# Online Convex Optimization and Its Surprising Applications

Francesco Orabona

KAUST

Learning Theory Summer School, Copenhagen, Denmark, 2026

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

# Aims of the Lecture

- Provide an introduction to Online Convex Optimization
- *Almost* rigorous: details are missing, but theorems are correct

- (1-slide) Proofs! Because it is the only way to design online learning algorithms
- Ideally, when in 1 week all this material will disappear from your memory, you can still use the slides as a "cheat sheet"

- Most of the material is based on my upcoming book on online learning (`https://arxiv.org/abs/1912.13213`), my blog posts (`https://parameterfree.com`), and some recent papers

# Outline of the Lecture

# Online Convex Optimization

Online Learning

1. In each round, output $\boldsymbol{x}_t \in \mathcal{V}$ — Choose $\boldsymbol{x}_t$ before observing $\ell_t$
2. Pay $\ell_t(\boldsymbol{x}_t)$ — No assumptions on how $\ell_t$ is generated!
3. Update $\boldsymbol{x}_{t+1}$ based on received information on $\ell_t$

Regret minimization

$$\min_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_T \in \mathcal{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) \qquad \text{equivalently} \qquad \min_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_T \in \mathcal{V}} \underbrace{\sum_{t=1}^{T} \ell_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u})}_{\mathit{Regret}_T(\boldsymbol{u})}$$

- The algorithm is *no-regret* if $\lim_{T\to\infty} \frac{1}{T} \mathit{Regret}_T(\boldsymbol{u}) \leq 0$ for all $\boldsymbol{u} \in \mathcal{V}$ and any sequence of losses in a certain family

# Why Online Convex Optimization?

- It is a strict generalization of the learning with expert setting
- It generalizes the setting of batch and stochastic convex optimization, in 99% of the cases without losing anything
- It provides a different mindset for designing optimization algorithms

- It is connected to a number of topics: Adaptation, Generalization, PAC-Bayes, Compression, Betting, etc.

# Some Famous Online Learning Algorithms

- Online Gradient Descent [Zinkevich, ICML'03]
- AdaGrad [Duchi et al., COLT'10, JMLR'11; McMahan&Streeter, COLT'10]
- AMSGrad [Reddi et al., ICLR'18]

These algorithms are designed to work in the adversarial setting and have a $\mathcal{O}(\sqrt{T})$ regret bound

We will see that they can also be used as stochastic optimization algorithms with a $\mathcal{O}(\frac{1}{\sqrt{T}})$ convergence rate

# Assumptions and Definitions

- Losses: $\ell_t : \mathbb{R}^d \to \mathbb{R}$, convex, 1-Lipschitz
- Feasible set: $\mathcal{V} \subseteq \mathbb{R}^d$, closed, convex, non-empty
- Iterates: All technical conditions for iterates $\boldsymbol{x}_t$ to exists hold

# Mainly Two Main Meta-Algorithms

- Online Mirror Descent (OMD)
- Follow-the-Regularized-Leader (FTRL)

- These two meta-algorithms cover 90% of the (online) optimization algorithms
- Examples
  - Online Gradient Descent = special case of OMD
  - Dual Averaging = Special case of FTRL with linearized losses
  - Regularized Dual Averaging = Special case of FTRL with linearized losses
  - “Lazy version” of online gradient descent = FTRL
  - Newton algorithm = OMD with distance induced by the Hessian
  - Accelerated algorithm = two OCO algorithms playing against each other
  - Frank-Wolfe algorithm = two OCO algorithms playing against each other
  - etc.

# Online Subgradient Descent

# Projected Online Gradient Descent

**Require:** Feasible set $\mathcal{V} \subseteq \mathbb{R}^d$, $\boldsymbol{x}_1 \in \mathcal{V}$, $\eta_1, \cdots, \eta_T > 0$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \mathcal{V}$
3: Pay $\ell_t(\boldsymbol{x}_t)$
4: Set $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{x}_t)$
5: $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{V}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) = \mathit{argmin}_{\boldsymbol{y} \in \mathcal{V}} \ \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{y}\|_2$
6: **end for**

[Zinkevich, ICML'03]

