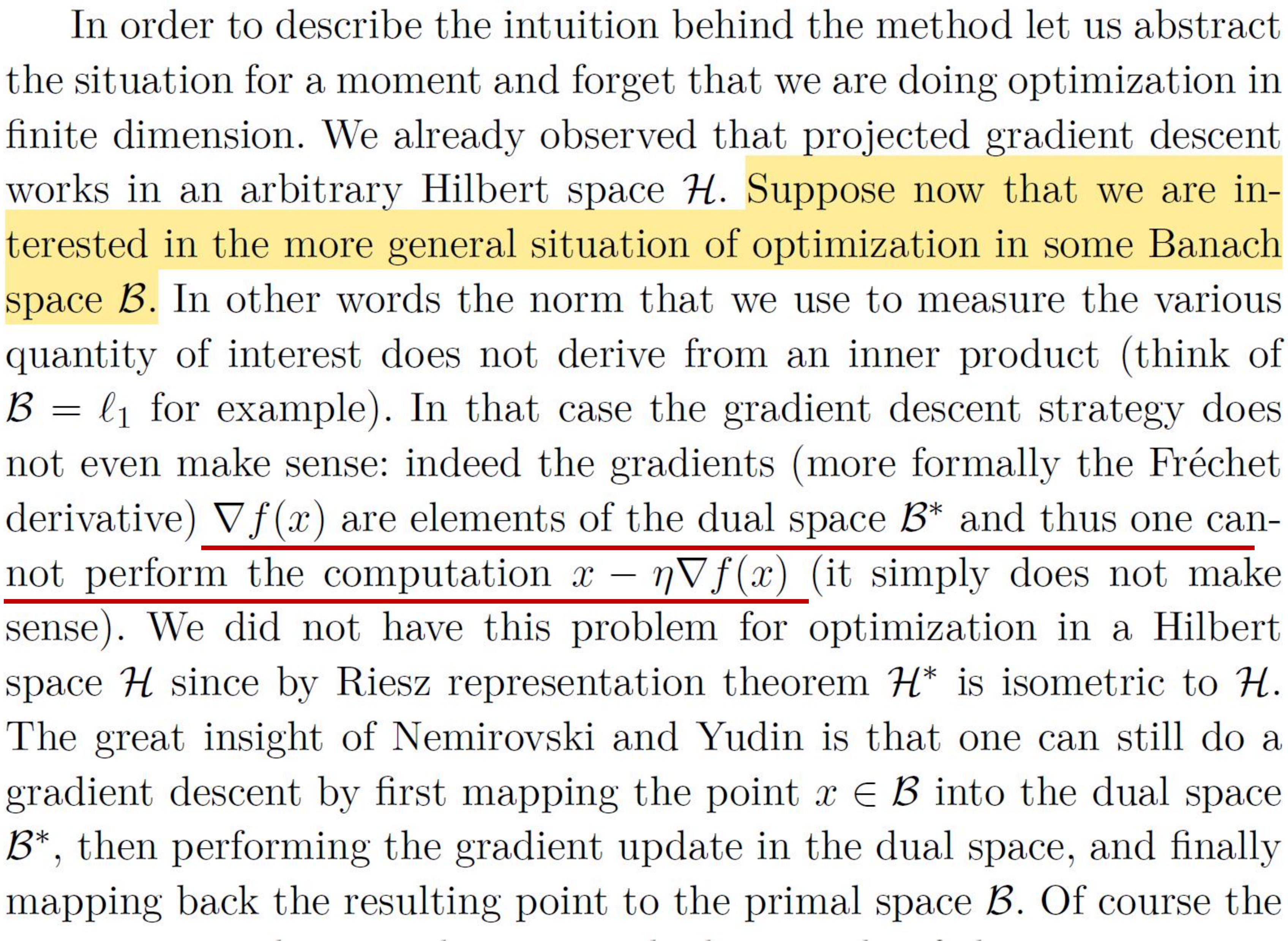
Primal-Dual View for Mirror Descent

- 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 $\mathcal{B}^*$



In order to describe the intuition behind the method let us abstract the situation for a moment and forget that we are doing optimization in finite dimension. We already observed that projected gradient descent works in an arbitrary Hilbert space $\mathcal{H}$. Suppose now that we are interested in the more general situation of optimization in some Banach space $\mathcal{B}$. In other words the norm that we use to measure the various quantity of interest does not derive from an inner product (think of $\mathcal{B} = \ell_1$ for example). In that case the gradient descent strategy does not even make sense: indeed the gradients (more formally the Fréchet derivative) $\nabla f(x)$ are elements of the dual space $\mathcal{B}^*$ and thus one cannot perform the computation $x - \eta \nabla f(x)$ (it simply does not make sense). We did not have this problem for optimization in a Hilbert space $\mathcal{H}$ since by Riesz representation theorem $\mathcal{H}^*$ is isometric to $\mathcal{H}$. The great insight of Nemirovski and Yudin is that one can still do a gradient descent by first mapping the point $x \in \mathcal{B}$ into the dual space $\mathcal{B}^*$, then performing the gradient update in the dual space, and finally mapping back the resulting point to the primal space $\mathcal{B}$. Of course the

