Lecture 9. Optimistic Online Mirror Descent

Advanced Optimization (Fall 2025)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Optimistic OMD
- Deployment for Problem-dependent Regret
  - Small-loss bound
  - Gradient-variance bound
  - Gradient-variation bound
- Implications to Offline Optimization
  - Smooth optimization
  - Accelerated optimization

Part 1. Optimistic OMD

- Optimistic Online Learning
- Conceptual OMD
- Optimistic OMD

Optimistic Online Learning

- Standard (full-information) online learning protocol.

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 $\nabla f_t(\mathbf{x}_t)$, and further receives the optimistic vector M_{t+1}, and then updates the model.

- We need to encode ***"predictable"*** information in the update such that the overall algorithm can adapt to the niceness of environments.

Optimistic Online Mirror Descent

- Online Mirror Descent (OMD) provides a unified framework for online learning under the worst-case scenarios.

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

A Summary of OMD Deployment

- Our previous algorithms and regret can all be covered 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	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x}\|\mathbf{x}_t)$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T\log N})$

General Regret Analysis for OMD

Online Mirror Descent

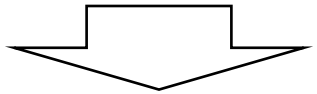
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

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*

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

Optimistic Online Mirror Descent

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



Optimistic Online Mirror Descent

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

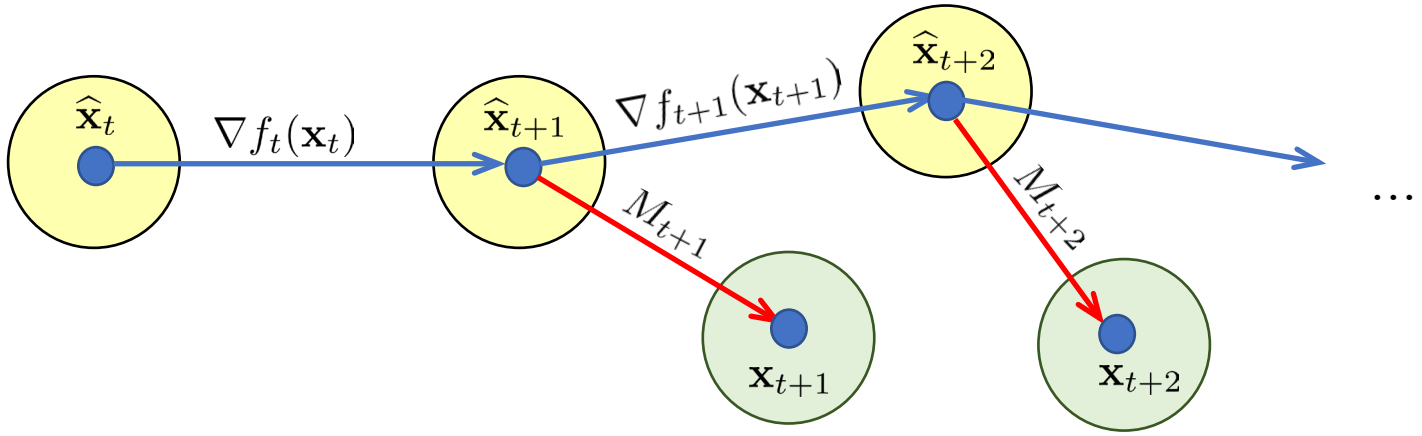
$$\widehat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_t) \right\}$$

where $M_t \in \mathbb{R}^d$ is the optimistic vector at round t (before seeing $\mathbf{x}_t$).

Understanding Optimistic OMD

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

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



Conceptual Online Mirror Descent

- Start with OMD:

OMD updates:

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

- An ideal situation: We can obtain the gradient for the next iteration in advance

Conceptual Online Mirror Descent

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

⇨ We can prove that $\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \mathcal{O}\left(\frac{1}{\eta}\right) = \mathcal{O}(1)$

Conceptual Online Mirror Descent

Conceptual Online Mirror Descent

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

⇨ We can prove that $\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \mathcal{O}\left(\frac{1}{\eta}\right) = \mathcal{O}(1)$

Proof.

By the first-order optimality $\langle \eta \nabla f_{t+1}(\mathbf{x}_{t+1}) + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle \le 0,$

Rearranging and applying the three-point lemma of Bregman divergence:

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

Summing up and telescoping: $\sum_{t=1}^T \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle \le \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_0)}{\eta}.$ □

Optimistic Online Mirror Descent

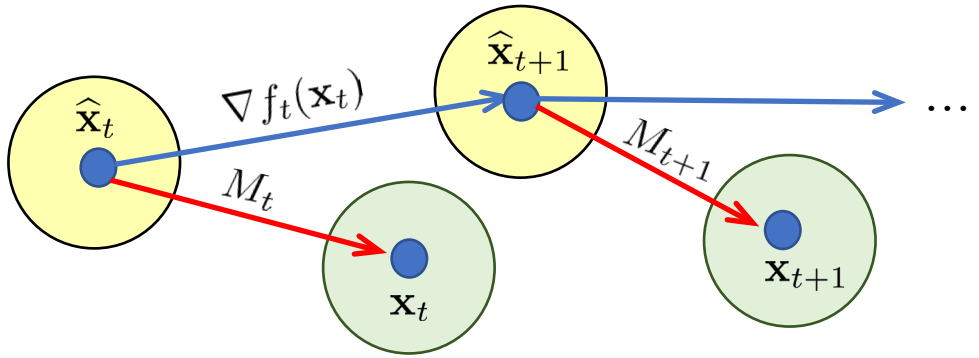
OMD: $\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \nabla f_{t-1}(\mathbf{x}_{t-1}), \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_{t-1}) \right\}$ ***Worst-case regret***

Conceptual OMD:

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

Constant regret, but infeasible

⇨ **Idea:** Use a guess M_t for $\nabla f_t(\mathbf{x}_t)$ in the lookahead update, then perform corrective update with the actual $\nabla f_t(\mathbf{x}_t)$



Optimistic Online Mirror Descent

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

Worst-case regret

Conceptual OMD:

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

Constant regret, but infeasible

Optimistic Online Mirror Descent

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

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

where $M_t \in \mathbb{R}^d$ is the optimistic vector at round t (before seeing $\mathbf{x}_t$).

Optimistic Online Mirror Descent

Optimistic Online Mirror Descent

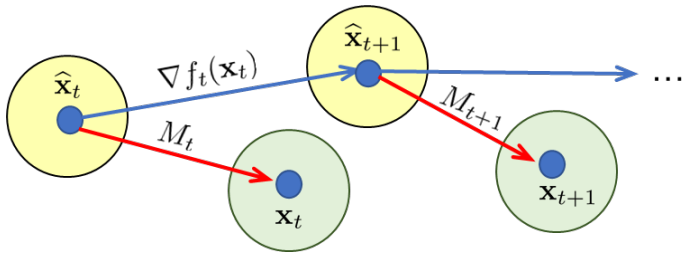
$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle M_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_t) \right\}$$

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

where $M_t \in \mathbb{R}^d$ is the optimistic vector at round t (before seeing $\mathbf{x}_t$).

Case 1. $M_t = \mathbf{0}$, recover OMD (i.e., $\mathbf{x}_t = \widehat{\mathbf{x}}_t$) ⇨ ***Worst-case regret***

Case 2. $M_t = \nabla f_t(\mathbf{x}_t)$, recover Conceptual OMD (i.e., $\mathbf{x}_t = \widehat{\mathbf{x}}_{t+1}$) ⇨ ***Constant regret***



⇨

The regret bound of Optimistic OMD should scale with the "estimate accuracy" of M_t for $\nabla f_t(\mathbf{x}_t)$

Optimistic OMD: Regret Analysis

Theorem 7 (Regret for Optimistic OMD). *Assume ψ is 1-strongly convex w.r.t. $\|\cdot\|$, the regret of Optimistic OMD w.r.t. any comparator $\mathbf{u} \in \mathcal{X}$ is bounded as:*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \sum_{t=1}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - M_t\right\|_\star^2 \quad \text{(quality of guess)}$$

$$+ \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_t) - \mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_{t+1}) \right) \quad \text{(telescoping term)}$$

$$- \sum_{t=1}^{T} \frac{1}{\eta_t} \left(\mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \mathbf{x}_t) + \mathcal{D}_\psi(\mathbf{x}_t, \widehat{\mathbf{x}}_t) \right). \quad \text{(negative term)}$$

The proof still relies on the ***stability lemma*** and the ***Bregman proximal inequality***, but now it requires taking the two-step updates (with optimism) into account.

Example: Optimistic OGD

- Consider the Euclidean regularizer $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, i.e.,:

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta \langle M_t, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2$$

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

$$\Longrightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \underbrace{\eta \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2}_{\text{(quality of guess)}} + \frac{\|\mathbf{u} - \widehat{\mathbf{x}}_1\|_2^2}{2\eta} \underbrace{- \frac{1}{4\eta}\sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2}_{\text{(negative term)}}$$

$$\le \eta \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 + \frac{D^2}{2\eta} \le \mathcal{O}\left(\sqrt{1 + \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2}\right)$$

$\left(\eta = \frac{D}{\sqrt{\sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2}}\right.$, which is not available)

→ ***self-confident tuning***

Part 2. Deploying Optimistic OMD

- Small-Loss Bound
- Gradient-Variance Bound
- Gradient-Variation Bound

Part 2. Deploying Optimistic OMD

- Small-Loss Bound
- Gradient-Variance Bound
- Gradient-Variation Bound

Small-Loss Bound

- Recall the guarantee of optimistic OGD:

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

- Consider the Euclidean regularizer $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, i.e.,:

$$\Longrightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \mathcal{O}\left(\sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2} \right)$$

$$\text{Setting } M_t = 0 \Longrightarrow \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \mathcal{O}\left(\sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2^2} \right)$$

Small-Loss Bound

Lemma. *For $x, y, a \in \mathbb{R}_+$ that satisfy $x - y \le \sqrt{ax}$, it implies $x - y \le \sqrt{ay} + a$.*

- Employing the ***self-bounding property*** of smooth and non-negative functions.

Corollary 1. *For an L-smooth and non-negative function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we have that*

$$\|\nabla f(\mathbf{x})\|_2 \le \sqrt{2Lf(\mathbf{x})}, \quad \forall \mathbf{x} \in \mathcal{X}.$$

Setting $M_t = 0$ in Optimistic OMD (with Euclidean regularizer):

$$\Longrightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \mathcal{O}\left(\sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2^2}\right) \le \mathcal{O}\left(\sqrt{1 + L\sum_{t=1}^{T} f_t(\mathbf{x}_t)}\right) \quad \text{(self-bounding property)}$$

$$\Longrightarrow \quad \mathrm{REG}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) = \mathcal{O}\left(D\sqrt{L\sum_{t=1}^{T} f_t(\mathbf{u}) + 1 + G^2}\right). \quad \text{(converting trick)}$$

□

Small-Loss Bound

- Since we are using optimistic OMD with a fixed step size, the algorithm requires $G_T \triangleq \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2^2$ when achieving small-loss bound.
- This is can be rectified by the ***self-confident tuning***. We can use the optimistic OMD with time-varying step sizes.

Theorem 8 (Small-loss Bound). *Assume that $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ and f_t is L-smooth and non-negative for all $t \in [T]$, when setting $\eta_t = \frac{D}{\sqrt{1+G_t}}$ and $M_t = \mathbf{0}$, the regret of Optimistic OMD to any comparator $\mathbf{u} \in \mathcal{X}$ is bounded as*

$$\textsc{Reg}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \mathcal{O}\left(\sqrt{1 + F_T}\right),$$

where $G_t = \sum_{s=1}^{t} \|\nabla f_s(\mathbf{x}_s)\|_2^2$ is the empirical cumulative gradient norm.