# Guarantee for OGD (1)

## Lemma

*Let $\ell_t : \mathcal{V} \to \mathbb{R}$ differentiable in an open set that contains $\mathcal{V}$. Then, $\forall \boldsymbol{u} \in \mathcal{V}$, OGD satisfies*

$$\eta_t(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \eta_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle \leq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 + \frac{\eta_t^2}{2}\|\boldsymbol{g}_t\|_2^2 \,.$$

## Proof.

$$\begin{aligned}
\|\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 - \|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 &\overset{\Pi \text{ is non expansive}}{\leq} \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{u}\|_2^2 - \|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 \\
&= -2\eta_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle + \eta_t^2 \|\boldsymbol{g}_t\|_2^2 \\
&\overset{\text{Convexity}}{\leq} -2\eta_t(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) + \eta_t^2 \|\boldsymbol{g}_t\|_2^2 \,.
\end{aligned}$$

□

# Guarantee for OGD (2)

## Theorem

*Let $\ell_1, \cdots, \ell_T$ differentiable in open sets containing $\mathcal{V}$. Pick any $\boldsymbol{x}_1 \in \mathcal{V}$ and assume $\eta_t = \eta,\ t = 1, \ldots, T$. Then, $\forall \boldsymbol{u} \in \mathcal{V}$, OGD satisfies*

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \frac{\|\boldsymbol{u} - \boldsymbol{x}_1\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2 - \frac{1}{2\eta} \|\boldsymbol{x}_{T+1} - \boldsymbol{u}\|_2^2 .$$

## Proof.

Dividing the inequality in the previous Lemma by $\eta$ and summing over $t = 1, \cdots, T$, we have

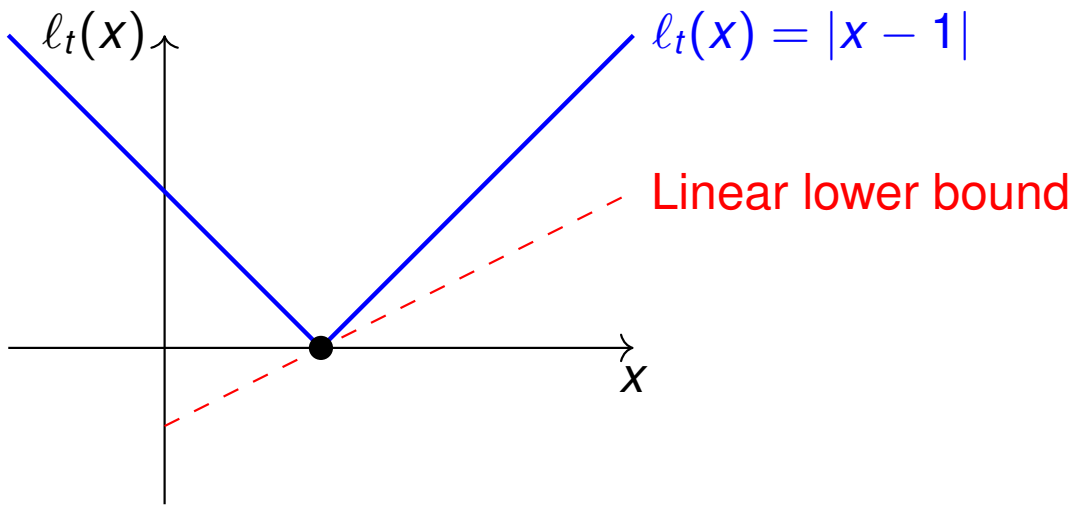
$$\begin{aligned}
\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) &\leq \sum_{t=1}^{T} \left( \frac{1}{2\eta} \|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 - \frac{1}{2\eta} \|\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2 \\
&= \frac{1}{2\eta} \|\boldsymbol{x}_1 - \boldsymbol{u}\|_2^2 - \frac{1}{2\eta} \|\boldsymbol{x}_{T+1} - \boldsymbol{u}\|_2^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2 .
\end{aligned}$$

□

# Non-Differentiable Convex Functions

- If the losses are convex, but not differentiable, we cannot calculate the gradients
- We only need gradients because they satisfy $\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u}) \leq \langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \rangle$

- Solution: use any vector $\boldsymbol{g}_t$ that satisfies $\ell_t(\boldsymbol{x}_t) - \ell(\boldsymbol{u}) \leq \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle$ for all $\boldsymbol{u} \in \mathcal{V}$
- $\boldsymbol{g}_t$ is called a **subgradient** of $\ell_t$ in $\boldsymbol{x}_t$
- The set of all subgradients $\ell$ in $\boldsymbol{x}$ is called **subdifferential** and it is denoted by $\partial \ell_t(\boldsymbol{x}_t)$



