Online Convex Optimization and Its Surprising Applications

Francesco Orabona

KAUST

Machine Learning Summer School, OIST, 2024

Aims of the Lecture

- Provide an introduction to Online Convex Optimization
- *Almost* rigorous: details are missing, but theorems are correct
- Connections: not written anywhere, but known to people in the field

- (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 online learning notes (`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 $\boldsymbol{x}_t \in V$ — Choose $\boldsymbol{x}_t$ before observing ℓ_t
2. Pay $\ell_t(\boldsymbol{x}_t)$ — No assumptions on how ℓ_t is generated!
3. Update $\boldsymbol{x}_{t+1}$ based on received information on ℓ_t

Regret minimization

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

- The algorithm is *no-regret* if $\frac{1}{T} Regret_T(\boldsymbol{u}) \to 0$ for all $\boldsymbol{u} \in V$ 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: Generalization, PAC-Bayes, Compression, Betting, etc.

Some Famous Online Learning Algorithms

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

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

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

Assumptions and Definitions

- Losses: $\ell_t : \mathbb{R}^d \to \mathbb{R}$, convex, 1-Lipschitz
- Feasible set: $V \subseteq \mathbb{R}^d$, 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 $V \subseteq \mathbb{R}^d$, $\boldsymbol{x}_1 \in V$, $\eta_1, \cdots, \eta_T > 0$

for $t = 1$ **to** T **do**

  Output $\boldsymbol{x}_t \in V$

  Pay $\ell_t(\boldsymbol{x}_t)$

  Set $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{x}_t)$

  $\boldsymbol{x}_{t+1} = \Pi_V(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) = \mathit{argmin}_{\boldsymbol{y} \in V} \; \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{y}\|_2$

end for

[Zinkevich, ICML'03]

Guarantee for OGD (1)

Lemma

Let $\ell_t : V \to \mathbb{R}$ differentiable in an open set that contains V. Then, $\forall \boldsymbol{u} \in V$, 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 V. Pick any $\boldsymbol{x}_1 \in V$ and assume $\eta_t = \eta,\ t = 1, \ldots, T$. Then, $\forall \boldsymbol{u} \in V$, 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 η and summing over $t = 1, \cdots, T$, we have

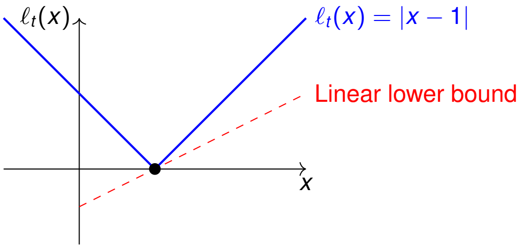
$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \sum_{t=1}^{T} \left(\frac{1}{2\eta} \|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 - \frac{1}{2\eta} \|\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{g}_t\|_2^2$$

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

□

Non-Differentiable Convex Functions

- If the losses are convex, but not differentiable, we cannot calculate the gradients
- We only need gradients because they satisfy
  $\ell_t(\boldsymbol{x}_t) - \ell(\boldsymbol{u}) \leq \langle \nabla \ell_t(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{u} \rangle$
- Solution: use <u>any</u> 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 $\boldsymbol{u} \in V$
- $\boldsymbol{g}_t$ is called a **subgradient** of ℓ_t in $\boldsymbol{x}_t$
- The set of all subgradients ℓ in $\boldsymbol{x}$ is called **subdifferential** and it is denoted by $\partial \ell_t(\boldsymbol{x}_t)$



Projected Online Subgradient Descent

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

1: **for** $t = 1$ **to** T **do**
2: Output $\boldsymbol{x}_t \in V$
3: Pay $\ell_t(\boldsymbol{x}_t)$
4: Set $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$
5: $\boldsymbol{x}_{t+1} = \Pi_V(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) = argmin_{\boldsymbol{y} \in V} \ \|\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t - \boldsymbol{y}\|_2$
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 $(\|\ell_t(\boldsymbol{x}) - \ell_t(\boldsymbol{u})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2)$
- 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: 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 gradient $\boldsymbol{g}_t$ 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

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

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.

$$
\begin{aligned}
\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F(\boldsymbol{u}) &\overset{\text{Jensen}}{\leq} \frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}[F(\boldsymbol{x}_t)] - F(\boldsymbol{u})) \\
&\overset{\text{convexity}}{\leq} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \mathbb{E}_t[\boldsymbol{g}_t], \boldsymbol{x}_t - \boldsymbol{u}\rangle] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}_t[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle]] \\
&\overset{\text{total expectation}}{=} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle] \\
&= \frac{\mathbb{E}[Regret_T(\boldsymbol{u})]}{T}
\end{aligned}
$$

□

Example: Stochastic Subgradient Descent

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

for $t = 1$ **to** T **do**

  Output $\boldsymbol{x}_t \in V$

  Receive stochastic gradient $\boldsymbol{g}_t$ such that $\mathbb{E}_t[\boldsymbol{g}_t] \in \partial F(\boldsymbol{x}_t)$

  $\boldsymbol{x}_{t+1} = \Pi_V(\boldsymbol{x}_t - \eta \boldsymbol{g}_t)$

end for

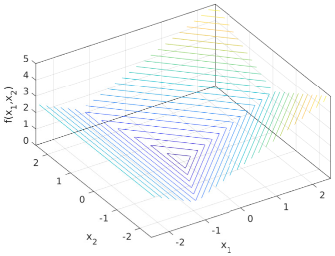
return $\bar{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$

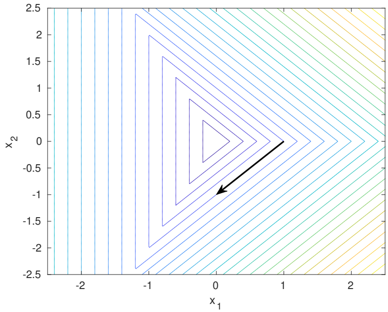
From the previous slides, we have

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

Beyond Online Subgradient Descent

Does Online Subgradient Descent Minimize the Functions? (1)