Small-Loss Bound

Proof. $\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \sum_{t=1}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - M_t\right\|_2^2$ (quality of guess, `term(a)`)

$$+ \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2 \right)$$ (telescoping term, `term(b)`)

$$- \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\widehat{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2 \right)$$ (negative term, `term(c)`)

For term (a),

$$\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 = D \sum_{t=2}^{T} \frac{\|\nabla f_t(\mathbf{x}_t)\|_2^2}{\sqrt{1 + G_t}} + G^2 \leq 2D \sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2^2} + G^2$$

(self-confident tuning lemma)

$$\leq D \sqrt{1 + 2L \sum_{t=1}^{T} f_t(\mathbf{x}_t)} + G^2$$ (self-bounding property)

Small-Loss Bound

Proof.

$$\texttt{term (b)} = \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2 \right)$$

$$\leq \frac{1}{2\eta_T} \sum_{t=1}^{T} \left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2 \right) \quad (\{\eta_1, \dots, \eta_T\} \text{ decreasing step size})$$

$$\leq \frac{1}{2\eta_T} \|\mathbf{u} - \widehat{\mathbf{x}}_1\|_2^2 \quad \text{(telescoping)}$$

$$\leq \frac{D}{2} \sqrt{1 + 2L \sum_{t=1}^{T} f_t(\mathbf{x}_t)} + \frac{D}{2} \quad \left(\text{by def of } \eta_T = \tfrac{D}{\sqrt{1+G_T}} \text{ and domain boundedness}\right)$$

$$\Longrightarrow \quad \textsc{Reg}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq 3D \sqrt{1 + 2L \sum_{t=1}^{T} f_t(\mathbf{x}_t)} + G^2 \quad \leq \mathcal{O}\left(D \sqrt{L \sum_{t=1}^{T} f_t(\mathbf{u}) + 1} + G^2 \right).$$

(converting trick)

□

Part 2. Deploying Optimistic OMD

- Small-Loss Bound

- Gradient-Variance Bound

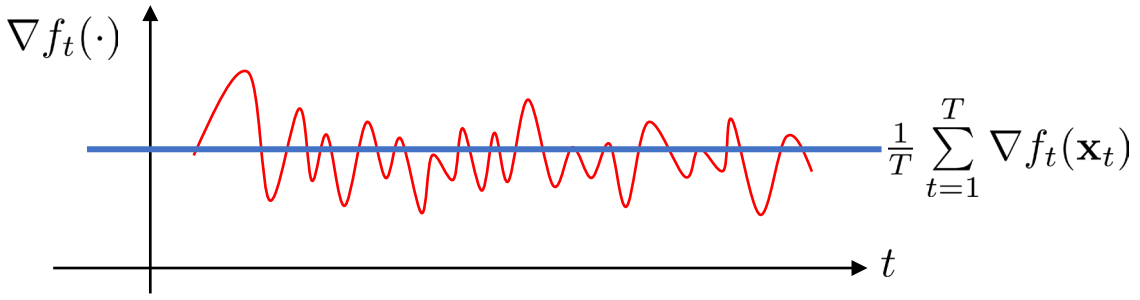
- Gradient-Variation Bound

Gradient-Variance Bound

Definition 2 (Gradient Variance). Let T be the time horizon and $\mathcal{X} \subseteq \mathbb{R}^d$ be the feasible domain. For the function sequence $f_1, \ldots, f_T$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$ for $t \in [T]$, its gradient variance is defined as

$$\mathrm{Var}_T = \sup_{\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}\in\mathcal{X}} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2$$

where $\boldsymbol{\mu}_T \triangleq \arg\min_{\boldsymbol{\mu}} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}\|_2^2 = \frac{1}{T}\sum_{t=1}^{T} \nabla f_t(\mathbf{x}_t)$.



Implicit prior on the enviornment:
there exists a ***latent mean gradient*** $\mathbb{E}[\nabla f_t(\mathbf{x}_t)]$.

e.g. SGD (sampled from a set of data)

e.g. Classification (sampled from training set)

Gradient-Variance Bound

Definition 2 (Gradient Variance). Let T be the time horizon and $\mathcal{X} \subseteq \mathbb{R}^d$ be the feasible domain. For the function sequence $f_1, \ldots, f_T$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$ for $t \in [T]$, its gradient variance is defined as

$$\mathrm{Var}_T = \sup_{\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}\in\mathcal{X}} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2$$

where $\boldsymbol{\mu}_T \triangleq \frac{1}{T}\sum_{t=1}^{T} \nabla f_t(\mathbf{x}_t)$ is the gradient mean.

Optimistic Online Mirror Descent

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle M_t, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2 \right\}$$

How to choose M_t?

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

Gradient-Variance Bound

Definition 2 (Gradient Variance). Let T be the time horizon and $\mathcal{X} \subseteq \mathbb{R}^d$ be the feasible domain. For the function sequence $f_1, \ldots, f_T$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$ for $t \in [T]$, its gradient variance is defined as

$$\mathrm{Var}_T = \sup_{\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}\in\mathcal{X}} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2$$

where $\boldsymbol{\mu}_T \triangleq \frac{1}{T}\sum_{t=1}^{T} \nabla f_t(\mathbf{x}_t)$ is the gradient mean.

Optimistic Online Mirror Descent

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle M_t, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2 \right\}$$

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

self-confident estimate of gradient mean: $\boldsymbol{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} \nabla f_s(\mathbf{x}_s)$

Gradient-Variance Bound

Theorem 9 (gradient-variance bound). *Assume that* $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, *when setting* $\eta_t = \frac{D}{\sqrt{1+\widetilde{\mathrm{Var}}_{t-1}}}$ *and* $M_t = \boldsymbol{\mu}_{t-1}$, *the regret of Optimistic OMD to any comparator* $\mathbf{u} \in \mathcal{X}$ *is bounded as*

$$\textsc{Reg}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \widetilde{\mathcal{O}}\left(\sqrt{1+\mathrm{Var}_T}\right)$$

where $\widetilde{\mathrm{Var}}_{t-1} = \sum_{s=1}^{t-1} \|\nabla f_s(\mathbf{x}_s) - \boldsymbol{\mu}_s\|_2^2$ *is the self-confident estimate of variance* Var_T, *and* $\boldsymbol{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} \nabla f_s(\mathbf{x}_s)$ *is the empirical gradient mean.*

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

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

(negative term)

Gradient-Variance Bound

Proof. For term (a),

$$\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 = \sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_{t-1}\right\|_2^2 + G^2 \qquad (\eta_1 \triangleq 1)$$

$$\leq 2\sum_{t=2}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2 + 2\sum_{t=2}^{T} \eta_t \left\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\right\|_2^2 + G^2$$

$$\leq 2D\sum_{t=2}^{T} \frac{\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2}{\sqrt{1 + \sum_{s=1}^{t-1} \|\nabla f_s(\mathbf{x}_s) - \boldsymbol{\mu}_s\|_2^2}} + 2D\sum_{t=2}^{T} \frac{9G^2}{t^2} + G^2 \qquad \left(\boldsymbol{\mu}_t = \frac{(t-1)\boldsymbol{\mu}_{t-1} + \nabla f_t(\mathbf{x}_t)}{t}\right)$$

$$(\|\boldsymbol{\mu}_t\|_2 \leq G, \forall t \in [T])$$

$$(\eta_t \leq D, \forall t \in [T])$$

$$\leq 2D\sum_{t=2}^{T} \frac{\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2}{\sqrt{1 + \sum_{s=1}^{t-1} \|\nabla f_s(\mathbf{x}_s) - \boldsymbol{\mu}_s\|_2^2}} + 18DG^2 \cdot \frac{\pi^2}{6} + G^2 \qquad \left(\sum_{x=1}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}\right)$$

Gradient-Variance Bound

Proof. $\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \le 2D \sum_{t=2}^{T} \frac{\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2}{\sqrt{1 + \sum_{s=1}^{t-1} \|\nabla f_s(\mathbf{x}_s) - \boldsymbol{\mu}_s\|_2^2}} + 18DG^2 \cdot \frac{\pi^2}{6} + G^2$

Lemma 4. *Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. Then*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{1 + \sum_{s=1}^{t-1} a_s}} \le 4\sqrt{1 + \sum_{t=1}^{T} a_t} + \max_{t \in [T]} a_t.$$

$\Rightarrow$ $\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \le 8D\sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2} + 8DG^2 + 18DG^2 \cdot \frac{\pi^2}{6} + G^2$

Recall that our goal is to obtain $\mathcal{O}\left(\sqrt{\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2}\right)$

Gradient-Variance Bound

Proof. $\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \leq 8D\sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2} + 8DG^2 + 18DG^2 \cdot \frac{\pi^2}{6} + G^2$

We need to measure the gap between $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2$ and $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2$

Let us consider ***another online learning process***: the online function is $h_t : \mathbb{R}^d \mapsto \mathbb{R}$,

$$h_t(\mathbf{a}) = \frac{1}{2} \|\nabla f_t(\mathbf{x}_t) - \mathbf{a}\|_2^2 ,$$

which is evidently a 1-strongly convex function with respect to $\|\cdot\|_2$.

Consider OGD over $\{h_t\}_{t=1}^T$ with step size $\{\eta_t\}_{t=1}^T$, which updates by

$$\mathbf{a}_{t+1} = \mathbf{a}_t - \eta_t \nabla h_t(\mathbf{a}_t) = \mathbf{a}_t - \eta_t \left(\mathbf{a}_t - \nabla f_t(\mathbf{x}_t)\right) = (1 - \eta_t)\mathbf{a}_t + \eta_t \nabla f_t(\mathbf{x}_t) \qquad (\bigstar)$$

Gradient-Variance Bound

Proof. $\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \le 8D\sqrt{1 + \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2} + 8DG^2 + 18DG^2 \cdot \frac{\pi^2}{6} + G^2$

We need to measure the gap between $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2$ and $\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2$

Consider OGD over $\{h_t\}_{t=1}^T$ with step size $\{\eta_t\}_{t=1}^T$, which updates by

$$\mathbf{a}_{t+1} = (1 - \eta_t)\mathbf{a}_t + \eta_t \nabla f_t(\mathbf{x}_t) \qquad (\bigstar)$$

On the other hand, by definition of gradient mean, we have

$$\boldsymbol{\mu}_t = \frac{t-1}{t}\boldsymbol{\mu}_{t-1} + \frac{1}{t}\nabla f_t(\mathbf{x}_t) \qquad \left(\boldsymbol{\mu}_t = \tfrac{1}{t}\textstyle\sum_{s=1}^{t} \nabla f_s(\mathbf{x}_s)\right)$$

Thus, set $\mathbf{a}_1 = \mathbf{0}$, $\eta_t = \frac{1}{t}$, then $\{\mathbf{a}_{t+1}\}_{t=1}^{T-1}$ sequence is ***equivalent*** to $\{\boldsymbol{\mu}_t\}_{t=1}^{T-1}$ sequence.

To summarize, we have $\mathbf{a}_{t+1} = \boldsymbol{\mu}_t$ for $t = 1, \ldots, T-1$.