# Projected Online Subgradient Descent

**Require:** Feasible set $\mathcal{V} \subseteq \mathbb{R}^d$, $\boldsymbol{x}_1 \in \mathcal{V}$, $\eta_1, \ldots, \eta_T > 0$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \mathcal{V}$
3: Pay $\ell_t(\boldsymbol{x}_t)$
4: Set $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$
5: $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{V}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) = \mathit{argmin}_{\boldsymbol{y} \in \mathcal{V}} \ \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{y}\|_2$
6: **end for**

Same guarantee of OGD:

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \frac{\|\boldsymbol{u} - \boldsymbol{x}_1\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2 - \frac{1}{2\eta} \|\boldsymbol{x}_{T+1} - \boldsymbol{u}\|_2^2 \ .$$

[Zhang, ICML'04]

# Learning rate in OSD

- The regret is $\sum_{t=1}^{T}(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \frac{\|\boldsymbol{u}-\boldsymbol{x}_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_2^2$
- Assume the function 1-Lipschitz w.r.t. the $L_2$ norm ($|\ell_t(\boldsymbol{x}) - \ell_t(\boldsymbol{u})| \leq \|\boldsymbol{x} - \boldsymbol{u}\|_2$)
- Then, $\sum_{t=1}^{T}(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \frac{\|\boldsymbol{u}-\boldsymbol{x}_1\|_2^2}{2\eta} + \frac{T\eta}{2}$
- Optimal learning rate: $\eta = \frac{\|\boldsymbol{u}-\boldsymbol{x}_1\|_2}{\sqrt{T}}$
- Any problem with this choice?

- Practical choice $\eta = \frac{\alpha}{\sqrt{T}}$ that gives $\mathit{Regret}_T(\boldsymbol{u}) \leq \frac{1}{2}\left(\frac{\|\boldsymbol{x}_1-\boldsymbol{u}\|_2^2}{\alpha} + \alpha\right)\sqrt{T}$

- Easy case: $\mathcal{V}$ has bounded diameter $D$, then $\eta = \frac{D}{\sqrt{T}}$ gives regret $D\sqrt{T}$

# Applications: From Online to Stochastic (or Batch) Optimization (1)

1: **for** $t = 1$ **to** $T$ **do**
2: Get $\boldsymbol{x}_t$ from an Online Convex Optimization algorithm
3: Receive stochastic subgradient $\boldsymbol{g}_t$ such that $\mathbb{E}_t[\boldsymbol{g}_t] \in \partial F(\boldsymbol{x}_t)$
4: Pass loss $\ell_t(\boldsymbol{x}) = \langle \boldsymbol{g}_t, \boldsymbol{x} \rangle$ to Online Learning Algorithm
5: **end for**
6: **return** $\bar{\boldsymbol{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t$

**Theorem**

$$\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F(\boldsymbol{u}) \leq \frac{\mathbb{E}[Regret_T(\boldsymbol{u})]}{T}, \quad \forall \boldsymbol{u} \in \mathcal{V}$$

Corollary: any result on regret translates to a result on convergence for stochastic optimization of convex functions

[Cesa-Bianchi et al., IEEE Trans. Inf. Theory 2004]

# Applications: From Online to Stochastic (or Batch) Optimization (1)

**Proof.**

$$\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F(\boldsymbol{u}) \overset{\text{Jensen}}{\leq} \frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}[F(\boldsymbol{x}_t)] - F(\boldsymbol{u}))$$

$$\overset{\text{convexity}}{\leq} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \mathbb{E}_t[\boldsymbol{g}_t], \boldsymbol{x}_t - \boldsymbol{u}\rangle]$$

$$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}_t[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle]]$$

$$\overset{\text{total expectation}}{=} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle]$$

$$= \frac{\mathbb{E}[\mathit{Regret}_T(\boldsymbol{u})]}{T}$$

□

# Example: Stochastic Subgradient Descent

**Require:** Feasible set $\mathcal{V} \subseteq \mathbb{R}^d$, $\boldsymbol{x}_1 \in \mathcal{V}$, $\eta = \frac{\alpha}{\sqrt{T}}$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \mathcal{V}$
3: Receive stochastic subgradient $\boldsymbol{g}_t$ such that $\mathbb{E}_t[\boldsymbol{g}_t] \in \partial F(\boldsymbol{x}_t)$
4: $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{V}}(\boldsymbol{x}_t - \eta \boldsymbol{g}_t)$
5: **end for**
6: **return** $\bar{\boldsymbol{x}}_T = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t$