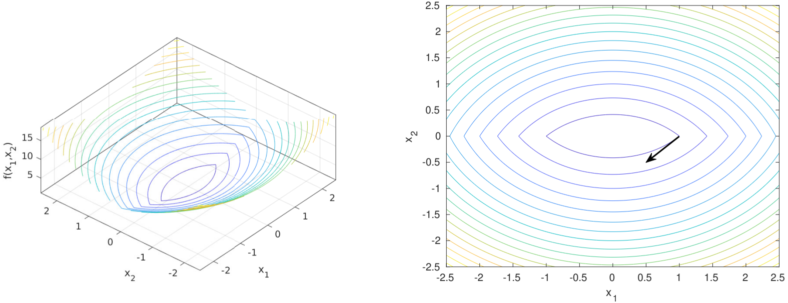




3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[-x_1, x_1 - x_2, x_1 + x_2]$. A negative subgradient is indicated by the black arrow

[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

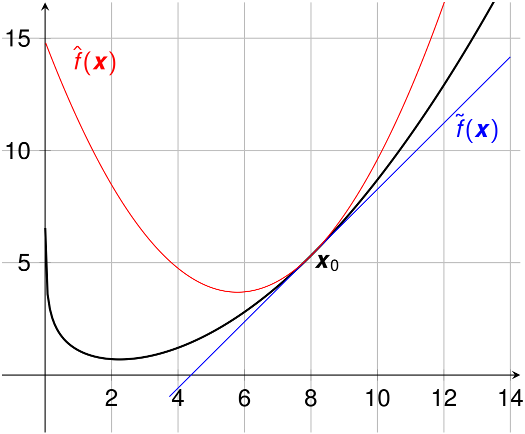
[Orabona, arXiv'19]

Intuition on OGD Update (1)

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

where Π_V is the Euclidean projection onto V, i.e., $\Pi_V(\boldsymbol{x}) = argmin_{\boldsymbol{y}\in V}\ \|\boldsymbol{x} - \boldsymbol{y}\|_2$

Intuition on OGD Update (2)



General Notion of Distances using Bregman Divergences

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

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. ψ 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

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

and we can change the last term with a Bregman Divergence

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

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

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

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

Strongly Convex Functions

Definition

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

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

Lemma (For OMD proof)

If ψ is λ-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 μ-strongly convex with respect to a norm $\|\cdot\|$ over its domain. Let $\boldsymbol{x}^\star = \mathrm{argmin}_{\boldsymbol{x}} \ f(\boldsymbol{x})$. Then, for all $\boldsymbol{x} \in \mathrm{dom}\, \partial f$, and $\boldsymbol{g} \in \partial f(\boldsymbol{x})$, we have

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

Regret Guarantee of OMD

Theorem

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

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

Proof.

One can show

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

The last term can be bounded as

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

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

Example: Online Subgradient Descent

- Set $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$
- ψ 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 $\boldsymbol{u}$:

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

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

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

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

Require: $\eta > 0$

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

Regret for any $\boldsymbol{u}$:

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

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

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

1: **for** $t = 1$ **to** T **do**

2: Output $\boldsymbol{x}_t \in argmin_{\boldsymbol{x} \in V} \ \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$

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

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

Proof.

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

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

An Explicit Regret with Strongly Convex Functions (1)

Lemma

Let $\psi_t : X \to \mathbb{R}$ and denote by $F_t(\boldsymbol{x}) = \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$. Assume $V \subseteq X$ be convex. Assume $\partial \ell_t(\boldsymbol{x}_t)$ to be non-empty and $F_t + \ell_t$ to be closed, subdifferentiable, and λ_t-strongly convex w.r.t. $\|\cdot\|$ in V. Then, we have

$$F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t) \le \|\boldsymbol{g}_t\|_\star^2/(2\lambda_t) + \psi_t(\boldsymbol{x}_{t+1}) - \psi_{t+1}(\boldsymbol{x}_{t+1}), \ \forall \boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t) \,.$$

Proof.

Define $\boldsymbol{x}_t^\star := argmin_{\boldsymbol{x} \in V} \ F_t(\boldsymbol{x}) + \ell_t(\boldsymbol{x})$, and $\boldsymbol{g}_t' \in \partial(F_t + \ell_t + i_V)(\boldsymbol{x}_t)$. Then

$$\begin{aligned}
&F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t) \\
&= (F_t(\boldsymbol{x}_t) + \ell_t(\boldsymbol{x}_t)) - (F_t(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_{t+1})) + \psi_t(\boldsymbol{x}_{t+1}) - \psi_{t+1}(\boldsymbol{x}_{t+1}) \\
&\le (F_t(\boldsymbol{x}_t) + \ell_t(\boldsymbol{x}_t)) - (F_t(\boldsymbol{x}_t^\star) + \ell_t(\boldsymbol{x}_t^\star)) + \psi_t(\boldsymbol{x}_{t+1}) - \psi_{t+1}(\boldsymbol{x}_{t+1}) \\
&\le \|\boldsymbol{g}_t'\|_\star^2/(2\lambda_t) + \psi_t(\boldsymbol{x}_{t+1}) - \psi_{t+1}(\boldsymbol{x}_{t+1}),
\end{aligned}$$

where in the second inequality we used the lemma in the previous slide. Observing that $\boldsymbol{x}_t = argmin_{\boldsymbol{x} \in V} \ F_t(\boldsymbol{x})$, we have $\mathbf{0} \in \partial(F_t + i_V)(\boldsymbol{x}_t)$. Hence, we have $\partial \ell_t(\boldsymbol{x}_t) \subseteq \partial(F_t + \ell_t + i_V)(\boldsymbol{x}_t)$. □

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

An Explicit Regret with Strongly Convex Functions (2)

Under the assumption of the previous slide and $\psi_{t+1}(\boldsymbol{x}) \geq \psi_t(\boldsymbol{x})$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \\
&= \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T}[F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t)] + F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u}) \\
&\leq \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T}\frac{\|\boldsymbol{g}_t\|_\star^2}{2\lambda_t}
\end{aligned}
$$

Example: Guessing Game

- In each round we have to guess a number y_t between 0 and 1
- Call your guess x_t
- Then, the y_t is revealed and you pay $\ell_t(x) = (x - y_t)^2$