Bubeck's Lecture Note, p297

Primal-Dual View for Mirror Descent

For a function $f : \mathcal{B} \to \mathbb{R}$, its Fréchet derivative at x is defined by

$$Df(\mathbf{x}) : \mathbf{h} \mapsto \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t}.$$

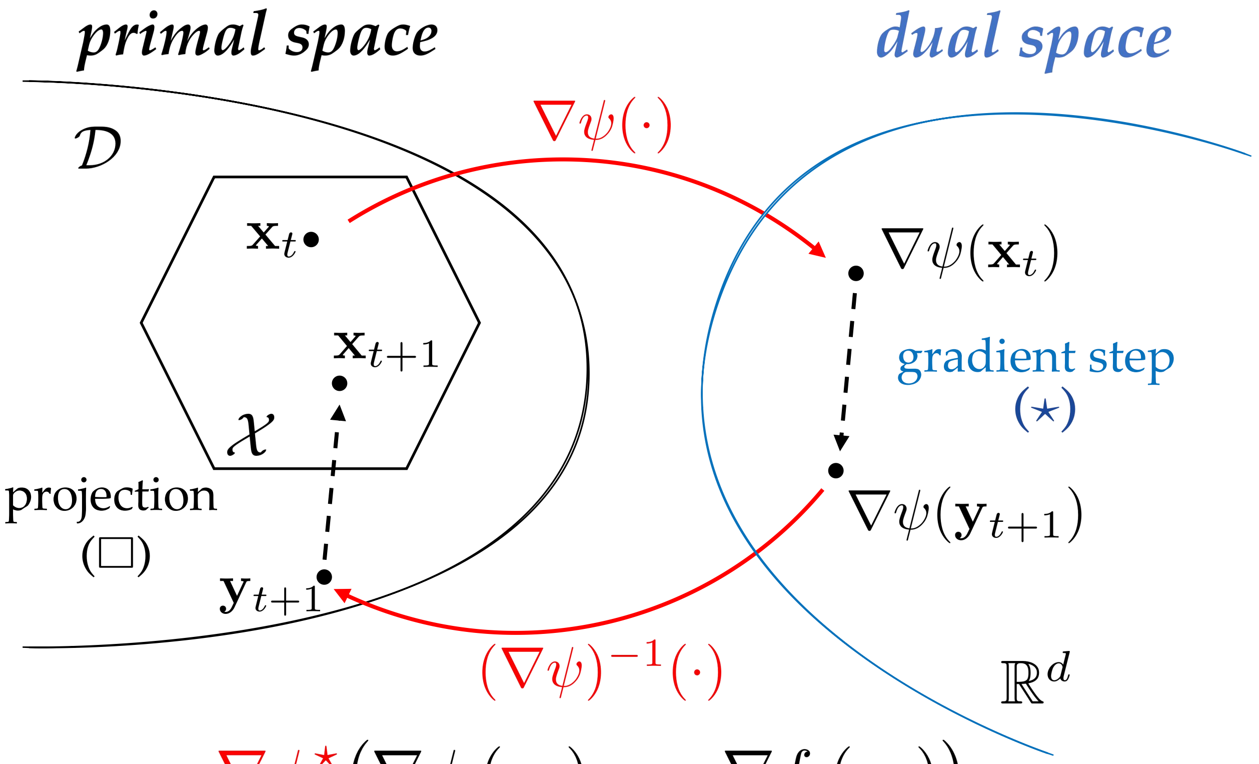
This derivative is a *linear operator from $\mathcal{B}$ to $\mathbb{R}$*: it takes a direction $\mathbf{h} \in \mathcal{B}$ as input and returns a real number as the directional rate of change of f at $\mathbf{x}$.

Consequently,

$$Df(\mathbf{x}) \in \mathcal{B}^*,$$

since the dual space $\mathcal{B}^*$ consists precisely of all continuous linear functionals mapping elements of $\mathcal{B}$ into $\mathbb{R}$.

Primal-Dual View for Mirror Descent



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

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

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

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

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

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

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 $\psi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it safisfies 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

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

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

$\Longleftrightarrow$

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

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

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

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

Mirror Map Calculation

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:

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

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. □

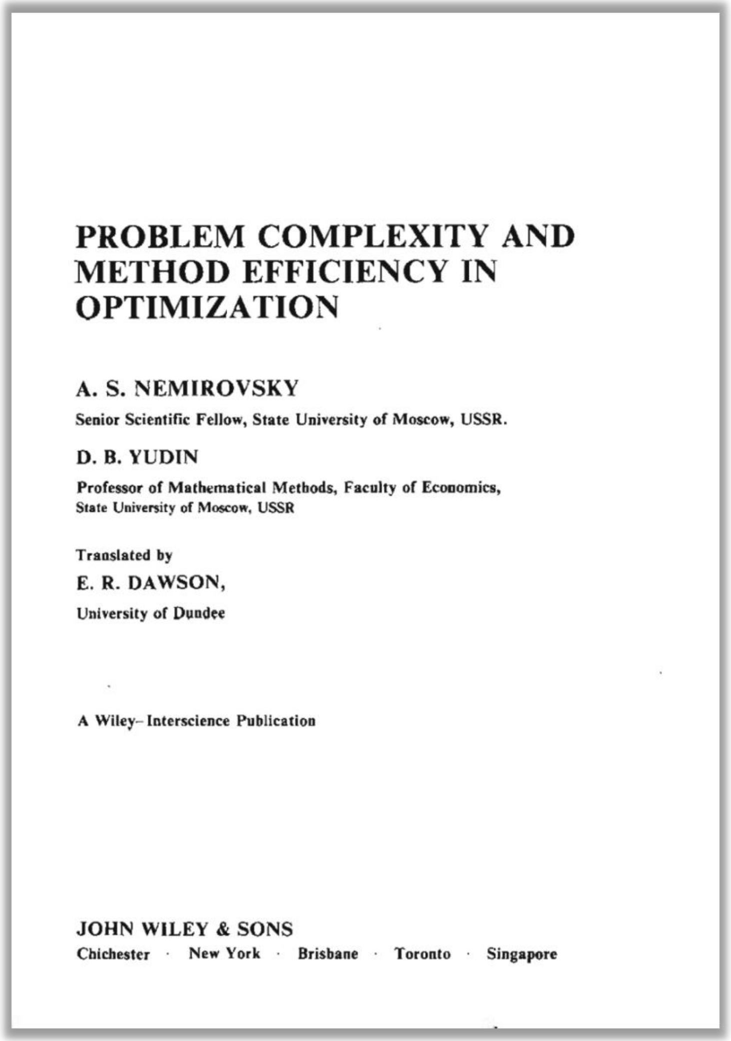
Mirror Descent for Optimization

- Our previous discussions (for offline optimization) essentially focus on the Euclidean norm case.

- They can be also extended to the general primal-dual norms by Mirror Descent framework, hence omitted…

- Mirror Descent can better capture the geometry of the spaces.

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

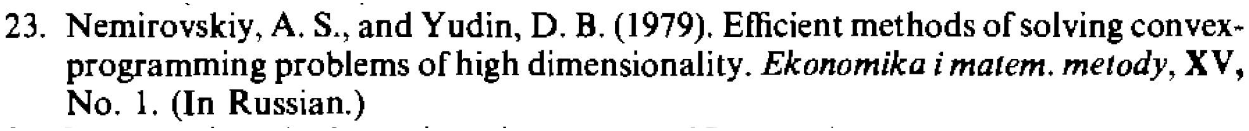
JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

A. S. Nemirovski (1947 -

D. B. Yudin (1919 - 2006)

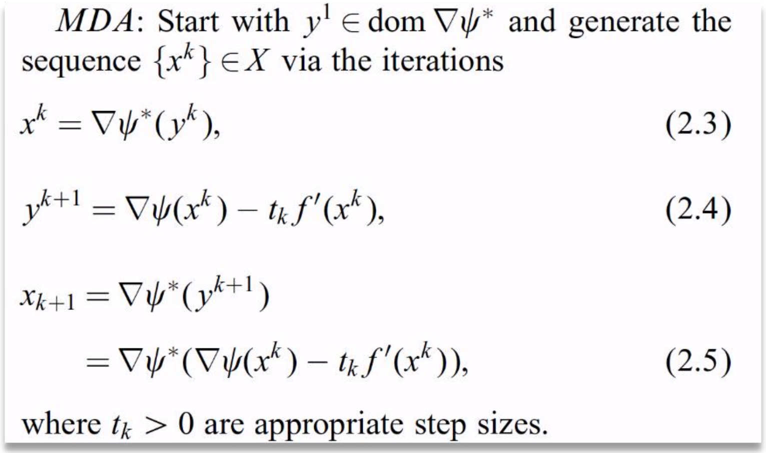
A.S. Nemirovski, D.B. Yudin, **Problem Complexity and Method Efficiency in Optimization**. Wiley-Interscience Series in Discrete Mathematics (A Wiley-Interscience Publication/Wiley, New York, 1983)



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.)

Mirror Descent: history bits

- Primal-Dual Interpretation and Proximal Interpretation



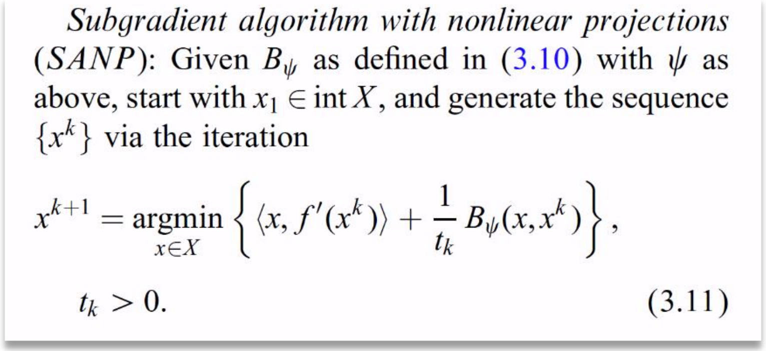
MDA: Start with $y^1 \in \operatorname{dom} \nabla\psi^*$ and generate the sequence $\{x^k\} \in X$ via the iterations

$$x^k = \nabla\psi^*(y^k), \tag{2.3}$$

$$y^{k+1} = \nabla\psi(x^k) - t_k f'(x^k), \tag{2.4}$$

$$\begin{aligned} x_{k+1} &= \nabla\psi^*(y^{k+1}) \\ &= \nabla\psi^*(\nabla\psi(x^k) - t_k f'(x^k)), \end{aligned} \tag{2.5}$$