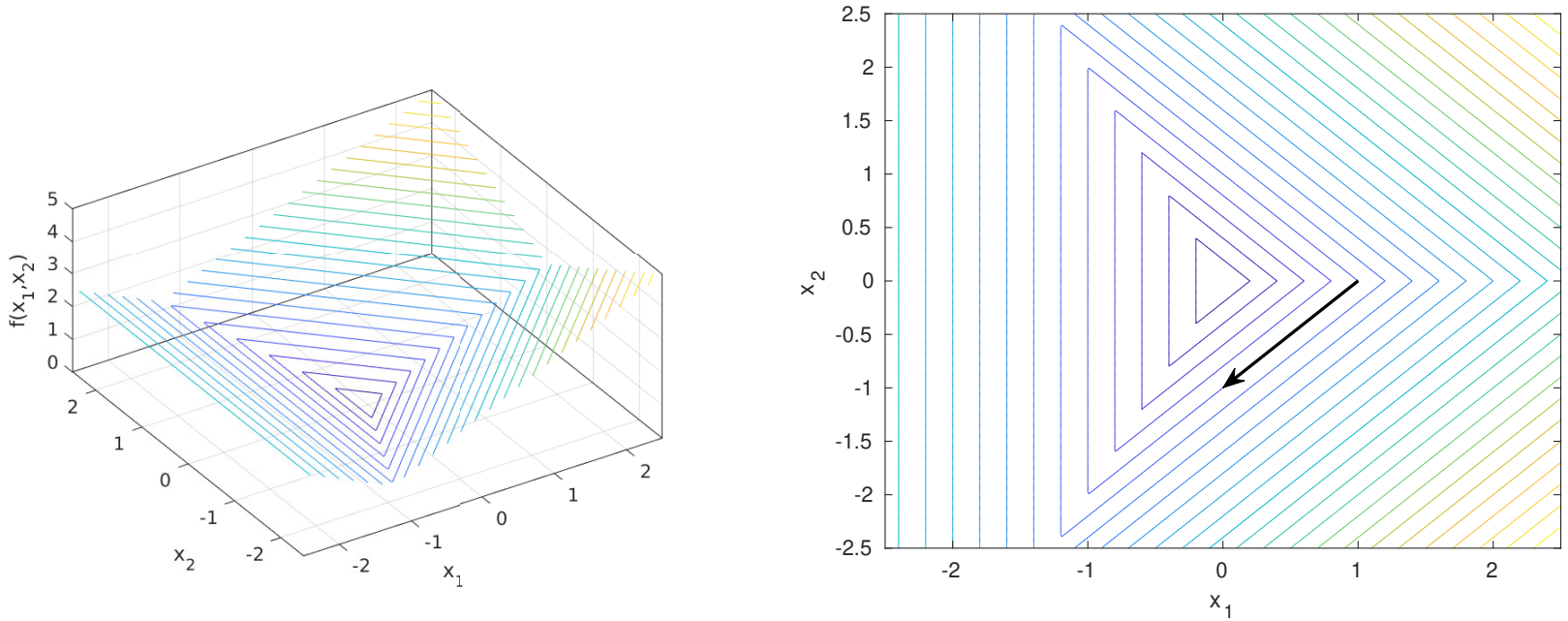
From the previous slides, we have

$$\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F(\boldsymbol{x}^\star) \leq \frac{1}{2\sqrt{T}} \left( \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|_2^2}{\alpha} + \alpha \right)$$