- Use FTRL, no regularizer: $\boldsymbol{x}_t = argmin_{x \in V} \sum_{i=1}^{t-1} \ell_i(x) = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i$
- $\ell_t(x) + \sum_{i=1}^{t-1} \ell_i(x)$ is $2t$ strongly convex w.r.t. $|\cdot|$
- Gradient is $2(x_t - y_t)$, hence $|g_t| \leq 2$
- Regret of FTRL: $\sum_{t=1}^{T}(x_t - y_t)^2 - \sum_{t=1}^{T}(y_t - u)^2 \leq \frac{1}{2} \sum_{t=1}^{T} \frac{2}{t} \leq \ln T + 1$

FTRL with Linearized Losses

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

- I can run FTRL with <u>any</u> 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 $\boldsymbol{x}_t \in argmin_{\boldsymbol{x} \in V} \ \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \langle \boldsymbol{g}_i, \boldsymbol{x} \rangle$
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

- $V = \mathbb{R}^d$
- $\psi_{t+1}(\boldsymbol{x}) = \frac{1}{\eta_{t+1}}\|\boldsymbol{x}\|_2^2 \Rightarrow \psi_{t+1}$ is $\frac{1}{\eta_{t+1}}$-strongly convex w.r.t. $\|\cdot\|_2$
- $\boldsymbol{x}_{t+1} = argmin_{\boldsymbol{x}} \ \frac{1}{2\eta_{t+1}}\|\boldsymbol{x}\|_2^2 + \sum_{i=1}^t \langle \boldsymbol{g}_i, \boldsymbol{x}\rangle = -\eta_{t+1}\sum_{i=1}^t \boldsymbol{g}_i$
- 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 η_t

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

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

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

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

What is the optimal tuning of γ?

Parameter-free Online Algorithms

What is a Parameter-free Algorithm?

Definition

We define a parameter-free online convex optimization algorithm as one that achieves optimal regret uniformly for any competitor vector $\boldsymbol{u}$, up to logarithmic factors

Examples

- Exponentiated Gradient: $Regret_T(\boldsymbol{u}) \leq \frac{KL(\boldsymbol{u};\boldsymbol{\pi})}{\eta} + \frac{T\eta}{2} \Rightarrow$
  NormalHedge: $Regret_T(\boldsymbol{u}) = O(\sqrt{T(KL(\boldsymbol{u};\boldsymbol{\pi})+1)})$ [Chaudhuri et al., NeurIPS'09][Chernov&Vovk, UAI'10][Orabona&Pál, NeurIPS'16]
- OSD: $Regret_T(\boldsymbol{u}) \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{u}\|_2^2}{2\eta} + \frac{\eta T}{2} \Rightarrow$
  KT (next slides): $Regret_T(\boldsymbol{u}) = O(\|\boldsymbol{x}_1 - \boldsymbol{u}\|_2 \sqrt{T \ln(1 + T\|\boldsymbol{x}_1 - \boldsymbol{u}\|_2/\epsilon)} + \epsilon)$

Simple Parameter-free FTRL

Theorem

Consider the 1-d OCO problem, $g_t \in [-1, 1]$, $V = \mathbb{R}_{\geq 0}$. *Set* $\psi_t(x) = x\sqrt{T}(\ln x - 1) + \frac{(t-1)x}{\sqrt{T}}$. *Assume* $T \geq 4$. *Then, FTRL has regret*

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) \leq \sqrt{T}(1 + u \ln u) - \frac{u}{\sqrt{T}}$$

Moreover, $x_t = \exp(-\sum_{i=1}^{t-1} g_i - \frac{t-1}{T})$

- Compare it with OSD: $\mathit{Regret}_T(u) \leq \frac{1}{2}\sqrt{T}(u^2/\alpha + \alpha)$
- “Impossible” tuning of learning rate of OSD would give $\mathit{Regret}_T(u) \leq |u|\sqrt{T}$

- **Important:** We get almost the optimal regret, *uniformly for all u*
- **Important:** The algorithm goes exponential fast if the subgradients are all in the same direction

[New, proved just for these slides :)]

Simple Parameter-free FTRL (2)

The regularizer is not strongly convex! But it still works:

Proof.

The formula for x_t comes from the definition of the FTRL update.
Let $\theta_t = -\sum_{i=1}^{t-1} g_i$. Then, in the FTRL regret bound we have

$$
\begin{aligned}
&F(x_t) - F_{t+1}(x_{t+1}) + g_t x_t \\
&= \sqrt{T}\exp\left(\frac{\theta_t - g_t}{\sqrt{T}} - \frac{t}{T}\right) - \sqrt{T}\exp\left(\frac{\theta_t}{\sqrt{T}} - \frac{t-1}{T}\right) + g_t \exp\left(\frac{\theta_t}{\sqrt{T}} - \frac{t-1}{T}\right) \\
&= \sqrt{T}\exp\left(\frac{\theta_t - g_t}{\sqrt{T}} - \frac{t}{T}\right) - \sqrt{T}\exp\left(\frac{\theta_t}{\sqrt{T}} - \frac{t-1}{T}\right)\left(1 - g_t \frac{1}{\sqrt{T}}\right) \\
&\leq \sqrt{T}\exp\left(\frac{\theta_t - g_t}{\sqrt{T}} - \frac{t}{T}\right) - \sqrt{T}\exp\left(\frac{\theta_t}{\sqrt{T}} - \frac{t-1}{T}\right)\exp\left(-g_t \frac{1}{\sqrt{T}} - g_t^2 \frac{1}{T}\right) \leq 0
\end{aligned}
$$

where we use the elementary inequality $1 + y \geq \exp(y - y^2)$ for $|y| \leq 1/2$ □

Did We Only Gain a Constant in the Rate?

- $\sqrt{T}(1 + u \ln u)$ vs $\frac{1}{2}\sqrt{T}(u^2/\alpha + \alpha)$
- The rate did not change, and it might seem like we only improved a constant
- Not so fast!