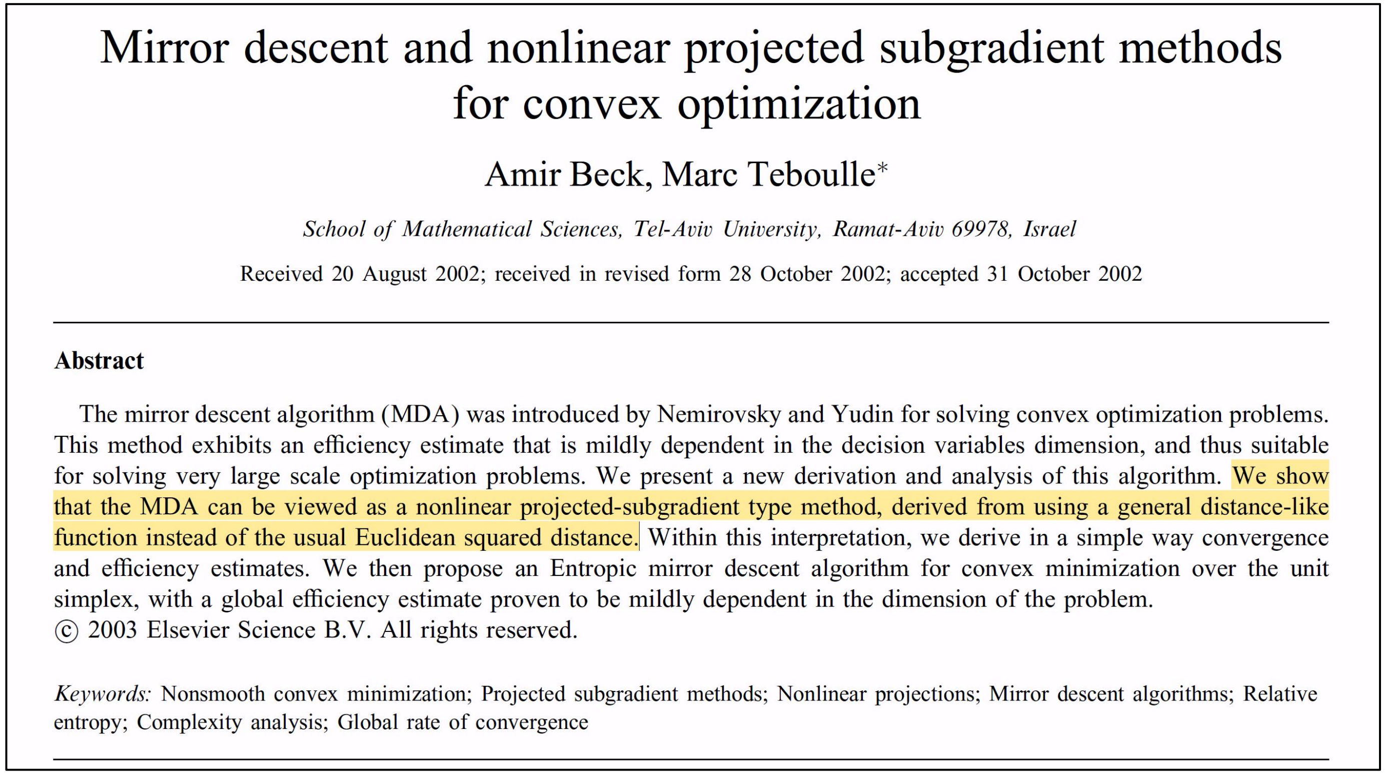
where $t_k > 0$ are appropriate step sizes.



Subgradient algorithm with nonlinear projections (*SANP*): Given B_ψ as defined in (3.10) with ψ as above, start with $x_1 \in \operatorname{int} X$, and generate the sequence $\{x^k\}$ via the iteration

$$x^{k+1} = \operatorname*{argmin}_{x \in X} \left\{ \langle x, f'(x^k) \rangle + \frac{1}{t_k} B_\psi(x, x^k) \right\},$$

$$t_k > 0. \tag{3.11}$$



Mirror descent and nonlinear projected subgradient methods for convex optimization

Amir Beck, Marc Teboulle*

School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel

Received 20 August 2002; received in revised form 28 October 2002; accepted 31 October 2002

Abstract

The mirror descent algorithm (MDA) was introduced by Nemirovsky and Yudin for solving convex optimization problems. This method exhibits an efficiency estimate that is mildly dependent in the decision variables dimension, and thus suitable for solving very large scale optimization problems. We present a new derivation and analysis of this algorithm. We show that the MDA can be viewed as a nonlinear projected-subgradient type method, derived from using a general distance-like function instead of the usual Euclidean squared distance. Within this interpretation, we derive in a simple way convergence and efficiency estimates. We then propose an Entropic mirror descent algorithm for convex minimization over the unit simplex, with a global efficiency estimate proven to be mildly dependent in the dimension of the problem.
© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Nonsmooth convex minimization; Projected subgradient methods; Nonlinear projections; Mirror descent algorithms; Relative entropy; Complexity analysis; Global rate of convergence

Amir Beck, Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters, 167-175, 2003.