In words, one pass of SGD minimizes the true risk

# Beyond Online Subgradient Descent

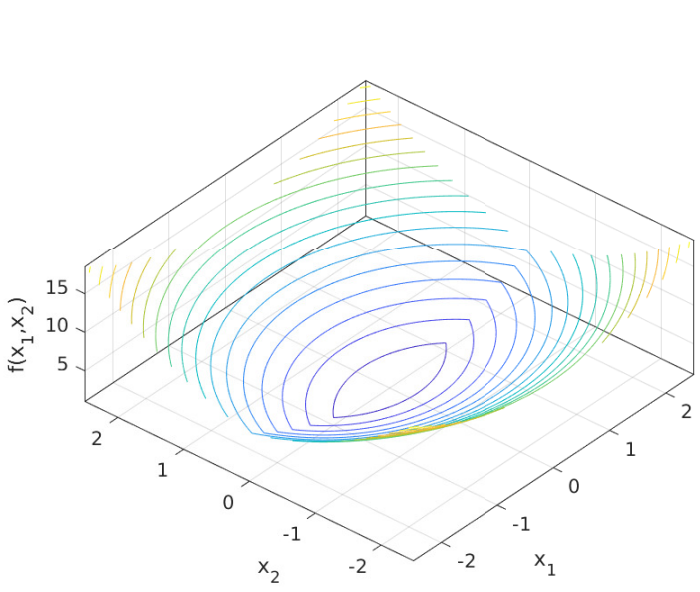
# Does Online Subgradient Descent Minimize the Functions? (1)

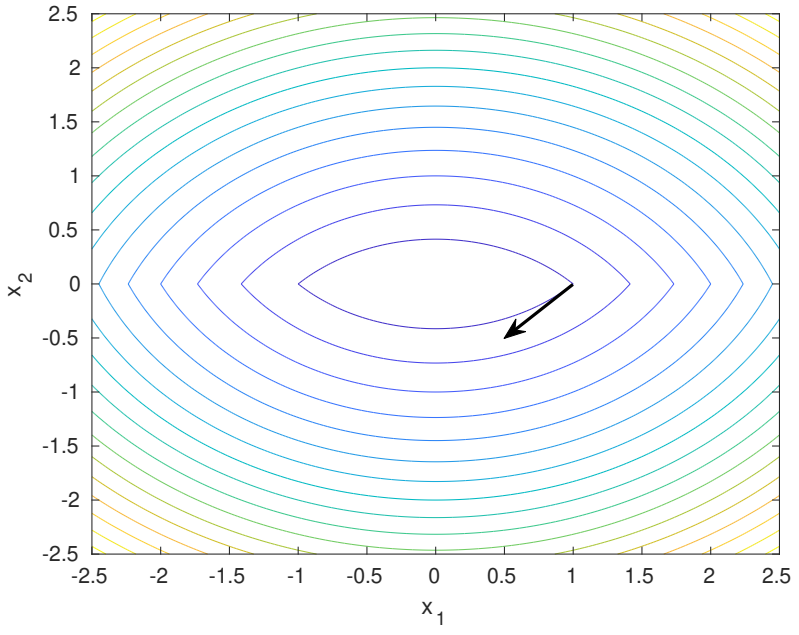



3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[-x_1, x_1 - x_2, x_1 + x_2]$. A negative subgradient is indicated by the black arrow

[Orabona, arXiv'19]

# Does Online Subgradient Descent Minimize the Functions? (2)




3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2]$.
A negative subgradient is indicated by the black arrow

[Orabona, arXiv'19]

# Understanding the Update of Online Subgradient Descent

$$
\begin{aligned}
\Pi_{\mathcal{V}}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) &= \underset{\boldsymbol{x} \in \mathcal{V}}{\mathit{argmin}} \ \|\boldsymbol{x} - \boldsymbol{x}_t + \eta_t \boldsymbol{g}_t\|_2^2 \\
&= \underset{\boldsymbol{x} \in \mathcal{V}}{\mathit{argmin}} \ \|\eta_t \boldsymbol{g}_t\|_2^2 + 2\eta_t \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle + \|\boldsymbol{x}_t - \boldsymbol{x}\|_2^2 \\
&= \underset{\boldsymbol{x} \in \mathcal{V}}{\mathit{argmin}} \ \underbrace{\ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle}_{\text{Linear approximation of } \ell_t} + \frac{1}{2\eta_t} \underbrace{\|\boldsymbol{x}_t - \boldsymbol{x}\|_2^2}_{\text{Stay close to } \boldsymbol{x}_t}
\end{aligned}
$$

where $\Pi_{\mathcal{V}}$ is the Euclidean projection onto $\mathcal{V}$, i.e., $\Pi_{\mathcal{V}}(\boldsymbol{x}) = \mathit{argmin}_{\boldsymbol{y} \in \mathcal{V}} \ \|\boldsymbol{x} - \boldsymbol{y}\|_2$

# General Notion of Distances using Bregman Divergences

$$\underset{\boldsymbol{x}\in\mathcal{V}}{\textit{argmin}}\ \ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle + \frac{1}{2\eta_t}\|\boldsymbol{x}_t - \boldsymbol{x}\|_2^2$$

Why the square Euclidean norm?