Gradient-Variance Bound

Proof. $h_t(\mathbf{a}) = \frac{1}{2}\|\nabla f_t(\mathbf{x}_t) - \mathbf{a}\|_2^2$, $\mathbf{a}_{t+1} \overset{(\bigstar)}{=} (1-\eta_t)\mathbf{a}_t + \eta_t \nabla f_t(\mathbf{x}_t)$, $\boldsymbol{\mu}_t = \frac{t-1}{t}\boldsymbol{\mu}_{t-1} + \frac{1}{t}\nabla f_t(\mathbf{x}_t)$

Thus, set $\mathbf{a}_1 = \mathbf{0}$, $\eta_t = \frac{1}{t}$, then $\{\mathbf{a}_{t+1}\}_{t=1}^{T-1}$ sequence is ***equivalent*** to $\{\boldsymbol{\mu}_t\}_{t=1}^{T-1}$ sequence.

Since $(\bigstar)$ is essentially OGD for 1-strongly convex, whose guarantee is:

$$
\begin{aligned}
\textsc{Reg}(\{h_t\}_{t=1}^{T-1}) &= \sum_{t=1}^{T-1} h_t(\boldsymbol{\mu}_t) - \sum_{t=1}^{T-1} h_t(\boldsymbol{\mu}) && \text{(holds for any point } \boldsymbol{\mu} \text{ in } \mathbb{R}^d) \\
&= \sum_{t=1}^{T-1} \frac{1}{2}\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2 - \sum_{t=1}^{T-1} \frac{1}{2}\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2 && \text{(taking } \boldsymbol{\mu}_T \text{ as the comparator)} \\
&\le \frac{(2G)^2}{2\alpha}(1 + \ln(T-1)) && \text{(regret bound of } \alpha\text{-strongly convex function does not rely} \\
&\le 2G^2(1 + \ln T) && \text{on domain diameter)}
\end{aligned}
$$

Gradient-Variance Bound

Proof. $h_t(\mathbf{a}) = \frac{1}{2}\|\nabla f_t(\mathbf{x}_t) - \mathbf{a}\|_2^2$, $\mathbf{a}_{t+1} \overset{(\bigstar)}{=} (1-\eta_t)\mathbf{a}_t + \eta_t \nabla f_t(\mathbf{x}_t)$, $\boldsymbol{\mu}_t = \frac{t-1}{t}\boldsymbol{\mu}_{t-1} + \frac{1}{t}\nabla f_t(\mathbf{x}_t)$

Thus, set $\mathbf{a}_1 = \mathbf{0}$, $\eta_t = \frac{1}{t}$, then $\{\mathbf{a}_{t+1}\}_{t=1}^{T-1}$ sequence is ***equivalent*** to $\{\boldsymbol{\mu}_t\}_{t=1}^{T-1}$ sequence.

Since $(\bigstar)$ is essentially OGD for 1-strongly convex, whose guarantee is:

$$\textsc{Reg}(\{h_t\}_{t=1}^{T-1}) = \sum_{t=1}^{T-1} \frac{1}{2}\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2 - \sum_{t=1}^{T-1} \frac{1}{2}\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2 \le 2G^2(1+\ln T)$$

$$\Longrightarrow \quad \sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \le 8D\sqrt{1 + \sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2} + 8DG^2 + 18DG^2 \cdot \frac{\pi^2}{6} + G^2$$

$$\le 8D\sqrt{1 + \sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_T\|_2^2 + 4G^2(1+\ln T)} + 8DG^2 + 18DG^2 \cdot \frac{\pi^2}{6} + G^2$$

Gradient-Variance Bound

Proof. We then analyze term (b) in the same way as before:

$$
\begin{aligned}
\texttt{term (b)} &= \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2 \right) \\
&= \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 + \frac{1}{2\eta_1} \|\mathbf{u} - \widehat{\mathbf{x}}_1\|_2^2 \\
&\leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) D^2 + \frac{1}{2\eta_1} D^2 \quad (\eta_t \leq \eta_{t-1} \text{ and } \|\mathbf{x} - \mathbf{y}\|_2 \leq D, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}) \\
&\leq \frac{D^2}{2\eta_T} + \frac{1}{2\eta_1} D^2 \leq \frac{D}{2} \sqrt{1 + \mathrm{Var}_T} + \frac{D}{2} \qquad \left(\frac{1}{\eta_T} = \frac{\sqrt{1+\widetilde{\mathrm{Var}}_{T-1}}}{D} \leq \frac{\sqrt{1+\mathrm{Var}_T}}{D}\right)
\end{aligned}
$$

Gradient-Variance Bound

Proof. Finally, putting three terms together achieves

$$\texttt{term}\,(\texttt{a}) \le 8D\sqrt{1+\mathrm{Var}_T + 4G^2(1+\ln T)} + (39D+1)G^2$$

$$\texttt{term}\,(\texttt{b}) \le \frac{D^2}{2\eta_T} + \frac{1}{2\eta_1}D^2 \le \frac{D}{2}\sqrt{1+\mathrm{Var}_T} + \frac{D}{2}$$

$$\texttt{term}\,(\texttt{c}) \ge 0$$

$$\Rightarrow \quad \mathrm{REG}_T = \texttt{term}\,(\texttt{a}) + \texttt{term}\,(\texttt{b}) - \texttt{term}\,(\texttt{c})$$
$$\le 9D\sqrt{1+\mathrm{Var}_T + 4G^2(1+\ln T)} + 39DG^2 + G^2 = \widetilde{\mathcal{O}}\left(\sqrt{1+\mathrm{Var}_T}\right).$$

Part 2. Deploying Optimistic OMD

- Small-Loss Bound
- Gradient-Variance Bound
- Gradient-Variation Bound

Gradient-Variation Bound

Definition 3 (Gradient Variation). Let T be the time horizon and $\mathcal{X} \subseteq \mathbb{R}^d$ be the feasible domain. For the function sequence $f_1, \ldots, f_T$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$ for $t \in [T]$, its gradient variation is defined as

$$V_T = \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|_2^2$$

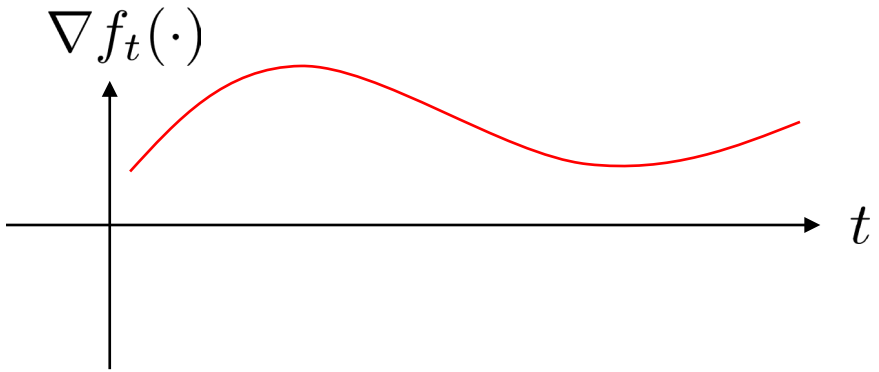
Gradient variation characterizes online functions' ***shifting intensity***.

- ***Adaptivity***: it can be small in slowly changing environments.
- ***Robustness***: $V_T \leq 4G^2T$ in the worst case. $(\|\nabla f_t(\mathbf{x})\| \leq G, \forall \mathbf{x} \in \mathcal{X} \text{ and } t \in [T])$

Gradient-Variation Bound

Definition 3 (Gradient Variation). Let T be the time horizon and $\mathcal{X} \subseteq \mathbb{R}^d$ be the feasible domain. For the function sequence $f_1, \ldots, f_T$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$ for $t \in [T]$, its gradient variation is defined as

$$V_T = \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|_2^2$$



Implicit assumption:

Gradient (online function) shifts slowly

e.g., age forecasting by portraits

Optimistic OMD for Gradient-Variation Bound

Optimistic Online Mirror Descent

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle M_t, \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2 \right\}$$

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

Question: How to choose M_t?

⇨ Imposing a prior on the change of the online functions

*setting M_t as the **last-round gradient*** $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$

Optimistic OMD for Gradient-Variation Bound

Optimistic Online Mirror Descent

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle M_t, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2$$

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

Optimistic OMD for Gradient-Variation Bound

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

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

Gradient-Variation Bound

Theorem 10 (Gradient Variation Regret Bound). *Assume that* $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ *and* f_t *is L-smooth for all* $t \in [T]$, *when setting* $\eta_t = \min\{\frac{1}{4L}, \frac{D}{\sqrt{1+\widetilde{V}_{t-1}}}\}$ *and* $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$, *the regret of Optimistic OMD to any comparator* $\mathbf{u} \in \mathcal{X}$ *is*

$$\textsc{Reg}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \mathcal{O}\left(\sqrt{1+V_T}\right)$$

where $\widetilde{V}_{t-1} = \sum_{s=2}^{t-1} \|\nabla f_s(\mathbf{x}_{s-1}) - \nabla f_{s-1}(\mathbf{x}_{s-1})\|_2^2$ *is the empirical estimates of* V_t.

Proof. $$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 + \sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2\right)$$
$$- \sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|\widehat{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2\right)$$
(negative term)

Proof

Proof. For term (a),

$$\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \leq \sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\right\|_2^2 + G^2 \qquad (\eta_1 \triangleq 1)$$

$$\leq 2\sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_{t-1})\right\|_2^2 + 2\sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\right\|_2^2 + G^2$$

$$\leq 2\sum_{t=2}^{T} \eta_t L^2 \left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|_2^2 + 2D\sum_{t=2}^{T} \frac{\left\|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\right\|_2^2}{\sqrt{1 + \sum_{s=2}^{t-1} \left\|\nabla f_s(\mathbf{x}_{s-1}) - \nabla f_{s-1}(\mathbf{x}_{s-1})\right\|_2^2}} + G^2$$

(L-smooth)

Proof

Proof. For term (a),

$$\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \le \sum_{t=2}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2 + G^2 \qquad (\eta_1 \triangleq 1)$$

$$\le 2\sum_{t=2}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_{t-1})\|_2^2 + 2\sum_{t=2}^{T} \eta_t \|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2 + G^2$$

$$\le 2\sum_{t=2}^{T} \eta_t L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 + 2D \sum_{t=2}^{T} \frac{\|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2}{\sqrt{1 + \sum_{s=2}^{t-1} \|\nabla f_s(\mathbf{x}_{s-1}) - \nabla f_{s-1}(\mathbf{x}_{s-1})\|_2^2}} + G^2$$

(L-smooth)

Lemma 4. *Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. Then*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{1 + \sum_{s=1}^{t-1} a_s}} \le 4\sqrt{1 + \sum_{t=1}^{T} a_t + \max_{t\in[T]} a_t}.$$

Proof

Proof. $\texttt{term (a)} \leq 2\sum_{t=2}^{T} \eta_t L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 + 8D\sqrt{1 + \sum_{t=2}^{T} \|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2} + (4D+1)G^2$

$$\leq 2\sum_{t=2}^{T} \eta_t L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 + 8D\sqrt{1 + V_T} + (4D+1)G^2$$

$(V_T = \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|_2^2)$

This term depends on our algorithm,
how to deal with it?

Proof

Proof. For the term (c), we have

$$\texttt{term (c)} = \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\widehat{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2 \right)$$

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

$$\geq \sum_{t=2}^{T} \frac{1}{2\eta_{t-1}} \left(\|\widehat{\mathbf{x}}_t - \mathbf{x}_{t-1}\|_2^2 + \|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 \right) \qquad \left(\tfrac{1}{\eta_t} \geq \tfrac{1}{\eta_{t-1}}\right)$$

$$\geq \sum_{t=2}^{T} \frac{1}{4\eta_{t-1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \qquad (a^2 + b^2 \geq (a+b)^2/2)$$

Does this term look familiar?