Example: Logistic Regression

- Consider logistic regression on a dataset of T samples: $\min_{\boldsymbol{x}} \ F(\boldsymbol{x}) := \frac{1}{T}\sum_{t=1}^{T} \ln(1+\exp(-y_t\langle \boldsymbol{x}, \boldsymbol{z}_t\rangle))$
- Assume that the dataset is linearly separable with margin at least 1 by a hyperplane $\boldsymbol{u}^\star$
- Does the minimum exist? Does the minimizer exist?
- Rate of Averaged OSD with $\boldsymbol{x}_1 = \boldsymbol{0}$: $\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] - F(\boldsymbol{x}^\star) \leq \frac{1}{2\sqrt{T}}\left(\|\boldsymbol{x}^\star\|_2^2/\alpha + \alpha\right)$, is it vacuous?

- Rewrite it as $\mathbb{E}[F(\bar{\boldsymbol{x}}_T)] \leq \min_{\boldsymbol{u}} \ F(\boldsymbol{u}) + \frac{1}{2\sqrt{T}}\left(\|\boldsymbol{u}\|_2^2/\alpha + \alpha\right)$
- The r.h.s. can be upper bounded by $\boldsymbol{u} = \boldsymbol{u}^\star \ln \frac{2\alpha\sqrt{T}}{\|\boldsymbol{u}^\star\|_2}$ that gives

$$
\begin{aligned}
F(\boldsymbol{u}) &\leq \frac{1}{T}\sum_{t=1}^{T} \ln\left(1+\exp\left(-\ln\frac{2\alpha\sqrt{T}}{\|\boldsymbol{u}^\star\|_2}\right)\right) \leq \frac{1}{T}\sum_{t=1}^{T} \exp\left(-\ln\frac{2\alpha\sqrt{T}}{\|\boldsymbol{u}^\star\|_2}\right)\\
&= \frac{\|\boldsymbol{u}^\star\|_2}{2\alpha\sqrt{T}}
\end{aligned}
$$

- Overall, rate is $O(\frac{\ln T}{\sqrt{T}})$ and $\|\boldsymbol{u}\|_2 = O(\ln T)$, so not a constant!

[Ji&Telgarsky, COLT'19][Blogpost Feb'24]

Example: Regression with Kernels

- Consider a "universal kernel" $k(\cdot,\cdot)$, e.g., Gaussian kernel
- Universal kernels can approximate any continuous target function uniformly on any compact subset of the input space
- Consider linear regression with kernels
- Same thing will happen: the solution might be at infinity

- $\min_{\boldsymbol{u}\in\mathcal{H}_k} F(\boldsymbol{u}) + \frac{\|\boldsymbol{u}\|^2_{\mathcal{H}_k}}{\sqrt{T}} = O(T^{-a})$ where 'a' measures how "smooth" is the optimal solution [tons of refs! See, e.g., Ying&Pontil, 2008] (see also Taiji's slides)
- Again $\|\boldsymbol{u}\|^2$ is not a constant!
- A parameter-free algorithm will achieve optimal convergence in the parameter 'a' without, knowing it [Orabona, NeurIPS'14]

Brief History of Parameter-free Algorithms

- Streeter&McMahan [NeurIPS'12]: Only in 1 dimension, suboptimal bound, not a complete understanding
- McMahan&Abernethy [NeurIPS'13]: 1 dimension, minimax strategy but suboptimal formulation
- Orabona [NeurIPS'13]: Still suboptimal, but extended to any number of dimensions, even infinite
- Nemirovski [Personal Communication 2013]: Run GD with a grid of learning rates, select best solution: suboptimal bound, only deterministic
- McMahan&Orabona [COLT'14] and Orabona [NeurIPS'14]: Optimal bound, any number of dimensions, unintuitive proofs
- Orabona&Pál [NeurIPS'16]: Coin-betting view
- Carmon&Hider [COLT'22]: from $\ln(\|\boldsymbol{u}\|_2)$ to $\ln\ln(\|\boldsymbol{x}^\star\|_2)$ in the stochastic setting

See also Tutorial at ICML'20 on "Parameter-free Online Optimization"
`https://parameterfree.com/icml-tutorial/`

Better Parameter-Free through Duality on Guarantee

- Online-to-batch conversion (deterministic case for simplicity):

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

Theorem (McMahan&Orabona, COLT'14)

An algorithm that produces $\boldsymbol{x}_t$ based on $\boldsymbol{g}_1, \dots, \boldsymbol{g}_{t-1}$ guarantees

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

$$\Updownarrow$$

$$-\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \geq \psi_T^\star\left(-\sum_{t=1}^{T}\boldsymbol{g}_t\right), \ \forall \boldsymbol{g}_1, \dots, \boldsymbol{g}_T$$

where $\psi_T^\star$ is the Fenchel conjugate of ψ_T defined as $\psi_T^\star(\boldsymbol{\theta}) = \sup_{\boldsymbol{x}} \langle \boldsymbol{\theta}, \boldsymbol{x}\rangle - \psi_T(\boldsymbol{x})$

The Main Tool

- Assume $\|\boldsymbol{g}_t\|_2 \leq 1$
- Set $\boldsymbol{x}_t = \frac{-\sum_{i=1}^{t-1} \boldsymbol{g}_i}{t}\left(1 - \sum_{i=1}^{t}\langle \boldsymbol{g}_i, \boldsymbol{x}_i\rangle\right)$
- Claim: $\boldsymbol{x}_t$ guarantees

$$-\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \geq \psi_T^\star\left(-\sum_{t=1}^{T} \boldsymbol{g}_t\right)$$

where $\psi_T^\star(\boldsymbol{\theta}) \approx \frac{1}{\sqrt{T}} \exp\left(\frac{\|\boldsymbol{\theta}\|_2^2}{2T}\right) - 1$

- This implies $\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{u}\rangle \leq \|\boldsymbol{x}^\star\|_2\sqrt{T\ln(\|\boldsymbol{u}\|_2 T + 1)} + 1$

- Where does the inequality in orange come from?

Optimization through Optimal Gambling

Krichevsky&Trofimov (KT) betting strategy:

- Observe sequence of coins outcomes $c_t \in [-1, 1]$, start with \$1, bet on x_t money, win/lose $x_t c_t$
- On round t bet a signed fraction of your money equal to $\frac{\sum_{i=1}^{t-1} c_i}{t}$
- Exponential amount of money

$$\text{Winnings of KT} = 1 + \sum_{t=1}^{T} x_t c_t \geq \frac{\exp\left(\frac{(\sum_{t=1}^{T} c_t)^2}{2T}\right)}{2\sqrt{T}}$$

- No assumptions on the coin!