I can use general notion of distances, in particular **Bregman divergences**

**Definition (Bregman Divergence [Bregman, 1967])**

Let $\psi : X \to \mathbb{R}$ be strictly convex and differentiable on $\operatorname{int} X \neq \{\}$. The **Bregman Divergence** w.r.t. $\psi$ is denoted by $B_\psi : X \times \operatorname{int} X \to \mathbb{R}$ defined as

$$B_\psi(\boldsymbol{x}; \boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \nabla\psi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \ .$$

# Online Mirror Descent

We start from the equivalent formulation of the OSD update

$$\boldsymbol{x}_{t+1} = \underset{\boldsymbol{x}\in\mathcal{V}}{\mathit{argmin}}\ \ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t\rangle + \frac{1}{2\eta_t}\|\boldsymbol{x}_t - \boldsymbol{x}\|_2^2$$

and we can change the last term with a Bregman Divergence

$$\boldsymbol{x}_{t+1} = \underset{\boldsymbol{x}\in\mathcal{V}}{\mathit{argmin}}\ \ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t\rangle + \frac{1}{\eta_t}B_\psi(\boldsymbol{x}; \boldsymbol{x}_t)$$

**Require:** $\psi : X \to \mathbb{R}$ strictly convex and differentiable on $\operatorname{int} X$, feasible set $\mathcal{V} \subseteq X \subseteq \mathbb{R}^d$, $\boldsymbol{x}_1 \in \operatorname{int} X \cap \mathcal{V}$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \mathcal{V}$
3: Pay $\ell_t(\boldsymbol{x}_t)$
4: Set $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$
5: Set $\boldsymbol{x}_{t+1} \in \mathit{argmin}_{\boldsymbol{x}\in\mathcal{V}}\ \langle \boldsymbol{g}_t, \boldsymbol{x}\rangle + \frac{1}{\eta_t}B_\psi(\boldsymbol{x}; \boldsymbol{x}_t)$
6: **end for**

[Nemirovskij&Yudin, 1983][Warmuth&Jagota, 1997][Beck&Teboulle, 2003]

# Strongly Convex Functions

## Definition

$f : \mathbb{R}^d \to (-\infty, +\infty]$ is $\lambda$-strongly convex w.r.t. $\|\cdot\|$ if

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{y} \rangle - \frac{\lambda}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \ \forall \boldsymbol{g} \in \partial f(\boldsymbol{x}) \ .$$

Intuition: Lower bounded by a quadratic

## Lemma (For OMD proof)

*If $\psi$ is $\lambda$-strongly convex w.r.t. $\|\cdot\|$ then $B_\psi(\boldsymbol{x}; \boldsymbol{y}) \geq \frac{\lambda}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$*

## Lemma (For FTRL proof)

*Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ closed, proper, subdifferentiable, and $\lambda$-strongly convex with respect to a norm $\|\cdot\|$ over its domain. Let $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$. Then, for all $\boldsymbol{x} \in \operatorname{dom} \partial f$, and $\boldsymbol{g} \in \partial f(\boldsymbol{x})$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \leq \frac{1}{2\lambda}\|\boldsymbol{g}\|_\star^2 \ .$$

# Regret Guarantee of OMD

## Theorem

*Let $\psi$ be $\lambda$-strongly convex w.r.t. $\|\cdot\|$. Pick any $\boldsymbol{x}_1 \in \operatorname{int} X \cap \mathcal{V}$ and assume $\eta_t = \eta,\ t = 1, \dots, T$. Then, $\forall \boldsymbol{u} \in \mathcal{V}$, OMD satisfies*

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \frac{B_\psi(\boldsymbol{u}; \boldsymbol{x}_1)}{\eta} + \frac{\eta}{2\lambda} \sum_{t=1}^{T} \|\boldsymbol{g}_t\|_\star^2 - \frac{1}{\eta} B_\psi(\boldsymbol{u}; \boldsymbol{x}_{T+1}) \,.$$