Proof

Proof. We then analysis term (b),

$$
\begin{aligned}
\texttt{term (b)} &= \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2 \right) \\
&\leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 + \frac{1}{2\eta_1} \|\mathbf{u} - \widehat{\mathbf{x}}_1\|_2^2 \\
&\leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) D^2 + \frac{1}{2\eta_1} D^2 \quad (\eta_t \leq \eta_{t-1} \text{ and } \|\mathbf{x} - \mathbf{y}\|_2 \leq D, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}) \\
&\leq \frac{D^2}{2\eta_T} \qquad \text{noting that } \eta_T = \min\left\{ \tfrac{1}{4L}, \tfrac{D}{\sqrt{1+\widetilde{V}_{T-1}}} \right\} \geq \min\left\{ \tfrac{1}{4L}, \tfrac{D}{\sqrt{1+V_T}} \right\} \\
&\leq \frac{1}{2} \max\{4LD^2, D\sqrt{1+V_T}\}
\end{aligned}
$$

Proof

Proof. Finally, putting three terms together yields

$$\texttt{term (a)} \le 2\sum_{t=2}^{T} \eta_t L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 + 4D\sqrt{1+V_T} + (4D+1)G^2$$

$$\texttt{term (b)} \le \frac{1}{2}\max\{4LD, D\sqrt{1+V_T}\}$$

$$\texttt{term (c)} \ge \sum_{t=2}^{T} \frac{1}{4\eta_{t-1}} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \quad \left(\eta_t = \min\left\{\tfrac{1}{4L}, \tfrac{D}{\sqrt{1+\widetilde{V}_{t-1}}}\right\}\right)$$

$$\Rightarrow \quad \mathrm{REG}_T = \texttt{term (a)} + \texttt{term (b)} - \texttt{term (c)}$$

$$\le 5D\sqrt{1+V_T} + (4D+1)G^2 + 2LD = \mathcal{O}\left(\sqrt{1+V_T}\right). \qquad \square$$

Problem-dependent Bounds

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta \langle M_t, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2$$

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

Different priors are imposed by designing suitable M_t for specific environments.

	Assumption(s)	**Setting of Optimism**	**Setting of η_t**	**Problem-dependent Regret Bound**
Small-loss Bound	L-Smooth + Non-negative	$M_t = \mathbf{0}$	$\approx \frac{D}{\sqrt{1+G_t}}$	$\mathcal{O}\left(\sqrt{1+F_T}\right)$
Variance Bound	–	$M_t = \widetilde{\boldsymbol{\mu}}_{t-1}$	$\approx \frac{D}{\sqrt{1+\mathrm{Var}_{t-1}}}$	$\widetilde{\mathcal{O}}\left(\sqrt{1+\mathrm{Var}_T}\right)$
Variation Bound	L-Smooth	$M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$	$\approx \frac{D}{\sqrt{1+\widetilde{V}_{t-1}}}$	$\mathcal{O}\left(\sqrt{1+V_T}\right)$

Connections: GV Algorithm and variance

By using algorithm for gradient-variation bound (OMD with $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$):

$$\underbrace{\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2}_{(\approx V_T)} \le 3\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2 \quad (\le 3\,\mathrm{Var}_T)$$

$$+ 3\sum_{t=1}^{T} \|\nabla f_{t-1}(\mathbf{x}_{t-1}) - \boldsymbol{\mu}_{t-1}\|_2^2 \quad (\le 3\,\mathrm{Var}_T)$$

$$+ 3\sum_{t=1}^{T} \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\|_2^2 \quad \left(\boldsymbol{\mu}_t = \tfrac{(t-1)\boldsymbol{\mu}_{t-1} + \nabla f_t(\mathbf{x}_t)}{t}\right)$$

$$(\|\boldsymbol{\mu}_t\|_2 \le G, \forall t \in [T])$$

$$\left(\le 3 \cdot \tfrac{\pi^2}{6}\right)$$

⇨ Optimistic OMD with last-round gradient as optimism (enjoying V_T-bound) can also attain gradient-variance bound (scaling with Var_T)

Connections: GV Algorithm and small-loss

By using algorithm for gradient-variation bound (OMD with $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$):

$$\underbrace{\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2}_{(\approx V_T)} \leq 2\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_2^2 + 2\sum_{t=2}^{T} \|\nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2 \qquad ((a+b)^2 \leq 2(a^2+b^2))$$

$$\leq 4L\sum_{t=1}^{T} f_t(\mathbf{x}_t) + 4L\sum_{t=2}^{T} f_{t-1}(\mathbf{x}_{t-1}) \qquad \text{(self-bounding property)}$$

$$\leq 8LF_T^X \qquad (F_T^X \triangleq \textstyle\sum_{t=1}^{T} f_t(\mathbf{x}_t))$$

further use converting trick to attain F_T bound

⇨ Optimistic OMD with last-round gradient as optimism (enjoying V_T-bound) can also attain small-loss bound (scaling with F_T)

History Bits: Gradient-Variation Bounds

JMLR: Workshop and Conference Proceedings vol 23 (2012) 6.1–6.20 25th Annual Conference on Learning Theory

Online Optimization with Gradual Variations

Chao-Kai Chiang[1,2] CHAOKAI@IIS.SINICA.EDU.TW
Tianbao Yang[3] YANGTIA1@MSU.EDU
Chia-Jung Lee[1] LEECJ@IIS.SINICA.EDU.TW
Mehrdad Mahdavi[3] MAHDAVIM@CSE.MSU.EDU
Chi-Jen Lu[1] CJLU@IIS.SINICA.EDU.TW
Rong Jin[3] RONGJIN@CSE.MSU.EDU
Shenghuo Zhu[4] ZSH@SV.NEC-LABS.COM
[1] *Institute of Information Science, Academia Sinica, Taipei, Taiwan.*
[2] *Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.*
[3] *Department of Computer Science and Engineering Michigan State University, East Lansing, MI, 48824, USA*
[4] *NEC Laboratories America Cupertino, CA, 95014, USA*

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We study the online convex optimization problem, in which an online algorithm has to make repeated decisions with convex loss functions and hopes to achieve a small regret. We consider a natural restriction of this problem in which the loss functions have a small deviation, measured by the sum of the distances between every two consecutive loss functions, according to some distance metrics. We show that for the linear and general smooth convex loss functions, an online algorithm modified from the gradient descend algorithm can achieve a regret which only scales as the square root of the deviation. For the closely related problem of prediction with expert advice, we show that an online algorithm modified from the multiplicative update algorithm can also achieve a similar regret bound for a different measure of deviation. Finally, for loss functions which are strictly convex, we show that an online algorithm modified from the online Newton step algorithm can achieve a regret which is only logarithmic in terms of the deviation, and as an application, we can also have such a logarithmic regret for the portfolio management problem.

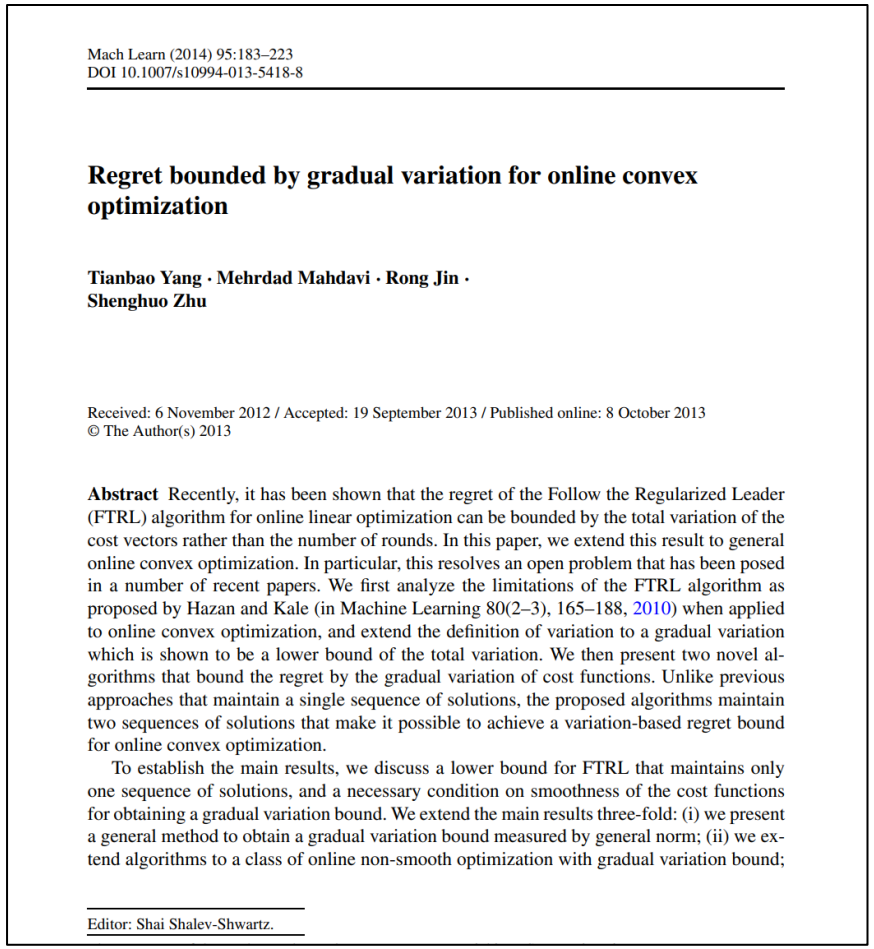
Keywords: Online Learning, Regret, Convex Optimization, Deviation.

1. Introduction

We study the online convex optimization problem in which a player has to make decisions iteratively for a number of rounds in the following way. In round t, the player has to choose a point x_t from some convex feasible set $\mathcal{X} \subseteq \mathbb{R}^N$, and after that the player receives a convex loss function f_t and suffers the corresponding loss $f_t(x_t) \in [0, 1]$. The player would like to have an online algorithm that can minimize its regret, which is the difference between the total loss it suffers and that of the best fixed point in hindsight. It is known

Chiang et al., Online Optimization with Gradual Variations. COLT 2012.

COLT 2012
best student paper award



Mach Learn (2014) 95:183–223
DOI 10.1007/s10994-013-5418-8

Regret bounded by gradual variation for online convex optimization

Tianbao Yang · Mehrdad Mahdavi · Rong Jin · Shenghuo Zhu

Received: 6 November 2012 / Accepted: 19 September 2013 / Published online: 8 October 2013

Abstract Recently, it has been shown that the regret of the Follow the Regularized Leader (FTRL) algorithm for online linear optimization can be bounded by the total variation of the cost vectors rather than the number of rounds. In this paper, we extend this result to general online convex optimization. In particular, this resolves an open problem that has been posed in a number of recent papers. We first analyze the limitations of the FTRL algorithm as proposed by Hazan and Kale (in Machine Learning 80(2–3), 165–188, 2010) when applied to online convex optimization, and extend the definition of variation to a gradual variation which is shown to be a lower bound of the total variation. We then present two novel algorithms that bound the regret by the gradual variation of cost functions. Unlike previous approaches that maintain a single sequence of solutions, the proposed algorithms maintain two sequences of solutions that make it possible to achieve a variation-based regret bound for online convex optimization.