- We need to prove that $-\sum_{t=1}^{T} g_t x_t \geq \psi_T^\star\left(-\sum_{t=1}^{T} g_t\right)$
- In 1d, set $c_t = -g_t$ and assume $|g_t| \leq 1$ then we have it!
- It works in the vector case too

[Krichevsky&Trofimov, 1981][Orabona&Pal, NeurIPS'16]

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + \frac{-\sum_{i=1}^{t-1} \boldsymbol{g}_i}{t} \left(1 - \sum_{i=1}^{t-1} \langle \boldsymbol{g}_i, \boldsymbol{x}_i \rangle\right)$$

- No need to know the Lipschitz constant [Cutkosky, COLT'19]
- It works in any number of dimensions, even Hilbert spaces
- It works with stochastic subgradients
- It can work with constrained sets [Cutkosky&Orabona, COLT'18]

- It can adapt to the strong convexity in the stochastic setting (bounded stochastic subgradients and domain) [Cutkosky&Orabona, COLT'18]

Surprising Applications of Online Learning

Online Learning is Much More than Online Learning

- Online Convex Optimization might seem only concerned with losses&regret
- In reality, it is about proving inequalities on arbitrary sequences of data
- In my opinion, the inequalities are more important than the algorithms

- Here, I'll try to convince you of this view

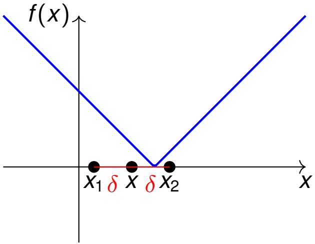
From Online Convex Optimization to Non-Convex Non-Smooth Optimization

Non-convex Optimization

- For convex optimization, we study $F(\boldsymbol{x}_T) - F(\boldsymbol{u})$
- For non-convex smooth optimization, we study $\mathbb{E}_i[\|\nabla F(\boldsymbol{x}_i)\|_2^2]$
- What can we do for non-smooth non-convex? Example: ConvNets with ReLUs

Definition (Zhang et al. ICML'20)

A point $\boldsymbol{x}$ is an (δ, ϵ)-stationary point of an almost-everywhere differentiable function F if there is a finite subset S of the ball of radius δ centered at $\boldsymbol{x}$ such that for $\boldsymbol{y}$ selected uniformly at random from S, $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$ and $\|\mathbb{E}[\nabla F(\boldsymbol{y})]\| \leq \epsilon$



If δ is small enough, it codifies our intuition on points close to a minimum

Well-Behaved Functions

We will assume that the functions are <u>well-behaved</u> in the sense that

$$F(\boldsymbol{y}) - F(\boldsymbol{x}) = \int_0^1 \langle \nabla F(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle dt$$

Up to perturbing the function with some noise, this holds for locally Lipschitz functions

Using OCO for Non-convex Optimization

Require: An OCO algorithm, duration of cycle K, initial point $\boldsymbol{x}_0$

$j = 0$
for $t = 1$ **to** T **do**
    if $mod(t, K) == 1$ **then**
        Reset OCO algorithm
        $j = j + 1$
        $\bar{\boldsymbol{x}}_j = \boldsymbol{0}$
    end if
    Receive $\boldsymbol{m}_t$ from OCO algorithm
    $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \boldsymbol{m}_t$
    Sample s_t uniformly in $[0, 1]$
    $\boldsymbol{x}'_t = \boldsymbol{x}_{t-1} - s_t \boldsymbol{m}_t$
    Pass $\ell_t(\boldsymbol{x}) = -\langle \nabla F(\boldsymbol{x}'_t), \boldsymbol{x} \rangle$ to OCO algorithm
    $\bar{\boldsymbol{x}}_j = \bar{\boldsymbol{x}}_j + \boldsymbol{x}'_t / K$
end for
return $\bar{\boldsymbol{x}}_J$ uniformly at random between 1 and T/K

- **Important:** The OCO algorithm decides the updates not the iterates

[Cutkosky et al., ICML'23]

Main Result

Theorem

Let the OCO algorithm be OGD over the L_2 ball of radius D. Then, we have

$$\mathbb{E}\left[\frac{1}{T/K}\sum_{i=1}^{T/K}\left\|\frac{1}{K}\sum_{t=1}^{K}\nabla F(\boldsymbol{x}'_{(i-1)K+t})\right\|_2\right] \le \frac{F(\boldsymbol{x}_0)-\inf_{\boldsymbol{x}} F(\boldsymbol{x})}{DT} + \frac{1}{\sqrt{K}}$$

Moreover, set $D=\delta/K$, $K=\left(\frac{T\delta}{F(\boldsymbol{x}_0)-\inf_{\boldsymbol{x}} F(\boldsymbol{x})}\right)^{\frac{2}{3}}$, and return $\bar{\boldsymbol{x}}_J$ where J is uniformly at random, then in expectation $\bar{\boldsymbol{x}}_j$ is $(\delta, O((T\delta)^{-\frac{1}{3}}))$-stationary point.

- The choice of D: $\bar{\boldsymbol{x}}_j$ is the average of K points at distance at most δ
- With the chosen D, we have

$$\frac{F(\boldsymbol{x}_0)-\inf_{\boldsymbol{x}} F(\boldsymbol{x})}{DT} + \frac{1}{\sqrt{K}} = \frac{K(F(\boldsymbol{x}_0)-\inf_{\boldsymbol{x}} F(\boldsymbol{x}))}{T\delta} + \frac{1}{\sqrt{K}}$$

- $\mathbb{E}\left[\frac{1}{T/K}\sum_{i=1}^{T/K}\left\|\frac{1}{K}\sum_{t=1}^{K}\nabla F(\boldsymbol{x}'_{(i-1)K+t})\right\|_2\right] =$
  $\mathbb{E}\left[\left\|\frac{1}{K}\sum_{t=1}^{K}\nabla F(\boldsymbol{x}'_{(J-1)K+t})\right\|_2\right] = O\left((T\delta)^{-\frac{1}{3}}\right)$

[Cutkosky et al., ICML'23]

From Function Value to Gradients

In all optimization analyses we need to link function values to gradients:

- Convex functions: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$
- Non-convex M-smooth: $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{M}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$

- What can we use for non-convex non-smooth?