## Proof.

One can show

$$\begin{aligned} \eta_t(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) &\leq \eta \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle \\ &\leq B_\psi(\boldsymbol{u}; \boldsymbol{x}_t) - B_\psi(\boldsymbol{u}; \boldsymbol{x}_{t+1}) - B_\psi(\boldsymbol{x}_{t+1}; \boldsymbol{x}_t) + \langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \end{aligned}$$

The last term can be bounded as

$$\langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \leq \eta_t \|\boldsymbol{g}_t\|_\star \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\| \leq \frac{\eta_t^2 \|\boldsymbol{g}_t\|_\star^2}{2\lambda} + \frac{\lambda}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2$$

From strong convexity of $\psi$, we get $-B_\psi(\boldsymbol{x}_{t+1}; \boldsymbol{x}_t) \leq -\frac{\lambda}{2} \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2$. Putting all together and summing over time, we get the stated bound. □

# Example: Online Subgradient Descent

- Set $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$
- $\psi$ is 1-strongly convex w.r.t. the $L_2$ norm
- Dual norm of $L_2$ is $L_2$

$$B(\boldsymbol{x}; \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2 - \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \langle \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Regret for any $\boldsymbol{u}$:

$$\sum_{t=1}^{T}(\ell_t(\boldsymbol{x}_t) - \ell(\boldsymbol{u})) \leq \frac{B_\psi(\boldsymbol{u}; \boldsymbol{x}_1)}{\eta} + \frac{\eta}{2\lambda}\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_*^2$$

$$= \frac{\|\boldsymbol{x}_1 - \boldsymbol{u}\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_2^2$$

# Example: Exponentiated Gradient (a.k.a. Hedge, EWA, etc.)

- Set $\mathcal{V} = \Delta^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : x_i \geq 0, \|\boldsymbol{x}\|_1 = 1\}$
- Set $\psi(\boldsymbol{x}) = \sum_{i=1}^d x_i \ln x_i$
- $\psi$ is 1-strongly convex w.r.t. the $L_1$ norm
- Dual norm of $L_1$ is $L_\infty$
- Assume $\|\boldsymbol{g}_t\|_\infty \leq 1$

**Require:** $\eta > 0$

1: Set $\boldsymbol{x}_1 = [1/d, \ldots, 1/d]$
2: **for** $t = 1$ **to** $T$ **do**
3: Output $\boldsymbol{x}_t \in \Delta^{d-1}$
4: Pay $\ell_t(\boldsymbol{x}_t)$
5: Set $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$
6: $x_{t+1,j} = \frac{x_{t,j} \exp(-\eta g_{t,j})}{\sum_{i=1}^d x_{t,i} \exp(-\eta g_{t,i})}$, $j = 1, \ldots, d$
7: **end for**

Regret for any $\boldsymbol{u}$:
$$\sum_{t=1}^T (\ell_t(\boldsymbol{x}_t) - \ell(\boldsymbol{u})) \leq \frac{B_\psi(\boldsymbol{u}; \boldsymbol{x}_1)}{\eta} + \frac{\eta}{2\lambda} \sum_{t=1}^T \|\boldsymbol{g}_t\|_*^2 \leq \frac{\ln d}{\eta} + \frac{\eta T}{2}$$

Set $\eta = \sqrt{\frac{2 \ln d}{T}}$ to obtain the upper bound of $\sqrt{2T \ln d}$

[Kivinen&Warmuth, 1997]

# Follow-The-Regularized-Leader Algorithm

# Follow-the-Regularized-Leader

**Require:** Feasible set $\mathcal{V} \subseteq X \subseteq \mathbb{R}^d$, a sequence of regularizers $\psi_1, \ldots, \psi_T : X \to \mathbb{R}$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \mathit{argmin}_{\boldsymbol{x} \in \mathcal{V}} \ \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$
3: Receive $\ell_t : \mathcal{V} \to \mathbb{R}$ and pay $\ell_t(\boldsymbol{x}_t)$
4: **end for**

[Gordon, COLT'99][Shalev-Shwartz&Singer, COLT'06, NeurIPS'06][Shalev-Shwartz, PhD'07][Abernethy et al., COLT'08][Hazan&Kale, COLT'08]