Part 4. Follow-the-Regularized Leader

- Algorithmic Framework

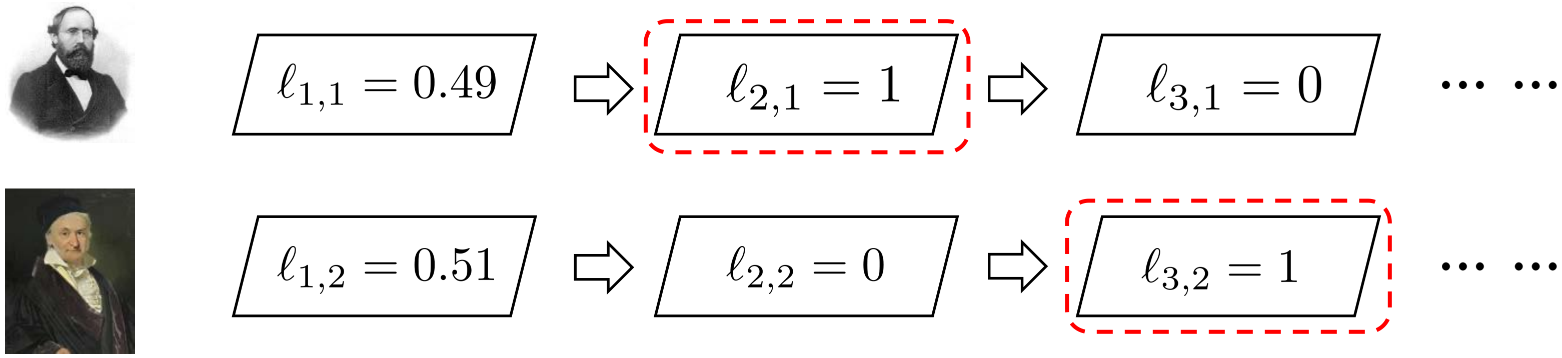
- Regret Analysis

- Primal-Dual Interpretation

Another OCO Framework: FTRL

- Recall: **Follow the Leader (FTL)**

> Select the expert that *performs best so far*, specifically,
>
> $$\boldsymbol{p}_t^{\mathrm{FTL}} = \operatorname*{arg\,min}_{\boldsymbol{p}\in\Delta_N} \langle \boldsymbol{p}, \boldsymbol{L}_{t-1}\rangle$$
>
> where $\boldsymbol{L}_{t-1} \triangleq \sum_{s=1}^{t-1} \boldsymbol{\ell}_s \in \mathbb{R}^N$ is the cumulative loss vector.



$\ell_{1,1} = 0.49$ ⇨ $\ell_{2,1} = 1$ ⇨ $\ell_{3,1} = 0$ … …

$\ell_{1,2} = 0.51$ ⇨ $\ell_{2,2} = 0$ ⇨ $\ell_{3,2} = 1$ … …

$$\mathrm{Reg}_T = \sum_{t=1}^T \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \min_{i\in[N]} \sum_{t=1}^T \ell_{t,i}$$

$$= T - \frac{T}{2} = \mathcal{O}(T)$$

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,

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

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

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

$\Longrightarrow$ 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*

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

(bias/range term)

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

(variance/stability)

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

($\mathbf{x}_{T+1} = \arg\min_{\mathbf{x}} F_{T+1}(\mathbf{x})$, thus ≤ 0)

General Analysis of FTRL

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

Lemma 4.

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

(bias/range term)

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

(variance/stability)

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

(negative term)

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

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

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

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

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

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

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

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

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

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*

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

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

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

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

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

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 $\psi_t(\mathbf{x}) \le \psi_{t+1}(\mathbf{x})$ for $t \in [T]$.Then, for FTRL satisfies*

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

Proof.

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

(range term)

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

(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 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.

$$\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{u}))$$

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

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

□

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: Equivalence Condition

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

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

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: Primal-Dual View

- Mirror Descent

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

- FTRL: Dual Averaging (or called *lazy mirror descent*)

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

averaging updates in dual space

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

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

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

FTRL vs OMD

- Consider PEA problem and Hedge algorithm as a case study.

lazy update $p_{t+1,i} \propto \exp\left(-\eta L_{t,i}\right), \forall i \in [N]$ where $L_{t,i} = \sum_{s=1}^{t} \ell_{s,i}$

greedy update $p_{t+1,i} \propto p_{t,i} \exp\left(-\eta \ell_{t,i}\right), \forall i \in [N]$ where $p_{0,i} = 1/N$

It can be verified that the two updates are ***exactly the same*** when ***learning rate is fixed***.