Key Observation

- We evaluate the gradient in $\boldsymbol{x}_t' = \boldsymbol{x}_{t-1} - s_t \boldsymbol{m}_t = \boldsymbol{x}_{t-1} + s_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$
- Hence, we have

$$\mathbb{E}_{s_t} \nabla F(\boldsymbol{x}_t') = \int_0^1 \nabla F(\boldsymbol{x}_{t-1} + t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}))dt$$

- This allows us to say that

$$\begin{aligned} F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1}) &= \int_0^1 \langle \nabla F(\boldsymbol{x}_{t-1} + t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})), \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \rangle dt \\ &= \langle \mathbb{E}_{s_t}[\nabla F(\boldsymbol{x}_t')], \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \rangle \end{aligned}$$

- This holds without assuming convexity nor smoothness!

[Zhang et al. ICML'20][Cutkosky et al., ICML'23]

Proof

Proof.

Using the key observation, for the first cycle we have

$$F(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1}) = \langle \mathbb{E}_{s_t}[\nabla F(\boldsymbol{x}_t')], \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \rangle = -\langle \mathbb{E}_{s_t}[\nabla F(\boldsymbol{x}_t')], \boldsymbol{m}_t \rangle$$
$$= \langle -\mathbb{E}_{s_t}[\nabla F(\boldsymbol{x}_t')], \boldsymbol{m}_t - \boldsymbol{u} \rangle - \langle \mathbb{E}_{s_t}[\nabla F(\boldsymbol{x}_t')], \boldsymbol{u} \rangle$$

Taking full expectation, summing over $t = 1, \ldots, K$, for any $\|\boldsymbol{u}\|_2 \leq D$, we have

$$\mathbb{E}[F(\boldsymbol{x}_K)] - F(\boldsymbol{x}_0) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{K} \langle -\nabla F(\boldsymbol{x}_t'), \boldsymbol{m}_t - \boldsymbol{u} \rangle\right]}_{Regret_K(\boldsymbol{u})} - \mathbb{E}\left[\sum_{t=1}^{K} \langle \nabla F(\boldsymbol{x}_t'), \boldsymbol{u} \rangle\right]$$
$$\leq D\sqrt{K} - \mathbb{E}\left[\sum_{t=1}^{K} \langle \nabla F(\boldsymbol{x}_t'), \boldsymbol{u} \rangle\right]$$

Choose $\boldsymbol{u} = D\frac{\sum_{t=1}^{K} \nabla F(\boldsymbol{x}_t')}{\left\|\sum_{t=1}^{K} \nabla F(\boldsymbol{x}_t')\right\|_2}$ to have $\sum_{t=1}^{K} \langle \nabla F(\boldsymbol{x}_t'), \boldsymbol{u} \rangle = -D\left\|\sum_{t=1}^{K} \nabla F(\boldsymbol{x}_t')\right\|_2$.
Summing over the cycles and dividing by DT ends the proof. □

More Results

Using the same reduction, but possibly changing the online learning algorithm, we also show

- $(\delta, O((T\delta)^{-\frac{1}{3}})$ for the stochastic setting too
- For smooth stochastic functions, it implies that best rate of SGD
- For smooth deterministic, it matches the optimal rates

- Recently, Ahn et al. [arXiv'24] used this framework to show that Adam can be casted as an FTRL algorithm constructing the updates

[Cutkosky et al., ICML'23]

Only a Hack or Something Fundamental?

One might wonder if the above reduction is only a hack or it discovers something more fundamental.

One way to convince you is to take a look at the resulting procedure

$$
\begin{aligned}
\boldsymbol{x}_t &= \boldsymbol{x}_{t-1} - \boldsymbol{m}_t \\
\boldsymbol{g}_t &= \nabla F(\boldsymbol{x}_t + (s_t - 1)\boldsymbol{m}_t) \\
\boldsymbol{m}_{t+1} &= Clip_D(\boldsymbol{m}_t + \eta \boldsymbol{g}_t)
\end{aligned}
$$

We recovered a version of SGD with momentum and clipping! The only really different part is that we perturb the iterate a bit before calculating the gradient

From Online Learning to Concentration Inequalities

Estimating the Mean of Random Variables

A classic problem in statistic

- Suppose to have a stream of random variables in $[0, 1]$, $X_1, X_2, \ldots$
- Assume that their expectation conditioned on the past is μ
- We want to estimate μ and give confidence intervals that holds with high probability *uniformly over time*
  - Given that it is uniform over time, we can decide to stop based on the data

- In formulas, find $[a_t, b_t]$ such that $\Pr\{\mu \in [a_t, b_t], \forall t \geq 1\} \geq 1 - \delta$
- Moreover, the width of the confidence intervals should go to zero as $\sim \frac{\sigma}{\sqrt{t}}$

Usual Approach: Concentration Inequalities

- Estimate the true mean by the empirical mean $\hat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t} X_i$
- Use a concentration inequality that holds uniformly over time to construct confidence intervals for $\hat{\mu}_t$
- We obtain $\mu \in \left[\hat{\mu}_t \pm K\frac{\sigma\sqrt{\ln\frac{t}{\delta}}}{\sqrt{t}}\right]$ with probability at least $1-\delta$
- Examples of this approach: Maurer&Pontil [COLT'09] + union bound

Unfortunately, We Often Get Vacuous Estimates

- But, the above estimates are vacuous when the number of samples is small
- For example, $\mu \in [0.3 \pm 3.7]$
- In other words, our confidence intervals could be useless in the small sample regime
- Ideally, we want non-vacuous confidence intervals even with one sample!

- Our approach: derive concentration inequalities from online gambling algorithms!