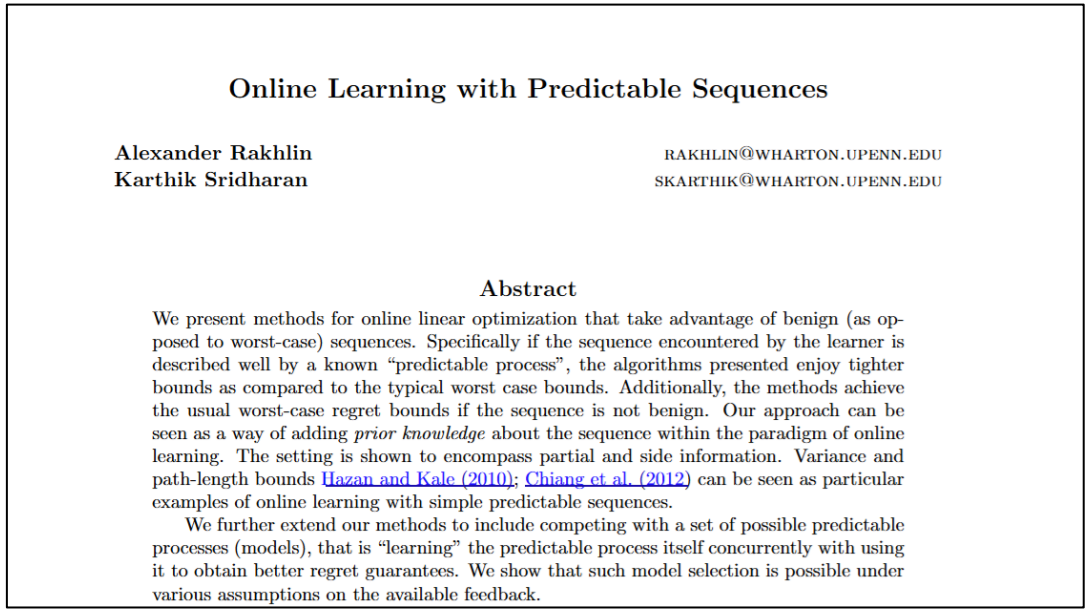
To establish the main results, we discuss a lower bound for FTRL that maintains only one sequence of solutions, and a necessary condition on smoothness of the cost functions for obtaining a gradual variation bound. We extend the main results three-fold: (i) we present a general method to obtain a gradual variation bound measured by general norm; (ii) we extend algorithms to a class of online non-smooth optimization with gradual variation bound;

Editor: Shai Shalev-Shwartz.

Yang et al., Regret bounded by gradual variation for online convex optimization. Machine Learning, 2014.

History Bits: Optimistic OMD

Optimistic OMD



Online Learning with Predictable Sequences

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
Karthik Sridharan SKARTHIK@WHARTON.UPENN.EDU

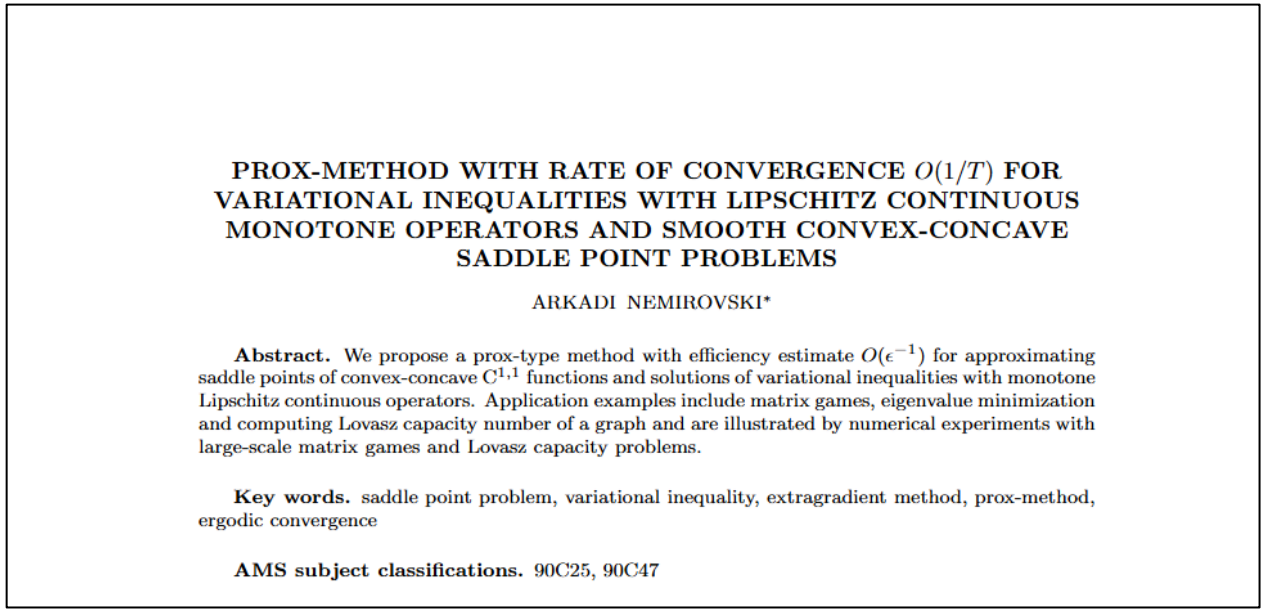
Abstract

We present methods for online linear optimization that take advantage of benign (as opposed to worst-case) sequences. Specifically if the sequence encountered by the learner is described well by a known "predictable process", the algorithms presented enjoy tighter bounds as compared to the typical worst case bounds. Additionally, the methods achieve the usual worst-case regret bounds if the sequence is not benign. Our approach can be seen as a way of adding *prior knowledge* about the sequence within the paradigm of online learning. The setting is shown to encompass partial and side information. Variance and path-length bounds Hazan and Kale (2010); Chiang et al. (2012) can be seen as particular examples of online learning with simple predictable sequences.

We further extend our methods to include competing with a set of possible predictable processes (models), that is "learning" the predictable process itself concurrently with using it to obtain better regret guarantees. We show that such model selection is possible under various assumptions on the available feedback.

Rakhlin & Sridharan, Online Learning with Predictable Sequences, COLT 2013.

Mirror Prox



PROX-METHOD WITH RATE OF CONVERGENCE $O(1/T)$ FOR VARIATIONAL INEQUALITIES WITH LIPSCHITZ CONTINUOUS MONOTONE OPERATORS AND SMOOTH CONVEX-CONCAVE SADDLE POINT PROBLEMS

ARKADI NEMIROVSKI*

Abstract. We propose a prox-type method with efficiency estimate $O(\epsilon^{-1})$ for approximating saddle points of convex-concave $C^{1,1}$ functions and solutions of variational inequalities with monotone Lipschitz continuous operators. Application examples include matrix games, eigenvalue minimization and computing Lovasz capacity number of a graph and are illustrated by numerical experiments with large-scale matrix games and Lovasz capacity problems.

Key words. saddle point problem, variational inequality, extragradient method, prox-method, ergodic convergence

AMS subject classifications. 90C25, 90C47

Nemirovski. Prox-Method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems. SIAM Journal on OPT., 2004.

Part 3. Implications to Offline Optimization

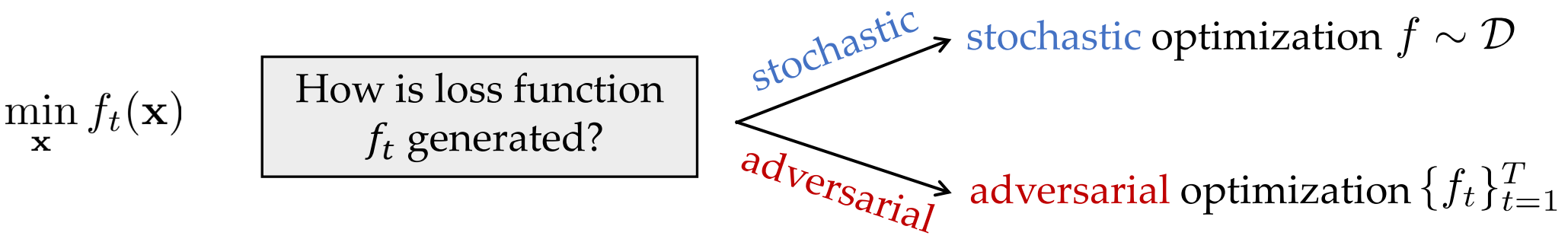
- Adaptive Optimization
- Smooth Optimization
- Accelerated Optimization

Part 3. Implications to Offline Optimization

- Adaptive Optimization
- Smooth Optimization
- Accelerated Optimization

Application to Adaptive Optimization

- **SEA (Stochastically Extended Adversarial) model**



[Sachs et al., NeurIPS'22]

Setup: at round $t \in [T]$, SEA optimizes $\min_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$

f_t is the *randomized function* sampled from underlying distribution $\mathcal{D}_t$: $f_t \sim \mathcal{D}_t$

F_t is the *expected function* of f_t: $F_t(\cdot) \triangleq \mathbb{E}_{f_t\sim\mathcal{D}_t}[f_t(\cdot)]$

Application to Adaptive Optimization

$$\mathbb{E}[\mathrm{REG}_T] \triangleq \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right]$$

- randomized function: $f_t \sim \mathcal{D}_t$
- expected function $F_t(\cdot) \triangleq \mathbb{E}_{f_t\sim\mathcal{D}_t}[f_t(\cdot)]$

complexity measures $\sigma_{1:T}^2 \triangleq \sum_{t=1}^{T} \max_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{f_t\sim\mathcal{D}_t}[\|\nabla f_t(\mathbf{x}) - \nabla F_t(\mathbf{x})\|^2]$, $\Sigma_{1:T}^2 \triangleq \mathbb{E}\left[\sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla F_t(\mathbf{x}) - \nabla F_{t-1}(\mathbf{x})\|^2\right]$

stochastic variance *adversarial change*

SEA can be solved by gradient-variation algorithms (*not implicit update*) over $\{f_t\}_{t=1}^T$.

$$\underbrace{\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})}_{\textit{gradient variation}} = [\nabla f_t(\mathbf{x}) - \nabla F_t(\mathbf{x})] + [\nabla F_t(\mathbf{x}) - \nabla F_{t-1}(\mathbf{x})] + [\nabla F_{t-1}(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})]$$

stochastic variance *adversarial change* *stochastic variance*

⇨ Approximately $V_T \approx \sigma_{1:T}^2 + \Sigma_{1:T}^2$.

For stochastic optimization, $\sigma_{1:T}^2 = \sigma^2 T$ and $\Sigma_{1:T}^2 = 0$.

For adversarial optimization, $\sigma_{1:T}^2 = 0$ and $\Sigma_{1:T}^2 = V_T$.

Optimistic OMD for the SEA model

Below, we focus on the optimization over bound the expected regret in terms of the linearized function, i.e., $\sum_{t=1}^{T}\langle\nabla f_t(\mathbf{x}_t),\mathbf{x}_t-\mathbf{u}\rangle$. For convex and smooth functions, we configure the algorithm with Euclidean regularizer

$$\psi_t(\mathbf{x})=\frac{1}{2\eta_t}\|\mathbf{x}\|_2^2 \quad \text{and step size} \quad \eta_t=\frac{D}{\sqrt{\delta+4G^2+\bar{V}_{t-1}}}, \tag{10}$$

where $\bar{V}_{t-1}=\sum_{s=1}^{t-1}\|\nabla f_s(\mathbf{x}_s)-\nabla f_{s-1}(\mathbf{x}_{s-1})\|_2^2$ (assuming $\nabla f_0(\mathbf{x}_0)=0$) and $\delta>0$ is a parameter to be specified later. Then, the optimistic OMD updates in (7) and (8) become

$$\widehat{\mathbf{x}}_{t+1}=\Pi_{\mathcal{X}}\left[\widehat{\mathbf{x}}_t-\eta_t\nabla f_t(\mathbf{x}_t)\right], \quad \mathbf{x}_{t+1}=\Pi_{\mathcal{X}}\left[\widehat{\mathbf{x}}_{t+1}-\eta_{t+1}\nabla f_t(\mathbf{x}_t)\right], \tag{11}$$

Theorem 1. *Under Assumptions 1, 2, 4 and 5, optimistic OMD with regularizer (10) and updates (11) enjoys the following guarantee:*

$$\mathbb{E}[\mathbf{Reg}_T(\mathbf{u})]\le 5\sqrt{10}D^2L+\frac{5\sqrt{5}DG}{2}+5\sqrt{2}D\sqrt{\sigma_{1:T}^2}+5D\sqrt{\Sigma_{1:T}^2}=\mathcal{O}\left(\sqrt{\sigma_{1:T}^2}+\sqrt{\Sigma_{1:T}^2}\right),$$

where we set $\delta=10D^2L^2$ in (10).