Essentially: lazy update ← FTRL; greedy update ← OMD

FTRL vs OMD

- Consider PEA problem and Hedge algorithm as a case study.

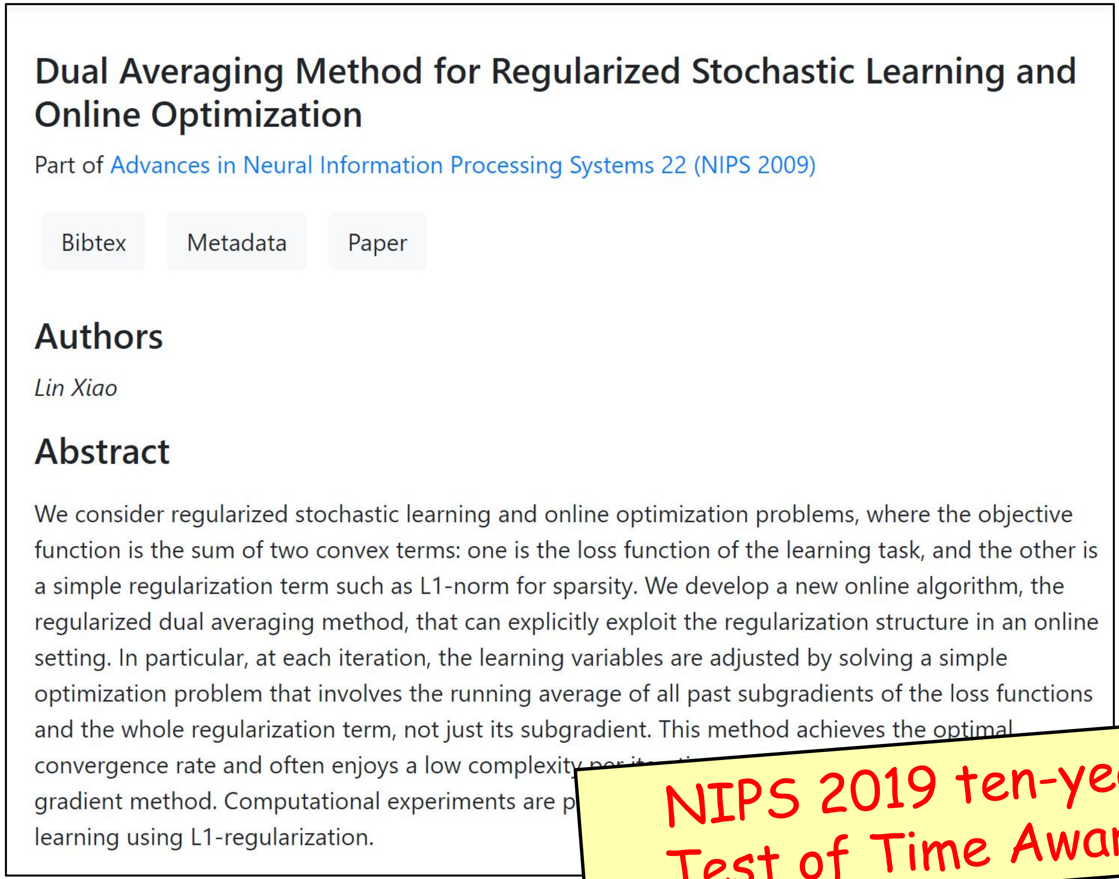
lazy update $p_{t+1,i} \propto \exp\left(-\eta_t L_{t,i}\right), \forall i \in [N]$ where $L_{t,i} = \sum_{s=1}^{t} \ell_{s,i}$

greedy update $p_{t+1,i} \propto p_{t,i} \exp\left(-\eta_t \ell_{t,i}\right), \forall i \in [N]$ where $p_{0,i} = 1/N$

However, the two updates are ***significantly different*** when ***learning rate is changing***.