# Guarantee for FTRL

## Lemma

*Let $\psi_1, \dots, \psi_T : X \to \mathbb{R}$ be a sequence of regularization functions and $\mathcal{V} \subseteq X \subseteq \mathbb{R}^d$. Denote by $F_t(\boldsymbol{x}) = \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$. Set $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x} \in \mathcal{V}} F_t(\boldsymbol{x})$. Then, for any $\boldsymbol{u} \in \mathbb{R}^d$, we have*

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

## Proof.

Just sum simplify the sums and use the fact that $F_1(\boldsymbol{x}_1) = \min_{\boldsymbol{x} \in \mathcal{V}} \psi_1(\boldsymbol{x})$. □

[McMahan, JMLR'17][Orabona, arXiv'19]

# An Explicit Regret with Strongly Convex Functions

Assume that $\psi_t + \sum_{i=1}^{t} \ell_i$ is $\lambda_t$ strongly convex w.r.t. $\|\cdot\|$ and $\psi_{t+1}(\boldsymbol{x}) \geq \psi_t(\boldsymbol{x})$, we have

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

$$= \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in \mathcal{V}} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T} [F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t)] + F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u})$$

$$\overset{\text{Strong convexity}}{\leq} \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in \mathcal{V}} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T} \frac{\|\boldsymbol{g}_t\|_\star^2}{2\lambda_t}$$

# FTRL with Linearized Losses

- FTRL needs to solve a convex optimization problem at each step

- I can run FTRL with any sequence of losses
- I can also construct some losses
- For example, I might want to run FTRL on $\hat{\ell}_t(\boldsymbol{x}) = \ell_t(\boldsymbol{x}_t) + \langle \boldsymbol{g}_t, \boldsymbol{x} - \boldsymbol{x}_t \rangle$ where $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$

**Require:** A sequence of regularizers $\psi_1, \dots, \psi_T : X \to \mathbb{R}$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \mathit{argmin}_{\boldsymbol{x} \in \mathcal{V}} \ \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \langle \boldsymbol{g}_i, \boldsymbol{x} \rangle$
3: Pay $\ell_t(\boldsymbol{x}_t)$
4: Get $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$
5: **end for**

Same regret because

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u} \rangle$$

# FTRL with Linearized Losses vs OSD

- $\mathcal{V} = \mathbb{R}^d$
- $\psi_{t+1}(\boldsymbol{x}) = \frac{1}{2\eta_{t+1}} \|\boldsymbol{x}\|_2^2$
- $\boldsymbol{x}_{t+1} = \mathit{argmin}_{\boldsymbol{x}} \ \frac{1}{2\eta_{t+1}} \|\boldsymbol{x}\|_2^2 + \sum_{i=1}^t \langle \boldsymbol{g}_i, \boldsymbol{x} \rangle = -\eta_{t+1} \sum_{i=1}^t \boldsymbol{g}_i$
- Compare it with OSD with $\boldsymbol{x}_1 = \boldsymbol{0}$: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}_t = -\sum_{i=1}^t \eta_i \boldsymbol{g}_i$

- **Important:** In FTRL the gradients are used with the same weight
- **Important:** In FTRL we don't take "jumps" of size $\eta_t$

# Example: FTRL with Linearized Loss and Euclidean Regularization

- $\mathcal{V} = \mathbb{R}^d$
- $\psi(\boldsymbol{x}) = \frac{\gamma}{2}\|\boldsymbol{x}\|_2^2$
- $\psi$ is $\gamma$-strongly convex w.r.t. $L_2$ norm
- Dual norm of $L_2$ norm is $L_2$ norm

$$\boldsymbol{x}_t = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{V}} \frac{\gamma}{2}\|\boldsymbol{x}\|_2^2 + \sum_{i=1}^{t-1}\langle \boldsymbol{g}_i, \boldsymbol{x}\rangle = \frac{-\sum_{i=1}^{t-1}\boldsymbol{g}_i}{\gamma}$$

$$\sum_{t=1}^{T}(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \le \sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle \le \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in\mathcal{V}} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T}\frac{\|\boldsymbol{g}_t\|_\star^2}{2\lambda_t}$$

$$= \frac{\gamma}{2}\|\boldsymbol{u}\|_2^2 + \sum_{t=1}^{T}\frac{\|\boldsymbol{g}_t\|_2^2}{2\gamma}$$

What is the optimal tuning of $\gamma$?