Key Idea: Confidence Intervals from Betting

Fact 1 A concentration inequality says that the empirical average cannot be too far from the true expectation

Fact 2 Starting from \$1, you cannot gain money betting on a fair coin
Ville's inequality (1939): $\Pr\{\max_t \textit{Wealth}_t \geq \frac{1}{\delta}\} \leq \delta$

- Start with \$1
- "Imagine" using a betting algorithm to bet on the outcome of $X_i - \mu$
- Using KT we have $\textit{Wealth}_t \geq \frac{1}{2\sqrt{t}} \exp\left(\frac{(\sum_{i=1}^t (X_i-\mu))^2}{2t}\right)$
- Fact 1 + Fact 2 + KT:

$$\Pr\left\{\max_t \frac{1}{2\sqrt{t}} \exp\left(\frac{(\sum_{i=1}^t (X_i-\mu))^2}{2t}\right) \geq \frac{1}{\delta}\right\} \leq \Pr\left\{\max_t \textit{Wealth}_t \geq \frac{1}{\delta}\right\} \leq \delta$$

- Solve inequality: $\Pr\left\{\max_t \left|\mu - \frac{1}{t}\sum_{i=1}^t X_i\right| \geq \sqrt{\frac{2\ln\frac{2\sqrt{t}}{\delta}}{t}}\right\} \leq \delta$
- Equivalently, with probability at least $1-\delta$ and for any t we have

$$\left|\mu - \frac{1}{t}\sum_{i=1}^t X_i\right| \leq \sqrt{\frac{2\ln\frac{2\sqrt{t}}{\delta}}{t}}$$

[Jun&Orabona, COLT'19]

Game-Theoretic Probabilities and Concentrations

- Very general testing framework in Shafer&Vovk'05,'19 books, but no specific application to derive new concentrations
- Hendriks (arXiv'18) first to numerically evaluate a specific betting strategy to derive confidence intervals
- Waudby-Smith&Ramdas [arXiv'21] proposed to use heuristic betting algorithms

- Jun&Orabona [COLT'19] were the first ones to use regret guarantees of online betting algorithms to derive new concentrations inequalities
- Rakhlin&Sridharan (COLT'17) showed equivalent between martingale tail bounds and regret guarantees, but does not derive time-uniform concentrations because it does not use non-negative martingales
- Cover (Tech Report'74) recasted a statistical test as a betting game

- Next step is obvious: What we get using the optimal betting scheme?

PRECiSE: Portfolio REgret for Confidence SEquences

- Which betting algorithm should we use?
- We show that Universal Portfolio [Cover&Ordentlich, 1996] with 2 stocks is optimal for this setting
- We obtain a new time-uniform concentration: With probability at least $1-\delta$, for any t we have

$$\psi_t^\star \leq \ln \frac{1}{\delta} + \mathit{Regret}_t$$

where

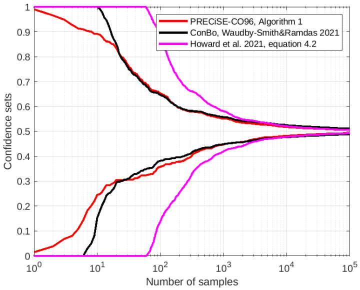
$$\psi_t^\star := \max_{\lambda \in [-\frac{1}{1-\mu}, \frac{1}{\mu}]} \sum_{i=1}^{t} \ln[1 + \lambda(X_i - \mu)] .$$

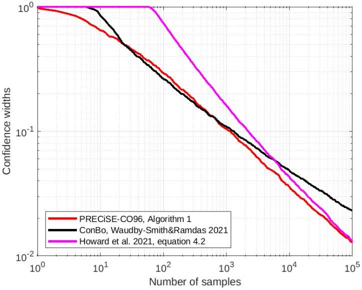
is the optimal log wealth with constant betting and $\mathit{Regret}_t \leq \ln \sqrt{t}$ is the regret of Universal Portfolio

- We prove that the set of μ that satisfy the inequality is an interval, so we can invert the concentration numerically using binary search
- Never vacuous: interval width less than $1 - \frac{\delta}{2}$

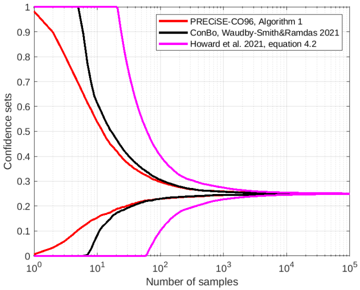
[Orabona&Jun, IEEE Trans. IT'23]

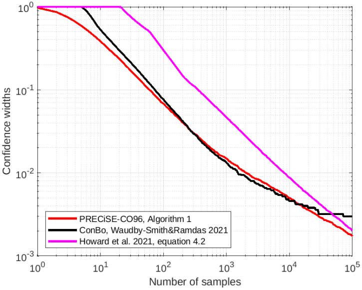
Experiments: Bernoulli(0.5)





Code: `https://github.com/bremen79/precise`





Code: `https://github.com/bremen79/precise`

From Online Betting to PAC-Bayes Bounds

PAC-Bayes Bounds

(Trying to follow Pierre's notation here!)
Definitions:

$$R(\theta) = \mathbb{E}_{(x,y)\sim P}[\ell(y, f_\theta(x))]$$

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y, f_\theta(X_i))$$

Assumption:
$0 \leq \ell \leq 1$

Theorem (McAllester, COLT'98)

Fix a prior distribution $\pi \in \mathcal{M}(\Theta)$. With probability at least $1-\delta$ on the data S, for any probability distribution ρ learnt on the data,

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R_n(\theta)] + \sqrt{\frac{KL(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{2n}}$$

Our Theorem

Theorem

Define optimal 'log-wealth' function:

$$\psi_n^\star(\theta) := \max_{\lambda \in [-\frac{1}{1-R(\theta)}, \frac{1}{R(\theta)}]} \sum_{i=1}^{n} \ln[1 + \lambda(\ell(Y_i, f_\theta(X_i)) - R(\theta))] \,.$$

Fix $\pi \in \mathcal{M}(\Theta)$, *then with probability at least* $1 - \delta$, ***simultaneously*** *for all n and* ρ*,*

$$\mathbb{E}_{\theta\sim\rho}[\psi_n^\star(\theta)] \leq \mathrm{KL}(\rho\|\pi) + \ln\frac{\sqrt{n}}{\delta} \,.$$