Essentially: lazy update ← FTRL; greedy update ← OMD

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

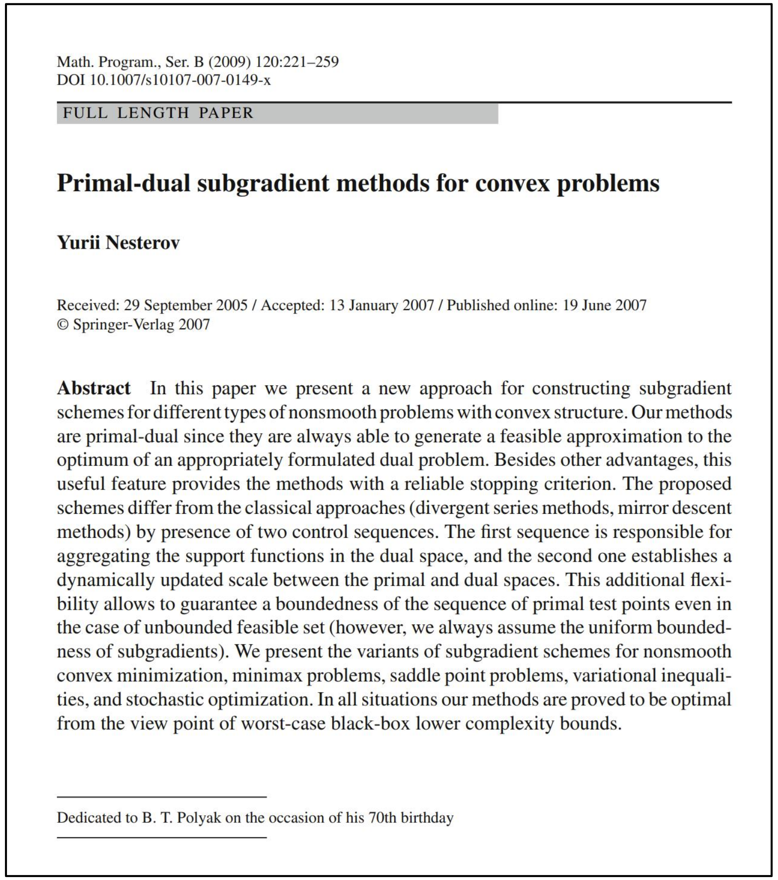
Abstract

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 it

gradient method. Computational experiments are p

learning using L1-regularization.

NIPS 2019 ten-year Test of Time Award!



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

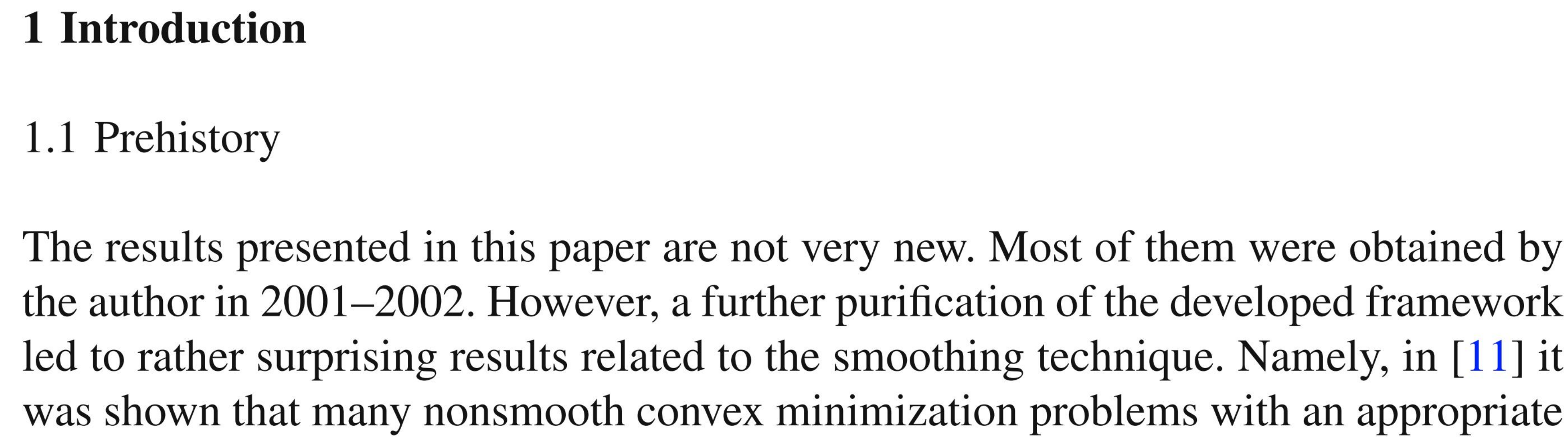
Received: 29 September 2005 / Accepted: 13 January 2007 / Published online: 19 June 2007
© Springer-Verlag 2007