Reference: Sijia Chen, Yu-Jie Zhang, Wei-Wei Tu, Peng Zhao, and Lijun Zhang. Optimistic Online Mirror Descent for Bridging Stochastic and Adversarial Online Convex Optimization. Journal of Machine Learning Research (JMLR), 25(178):1–62, 2024.

Part 3. Implications to Offline Optimization

- Adaptive Optimization
- Smooth Optimization
- Accelerated Optimization

Application to Offline Optimization

- Online algorithm with ***problem-independent*** bound:

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

- For an offline optimization problem $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$

  When the function is convex and ***Lipschitz***, we can use ***problem-independent*** regret with online-to-batch (O2B) conversion to obtain an averaged model with

$$\varepsilon_T \triangleq F\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right) - \min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}) \leq \mathcal{O}\left(\frac{\sqrt{T}}{T}\right) = \mathcal{O}\left(\sqrt{\frac{1}{T}}\right).$$

Gradient-Variation Bound Reflection

Definition 3 (Gradient Variation). Let T be the time horizon and $\mathcal{X} \subseteq \mathbb{R}^d$ be the feasible domain. For the function sequence $f_1, \ldots, f_T$ with $f_t : \mathcal{X} \mapsto \mathbb{R}$ for $t \in [T]$, its gradient variation is defined as

$$V_T = \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|_2^2$$

- This gradient-variation notion tightly connects the ***offline optimization*** and ***online optimization***.
- The gradient variation reveals the importance of ***smoothness*** for the first-order methods, as well as the crucial role of the ***negative term*** in analysis.

Application to Offline Optimization

- Online algorithm with ***gradient-variation*** regret bound:

$$\textsc{Reg}_T \triangleq \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \mathcal{O}\left(\sqrt{1+V_T}\right).$$

- For an offline optimization problem $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$

  When the function is convex and ***smooth***, we can use ***gradient-variation*** regret with online-to-batch (O2B) conversion to obtain an averaged model with

$$\varepsilon_T \triangleq F\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right) - \min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}) \leq \mathcal{O}\left(\frac{\sqrt{1+V_T(F,\ldots,F)}}{T}\right) = \mathcal{O}\left(\frac{1}{T}\right).$$

Part 3. Implications to Offline Optimization

- Adaptive Optimization
- Smooth Optimization
- Accelerated Optimization

Accelerated Methods

- Recall that ***accelerated*** rates can be achieved for smooth convex optimization using Nesterov's Accelerated GD.

Theorem 3. *Let f be convex and L-smooth. Nesterov's accelerated GD is configured as*

$$\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t), \quad \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t),$$

where $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$, *and* $\beta_t = \frac{\lambda_t - 1}{\lambda_{t+1}}$. *Then, we have*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right).$$

In our previous lecture, we prove this accelerated rate by the generalized one-step improvement property and a variety of algebraic tricks.

Acceleration by Optimistic OMD

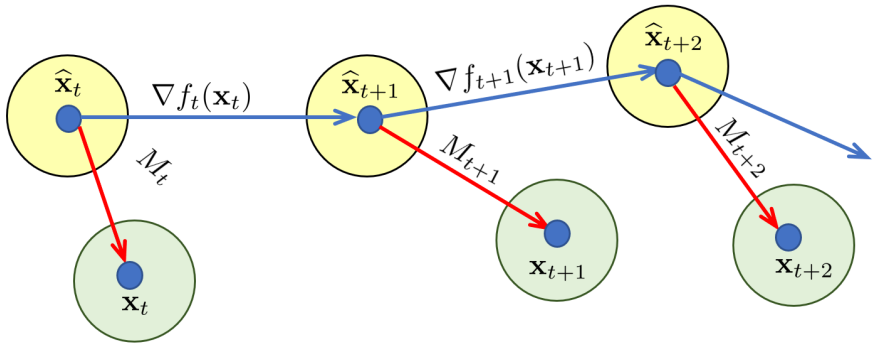
- We now present ***a new algorithm based on optimistic OMD*** with an accelerated rate for smooth convex optimization.

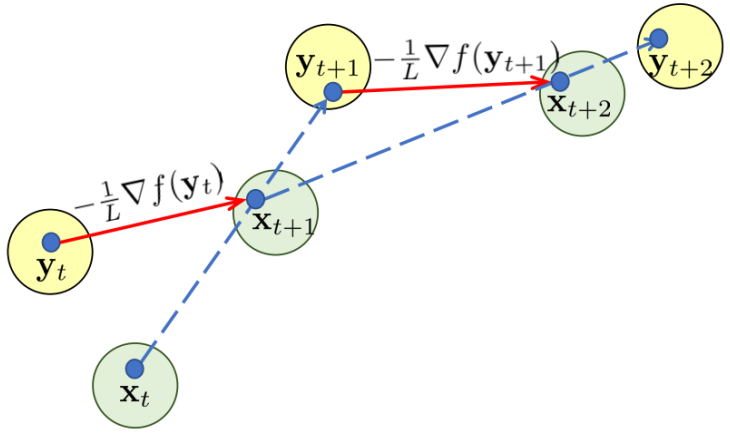
$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle M_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_t) \right\}$$

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

$$\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t)$$

$$\mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t)$$





Acceleration by Optimistic OMD

There are two key components:

- **Stabilized Online-to-Batch Conversion**

  This is used to reduce the offline optimization to online optimization, but now we need to carefully choose gradient evaluations to enhance the stability.

- **Optimism Design**

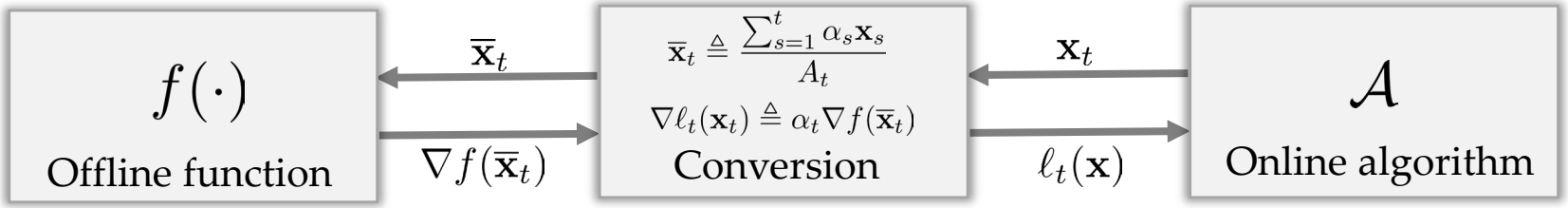
  This is used to achieve the desired vanishing regret in online optimization, in which the optimism design is crucial. It is essential to leverage the special structure of the problem.

Stabilized Online-to-Batch Conversion

- Reducing ***offline optimization*** as an ***online optimization***.

Algorithm 1 Stabilized Online-to-Batch Conversion Template

1: Online algorithm $\mathcal{A}_{\mathsf{OL}}$, weights $\{\alpha_t\}_{t=1}^T$ with $\alpha_t > 0$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Submit $\overline{\mathbf{x}}_t = \frac{\sum_{s=1}^t \alpha_s \mathbf{x}_s}{A_t}$ with $A_t \triangleq \sum_{s=1}^t \alpha_s$
4: Receive $\nabla f(\overline{\mathbf{x}}_t)$
5: Define $\ell_t(\mathbf{x}) = \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x} \rangle$ as t-th round online function to $\mathcal{A}_{\mathsf{OL}}$
6: Obtain $\mathbf{x}_{t+1}$ from $\mathcal{A}_{\mathsf{OL}}\left(\mathbf{x}_1, \{\ell_s(\cdot)\}_{s=1}^t\right)$
7: **end for**



Stabilized Online-to-Batch Conversion

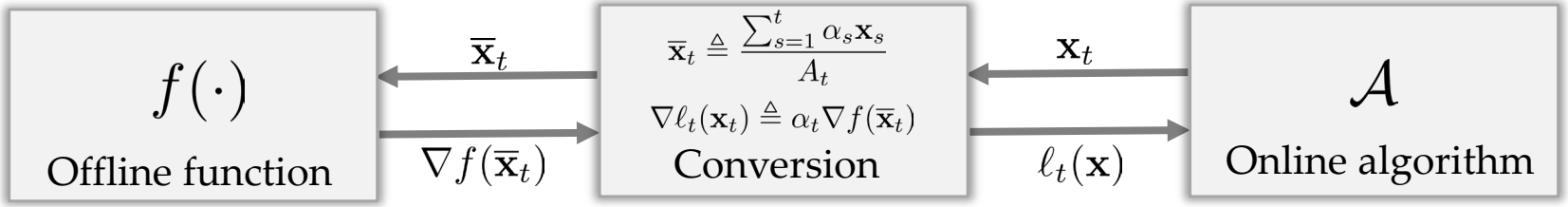
- Reducing ***offline optimization*** as an ***online optimization***.

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated):*

$$\overline{\mathbf{x}}_t = \frac{\sum_{s=1}^t \alpha_s \mathbf{x}_s}{A_t},$$

with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \leq \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$



Stabilized Online-to-Batch Conversion

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$

- We can set α_t larger to make the denominator larger, such that we may have a chance to achieve a faster rate than the standard $\mathcal{O}\big(\frac{1}{\sqrt{T}}\big)$ one.
- But when α_t is large, the regret of online learner is also hard to control – due to the large gradient magnitute $\alpha_t \nabla f(\bar{\mathbf{x}}_t)$

Stabilized Online-to-Batch Conversion

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$

Proof: First, by convexity we have

$$\begin{aligned}\sum_{t=1}^T \alpha_t (f(\overline{\mathbf{x}}_t) - f(\mathbf{x}^\star)) &\le \sum_{t=1}^T \alpha_t \langle \nabla f(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}^\star\rangle \\ &= \underbrace{\sum_{t=1}^T \alpha_t \langle \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle}_{\triangleq \mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)} + \sum_{t=1}^T \alpha_t \langle \nabla f(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t\rangle\end{aligned}$$

Stabilized Online-to-Batch Conversion

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$

Proof: First, by convexity we have

$$\sum_{t=1}^T \alpha_t (f(\overline{\mathbf{x}}_t) - f(\mathbf{x}^\star)) \le \mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star) + \sum_{t=1}^T \alpha_t \langle \nabla f(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t\rangle$$

Notice the following two facts

$$\begin{aligned} &\textstyle\sum_{s=1}^t \alpha_s \mathbf{x}_s = A_t \overline{\mathbf{x}}_t = A_{t-1}\overline{\mathbf{x}}_t + \alpha_t \overline{\mathbf{x}}_t \\ &\textstyle\sum_{s=1}^t \alpha_s \mathbf{x}_s = \sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_t = A_{t-1}\overline{\mathbf{x}}_{t-1} + \alpha_t \mathbf{x}_t \end{aligned} \quad \Longrightarrow \quad \alpha_t(\overline{\mathbf{x}}_t - \mathbf{x}_t) = A_{t-1}(\overline{\mathbf{x}}_{t-1} - \overline{\mathbf{x}}_t)$$

Stabilized Online-to-Batch Conversion

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$

Proof: Further using the convexity property, we get

$$\begin{aligned}\sum_{t=1}^T \alpha_t (f(\overline{\mathbf{x}}_t) - f(\mathbf{x}^\star)) &\le \mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star) - \sum_{t=1}^T A_{t-1}\langle \nabla f(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\rangle \\ &\le \mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star) - \sum_{t=1}^T A_{t-1}(f(\overline{\mathbf{x}}_t) - f(\overline{\mathbf{x}}_{t-1}))\end{aligned}$$

This implies that $A_T(f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star)) \le \mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)$ □

Stabilized Online-to-Batch Conversion

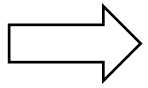
Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$

Set weights $\alpha_t = t$ for all $t \in [T]$, then $A_T = \mathcal{O}(T^2)$.