[Jang et al., COLT'23]

Consequences

I. By relaxing the 'log-wealth' term this inequality implies:

- McAllester's inequality [McAllester, COLT'98]
- Empirical Bernstein's PAC-Bayes inequality [Tolstikhin&Seldin, NeurIPS'13]
- Maurer's inequality of Bernoulli r.v.'s [Maurer, arXiv'04]
- Unexpected Bernstein's inequality [Mhammedi et al., NeurIPS'19]

II. With no relaxations, we can compute confidence sequences on μ_θ **efficiently**.

Examples of Relaxations

Our inequality:

$$\psi_n^\star(\theta) := \max_{\lambda \in [-\frac{1}{1-R(\theta)}, \frac{1}{R(\theta)}]} \sum_{i=1}^{n} \ln(1 + \lambda(\ell(Y_i, f_\theta(X_i)) - R(\theta))) \leq \mathrm{KL}(\rho\|\pi) + \ln \frac{\sqrt{n}}{\delta}$$

- $\ln(1+x) \geq x - x^2$ for $x \geq -0.68$ gives

$$|\mathbb{E}_{\theta\sim\rho}[R(\theta)] - \mathbb{E}_{\theta\sim\rho}[R_n(\theta)]| \leq 2\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{\sqrt{n}}{\delta}}{n}} \Rightarrow \text{McAllester's bound!}$$

- By convexity, $\max_\lambda \sum_{i=1}^n \ln\left(1 + \lambda(X_i - \mu)\right) \geq n\,\mathrm{kl}(\hat{\mu}, \mu)$, that gives

$$\mathrm{kl}\left(\mathbb{E}_{\theta\sim\rho}[R_n(\theta)], \mathbb{E}_{\theta\sim\rho}[R(\theta)]\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{\sqrt{n}}{\delta}}{n} \Rightarrow \text{Maurer's bound!}$$

- Similarly, you can get the other bounds too

Proof Sketch: Recall the Basic Bound

For any $\rho \ll \pi$ and measurable F:

$$\mathbb{E}_{\theta\sim\rho}[F(\theta)] \leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{\theta\sim\pi}[e^{F(\theta)}] \qquad \text{(Change-of-measure)}$$

For some fixed $\lambda > 0$ choose $F(\theta) = \lambda(R(\theta) - R_n(\theta))$. Then,

$$\begin{aligned}\lambda\mathbb{E}_{\theta\sim\rho}[R(\theta) - R_n(\theta)] &\leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{\theta\sim\pi}[e^{\lambda(R(\theta)-R_n(\theta))}] \\ &\leq \mathrm{KL}(\rho\|\pi) + \ln \frac{1}{\delta} + \ln \mathbb{E}_{\theta\sim\pi}[\mathbb{E}e^{\lambda(R(\theta)-R_n(\theta))}] \qquad \text{(Markov)}\end{aligned}$$

Concentration!

E.g., Hoeffding's lemma + tuning over λ gives McAllester's inequality.

Standard approach: λ is fixed. **Idea**: tune λ based on data...

... using an online betting game:

- A fictitious betting algorithm starts with wealth 1
- At round $i = 1, \ldots, n$ it bets a signed fraction of its wealth $B_i(\theta)$
- Observes outcome $\Delta_i(\theta) := \ell(Y_i, f_\theta(X_i)) - R(\theta)$
- Then it's log wealth is $\psi_n(\theta) := \sum_{i=1}^n \ln(1 + B_i(\theta)\Delta_i(\theta))$
- The regret of the algorithm is controlled, as before:

$$\psi_n^\star(\theta) - \psi_n(\theta) \leq \ln \sqrt{n}, \forall \theta$$

Recall that the optimal log-wealth is

$$\psi_n^\star(\theta) = \max_{\lambda \in [-\frac{1}{1-R(\theta)}, \frac{1}{R(\theta)}]} \sum_{i=1}^n \ln[1 + \lambda(\ell(Y_i, f_\theta(X_i)) - R(\theta))]$$

New Proof

For any $\rho \ll \pi$ and measurable F:

$$\mathbb{E}_{\theta\sim\rho}[F(\theta)] \le \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{\theta\sim\pi}[e^{F(\theta)}] \qquad \text{(Change-of-measure)}$$

Choose $F(\theta) = \psi_n(\theta, \mu_\theta)$ (optimal log-wealth). Then,

$$\mathbb{E}_{\theta\sim\rho}[\psi_n(\theta, \mu_\theta)] \le \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{\theta\sim\pi}[e^{\psi_n(\theta,\mu_\theta)}]$$

$$e^{\psi_n(\theta,\mu_\theta)} = \text{OptimalWealth} \le \text{WealthAnyOnlineAlgorithmA} \cdot \exp(\mathit{Regret}_n(A))$$

Concentration: WealthAnyOnlineAlgorithmA is a non-negative martingale

$$\Pr\left\{\sup_{n\ge 0} \text{WealthAnyOnlineAlgorithmA} \ge \frac{1}{\delta}\right\} \le \delta \qquad \text{(Ville's inequality)}$$

Putting all together

$$\mathbb{E}_{\theta\sim\rho}[\psi_n(\theta, \mu_\theta)] \le \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\exp(\ln\sqrt{n})\right) = \mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta} + \ln\sqrt{n}$$

Even More Surprising Applications

- Rademacher complexity bounds from Online Learning [Kakade et al., NeurIPS'08]
- From online learning to PAC-Bayes (but without the better bounds I showed) [Lugosi&Neu, arXiv'23]
- Better-than-KL PAC-Bayes bounds [Kuzborskij et al., arXiv'24]
- Parameter-free sampling [Sharrock&Nemeth, ICML'23][Sharrock et al. NeurIPS'23]

Summary

- Basic concepts and definitions of Online Learning
- OMD&FTRL
- Parameter-free algorithms
- Connection between regret guarantees and betting, concentrations, and generalization

Thank you!

Website: `https://francesco.orabona.com`

Blog: `https://parameterfree.com`

X/Twitter: @bremen79