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

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

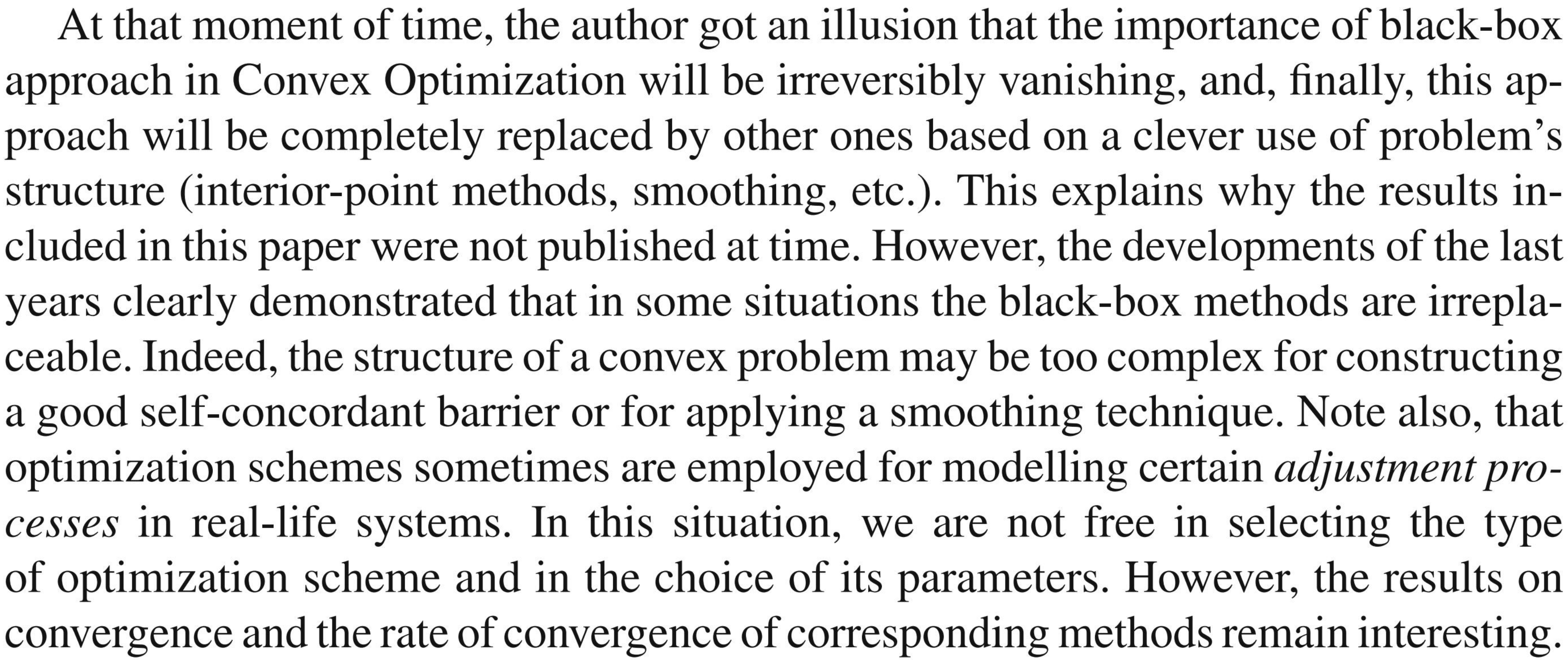
Yurii Nesterov. Primal-dual subgradient methods for convex problems, Math Programming B. 2005.



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。

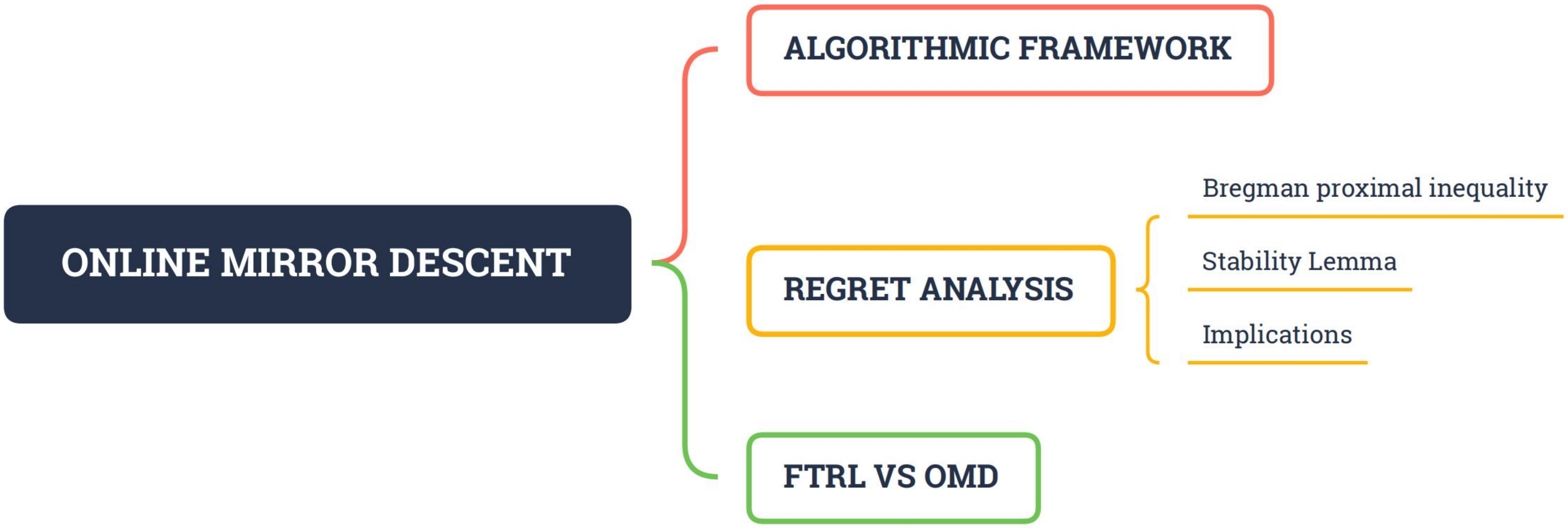
Congrats to WLA Prize (actually)

WLA PRIZE IN COMPUTER SCIENCE OR MA
2023年世界顶尖科学家协会奖“智能科学或数学奖”
WORLD LAUREATES FORUM
世界顶尖科学家论坛
6TH

2023年顶科协奖“智能科学或数学奖”

2023年11月6日

Summary



Q & A

Thanks!