We aim to use online algorithm ensuring $\mathcal{O}(1)$ regret.

This is kind of "crazy", as gradient magnitute $\alpha_t \nabla f(\bar{\mathbf{x}}_t)$ can be very large.



Optimistic OMD with a suitable optimism design!

Theorem 3. *Let f be convex and L-smooth. Nesterov's accelerated GD is configured as*

$$\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t), \quad \mathbf{y}_{t+1} = \mathbf{x}_{t+1} + \beta_t(\mathbf{x}_{t+1} - \mathbf{x}_t),$$

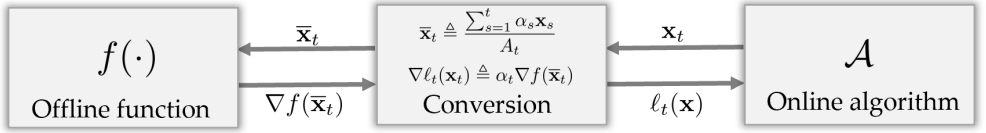
where $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$, and $\beta_t = \frac{\lambda_t - 1}{\lambda_{t+1}}$. Then, we have

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{2L\|\mathbf{x}_1 - \mathbf{x}^\star\|^2}{T^2} = \mathcal{O}\left(\frac{1}{T^2}\right).$$

Accelerated Rates by Optimistic OMD

$$\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^{t} \alpha_s \mathbf{x}_s$$

- Can we achieve an $\mathcal{O}(1)$ regret for stabilized online-to-batch conversion?



Yes! We can use the Optimistic Online Mirror Descent.

- Recall in gradient-variation regret, the negative term is crucial.

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta \langle M_t, \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2$$

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

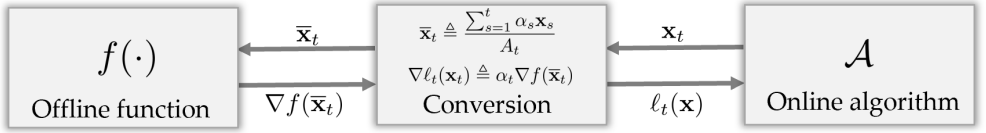
$$\Longrightarrow \sum_{t=1}^{T} \ell_t(\mathbf{x}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}) \leq \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \|\nabla \ell_t(\mathbf{x}_t) - M_t\|_2^2 - \frac{1}{4\eta}\sum_{t=1}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$$

(negative term)

Accelerated Rates by Optimistic OMD

$$\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^{t} \alpha_s \mathbf{x}_s$$

- Can we achieve an $\mathcal{O}(1)$ regret for stabilized online-to-batch conversion?



Yes! We can use the Optimistic Online Mirror Descent of the last lecture.

$\nabla \ell_t(\mathbf{x}_t) = \alpha_t \nabla f(\overline{\mathbf{x}}_t)$, $M_t = \alpha_t \nabla f(\widetilde{\mathbf{x}}_t)$, with $\widetilde{\mathbf{x}}_t$ to be determined:

$$\Longrightarrow \sum_{t=1}^{T} \ell_t(\mathbf{x}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}) \le \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \underbrace{\|\alpha_t \nabla f(\overline{\mathbf{x}}_t) - \alpha_t \nabla f(\widetilde{\mathbf{x}}_t)\|_2^2}_{\text{(L-smoothness)}} - \frac{1}{4\eta}\sum_{t=1}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$$

$$\le \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \alpha_t^2 L^2 \|\overline{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 - \frac{1}{4\eta}\sum_{t=1}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$$

Optimism Design

$$\sum_{t=1}^{T} \ell_t(\mathbf{x}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}) \le \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \alpha_t^2 L^2 \left\| \overline{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t \right\|_2^2 - \frac{1}{4\eta} \sum_{t=1}^{T} \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|_2^2$$

- Optimism design: approximate $\bar{\mathbf{x}}_t$ as possible as we can

by def $\bar{\mathbf{x}}_t = \frac{1}{A_t}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_t\right)$,
we set $\widetilde{\mathbf{x}}_t \triangleq \frac{1}{A_t}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1}\right)$

$\Rightarrow$ $\overline{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{\alpha_t}{A_t}(\mathbf{x}_t - \mathbf{x}_{t-1})$

$$\sum_{t=1}^{T} \ell_t(\mathbf{x}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}) \le \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \alpha_t^2 \frac{\alpha_t^2 L^2}{A_t^2} \left\| \mathbf{x}_t - \mathbf{x}_{t-1} \right\|_2^2 - \frac{1}{4\eta} \sum_{t=1}^{T} \left\| \mathbf{x}_{t+1} - \mathbf{x}_t \right\|_2^2$$

when $\alpha_t = t$, it becomes $\frac{\alpha_t^4}{A_t^2} = \Theta(1) \Rightarrow \eta \le \frac{1}{4L}$ ensures cancellation

$\Rightarrow$ Therefore, by setting $\eta = \frac{1}{4L}$, we have $\mathrm{Reg}_T^{\mathcal{A}} \le 2D^2 L = \mathcal{O}(1)$.

Stabilization Effect

- The crucial role of stabilization effect

Algorithm 1 Stabilized Online-to-Batch Conversion Template

1: Online algorithm $\mathcal{A}_{\mathsf{OL}}$, weights $\{\alpha_t\}_{t=1}^T$ with $\alpha_t > 0$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Submit $\bar{\mathbf{x}}_t = \frac{\sum_{s=1}^t \alpha_s \mathbf{x}_s}{A_t}$ with $A_t \triangleq \sum_{s=1}^t \alpha_s$
4: Receive $\nabla f(\bar{\mathbf{x}}_t)$
5: Define $\ell_t(\mathbf{x}) = \langle \alpha_t \nabla f(\bar{\mathbf{x}}_t), \mathbf{x} \rangle$ as t-th round online function to $\mathcal{A}_{\mathsf{OL}}$
6: Obtain $\mathbf{x}_{t+1}$ from $\mathcal{A}_{\mathsf{OL}}$ $(\mathbf{x}_1, \{\ell_s(\cdot)\}_{s=1}^t)$
7: **end for**

$$\text{by def } \bar{\mathbf{x}}_t = \frac{1}{A_t}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_t\right),$$
$$\text{we set } \widetilde{\mathbf{x}}_t \triangleq \frac{1}{A_t}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1}\right)$$

$$\Rightarrow \quad \bar{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{\alpha_t}{A_t}\left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)$$

$$\|\bar{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2 = \left(\frac{\alpha_t}{A_t}\right)^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 = \Theta\left(\frac{1}{t^2}\right) \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2$$

This indicates that the weighted average is $\Theta(1/t^2)$ more stable than the original sequence, enabling the negative term cancellation, which finally leads to a constant weighted regret.

Accelerated Rates by Optimistic OMD

- Combining the stabilized online-to-batch conversion and a careful optimism design (for constant regret), we achieve the acceleration.

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \leq \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\mathcal{A}}(\mathbf{x}^\star)}{A_T}.$$

$\Longrightarrow$ $\mathrm{Reg}_T^{\mathcal{A}} = \mathcal{O}(1), A_T^{-1} = \mathcal{O}(T^{-2})$, which leads to an $\mathcal{O}(T^{-2})$ convergence rate!

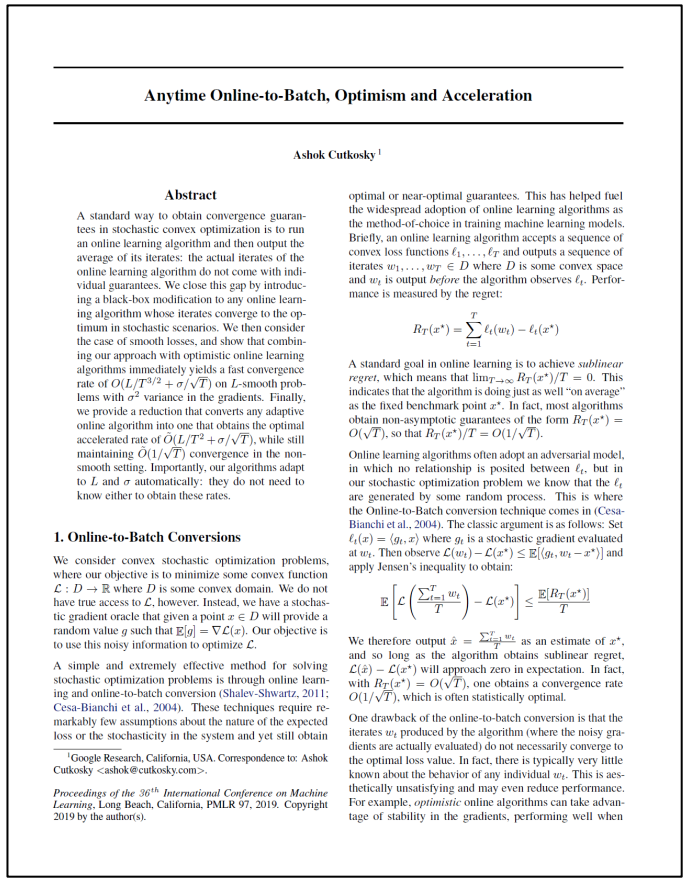
Accelerated Rates by Optimistic OMD

- Combining the stabilized online-to-batch conversion and a careful optimism design (for constant regret), we achieve the acceleration.

Algorithm 2 Simple Accelerated Method based on Optimistic OMD

Initialization: Set $\alpha_t = t$, $A_t = \sum_{s=1}^{t} \alpha_s$, $\eta = \frac{1}{4L}$.
for $t = 1, 2, \ldots, T$ **do**
  Submit $\widetilde{\mathbf{x}}_t \triangleq \frac{1}{A_t} \sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1}$
  Receive $\nabla f(\widetilde{\mathbf{x}}_t)$, set $M_t = \alpha_t \nabla f(\widetilde{\mathbf{x}}_t)$
  Update $\mathbf{x}_t = \arg\min_{\mathbf{x} \in \mathcal{X}} \eta \langle M_t, \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2$
  Submit $\overline{\mathbf{x}}_t = \frac{1}{A_t} \sum_{s=1}^{t} \alpha_s \mathbf{x}_s$
  Receive $\nabla f(\overline{\mathbf{x}}_t)$, set $\ell_t(\mathbf{x}) = \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x} \rangle$
  Update $\widehat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \eta \langle \nabla \ell_t(\mathbf{x}_t), \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x} - \widehat{\mathbf{x}}_t\|_2^2$
end for

History bits: Optimism for Acceleration



Anytime Online-to-Batch, Optimism and Acceleration

Ashok Cutkosky [1]

Abstract

A standard way to obtain convergence guarantees in stochastic convex optimization is to run an online learning algorithm and then output the average of its iterates: the actual iterates of the online learning algorithm do not come with individual guarantees. We close this gap by introducing a black-box modification to any online learning algorithm whose iterates converge to the optimum in stochastic scenarios. We then consider the case of smooth losses, and show that combining our approach with optimistic online learning algorithms immediately yields a fast convergence rate of $O(L/T^{3/2} + \sigma/\sqrt{T})$ on L-smooth problems with σ^2 variance in the gradients. Finally, we provide a reduction that converts any adaptive online algorithm into one that obtains the optimal accelerated rate of $\tilde{O}(L/T^2 + \sigma/\sqrt{T})$, while still maintaining $\tilde{O}(1/\sqrt{T})$ convergence in the non-smooth setting. Importantly, our algorithms adapt to L and σ automatically: they do not need to know either to obtain these rates.

1. Online-to-Batch Conversions

We consider convex stochastic optimization problems, where our objective is to minimize some convex function $\mathcal{L} : D \to \mathbb{R}$ where D is some convex domain. We do not have true access to $\mathcal{L}$, however. Instead, we have a stochastic gradient oracle that given a point $x \in D$ will provide a random value g such that $\mathbb{E}[g] = \nabla\mathcal{L}(x)$. Our objective is to use this noisy information to optimize $\mathcal{L}$.

A simple and extremely effective method for solving stochastic optimization problems is through online learning and online-to-batch conversion (Shalev-Shwartz, 2011; Cesa-Bianchi et al., 2004). These techniques require remarkably few assumptions about the nature of the expected loss or the stochasticity in the system and yet still obtain optimal or near-optimal guarantees. This has helped fuel the widespread adoption of online learning algorithms as the method-of-choice in training machine learning models. Briefly, an online learning algorithm accepts a sequence of convex loss functions $\ell_1, \dots, \ell_T$ and outputs a sequence of iterates $w_1, \dots, w_T \in D$ where D is some convex space and w_t is output *before* the algorithm observes ℓ_t. Performance is measured by the regret:

$$R_T(x^\star) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(x^\star)$$

A standard goal in online learning is to achieve *sublinear regret*, which means that $\lim_{T\to\infty} R_T(x^\star)/T = 0$. This indicates that the algorithm is doing just as well "on average" as the fixed benchmark point $x^\star$. In fact, most algorithms obtain non-asymptotic guarantees of the form $R_T(x^\star) = O(\sqrt{T})$, so that $R_T(x^\star)/T = O(1/\sqrt{T})$.

Online learning algorithms often adopt an adversarial model, in which no relationship is posited between ℓ_t, but in our stochastic optimization problem we know that the ℓ_t are generated by some random process. This is where the Online-to-Batch conversion technique comes in (Cesa-Bianchi et al., 2004). The classic argument is as follows: Set $\ell_t(x) = \langle g_t, x\rangle$ where g_t is a stochastic gradient evaluated at w_t. Then observe $\mathcal{L}(w_t) - \mathcal{L}(x^\star) \le \mathbb{E}[\langle g_t, w_t - x^\star\rangle]$ and apply Jensen's inequality to obtain:

$$\mathbb{E}\left[\mathcal{L}\left(\frac{\sum_{t=1}^{T} w_t}{T}\right) - \mathcal{L}(x^\star)\right] \le \frac{\mathbb{E}[R_T(x^\star)]}{T}$$

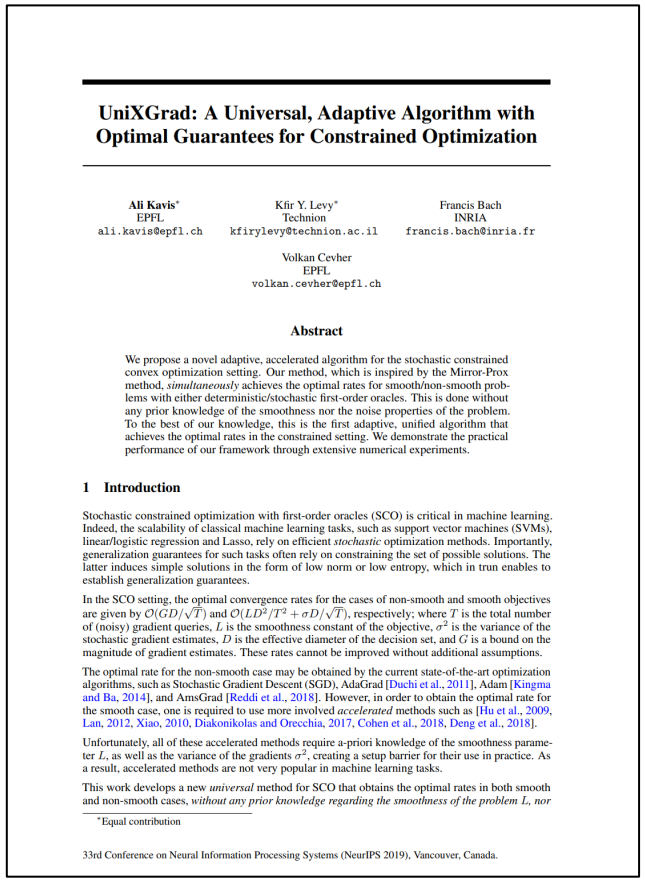
We therefore output $\hat{x} = \frac{\sum_{t=1}^{T} w_t}{T}$ as an estimate of $x^\star$, and so long as the algorithm obtains sublinear regret, $\mathcal{L}(\hat{x}) - \mathcal{L}(x^\star)$ will approach zero in expectation. In fact, with $R_T(x^\star) = O(\sqrt{T})$, one obtains a convergence rate $O(1/\sqrt{T})$, which is often statistically optimal.

One drawback of the online-to-batch conversion is that the iterates w_t produced by the algorithm (where the noisy gradients are actually evaluated) do not necessarily converge to the optimal loss value. In fact, there is typically very little known about the behavior of any individual w_t. This is aesthetically unsatisfying and may even reduce performance. For example, *optimistic* online algorithms can take advantage of stability in the gradients, performing well when

[1]Google Research, California, USA. Correspondence to: Ashok Cutkosky <ashok@cutkosky.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Anytime Online-to-Batch, Optimism and Acceleration. ICML 2019.



UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization

Ali Kavis*
EPFL
ali.kavis@epfl.ch

Kfir Y. Levy*
Technion
kfirylevy@technion.ac.il

Francis Bach
INRIA
francis.bach@inria.fr

Volkan Cevher
EPFL
volkan.cevher@epfl.ch

Abstract

We propose a novel adaptive, accelerated algorithm for the stochastic constrained convex optimization setting. Our method, which is inspired by the Mirror-Prox method, *simultaneously* achieves the optimal rates for smooth/non-smooth problems with either deterministic/stochastic first-order oracles. This is done without any prior knowledge of the smoothness nor the noise properties of the problem. To the best of our knowledge, this is the first adaptive, unified algorithm that achieves the optimal rates in the constrained setting. We demonstrate the practical performance of our framework through extensive numerical experiments.

1 Introduction

Stochastic constrained optimization with first-order oracles (SCO) is critical in machine learning. Indeed, the scalability of classical machine learning tasks, such as support vector machines (SVMs), linear/logistic regression and Lasso, rely on efficient *stochastic* optimization methods. Importantly, generalization guarantees for such tasks often rely on constraining the set of possible solutions. The latter induces simple solutions in the form of low norm or low entropy, which in trun enables to establish generalization guarantees.

In the SCO setting, the optimal convergence rates for the cases of non-smooth and smooth objectives are given by $\mathcal{O}(GD/\sqrt{T})$ and $\mathcal{O}(LD^2/T^2 + \sigma D/\sqrt{T})$, respectively; where T is the total number of (noisy) gradient queries, L is the smoothness constant of the objective, σ^2 is the variance of the stochastic gradient estimates, D is the effective diameter of the decision set, and G is a bound on the magnitude of gradient estimates. These rates cannot be improved without additional assumptions.

The optimal rate for the non-smooth case may be obtained by the current state-of-the-art optimization algorithms, such as Stochastic Gradient Descent (SGD), AdaGrad [Duchi et al., 2011], Adam [Kingma and Ba, 2014], and AmsGrad [Reddi et al., 2018]. However, in order to obtain the optimal rate for the smooth case, one is required to use more involved *accelerated* methods such as [Hu et al., 2009, Lan, 2012, Xiao, 2010, Diakonikolas and Orecchia, 2017, Cohen et al., 2018, Deng et al., 2018].

Unfortunately, all of these accelerated methods require a-priori knowledge of the smoothness parameter L, as well as the variance of the gradients σ^2, creating a setup barrier for their use in practice. As a result, accelerated methods are not very popular in machine learning tasks.

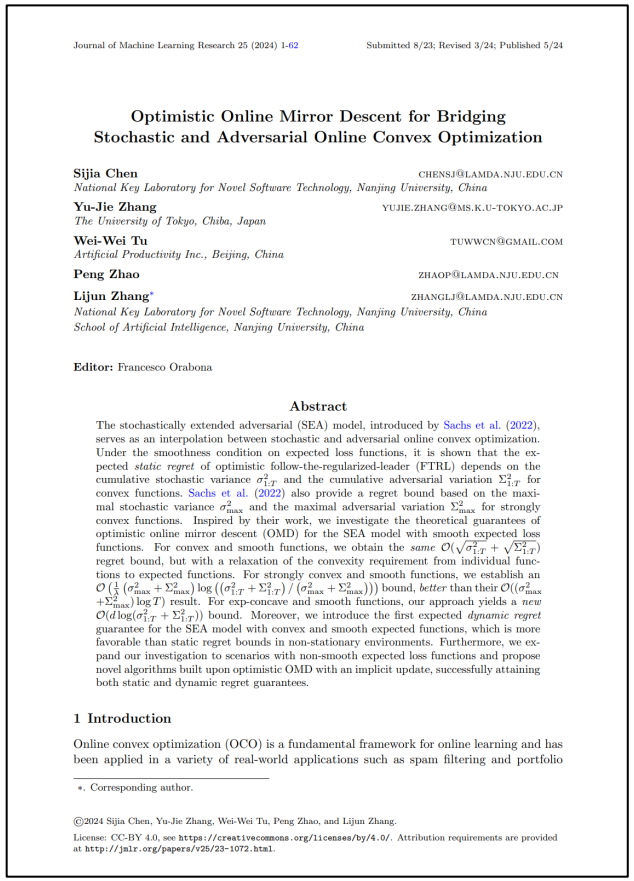
This work develops a new *universal* method for SCO that obtains the optimal rates in both smooth and non-smooth cases, *without any prior knowledge regarding the smoothness of the problem L, nor*

*Equal contribution

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization. NeurIPS 2019.

History bits: GV Regret for Optimization



Journal of Machine Learning Research 25 (2024) 1-62 Submitted 8/23; Revised 3/24; Published 5/24

Optimistic Online Mirror Descent for Bridging Stochastic and Adversarial Online Convex Optimization

Sijia Chen CHENSJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, China
Yu-Jie Zhang YUJIE.ZHANG@MS.K.U-TOKYO.AC.JP
The University of Tokyo, Chiba, Japan
Wei-Wei Tu TUWWCN@GMAIL.COM
Artificial Productivity Inc., Beijing, China
Peng Zhao ZHAOP@LAMDA.NJU.EDU.CN
Lijun Zhang* ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China

Editor: Francesco Orabona

Abstract

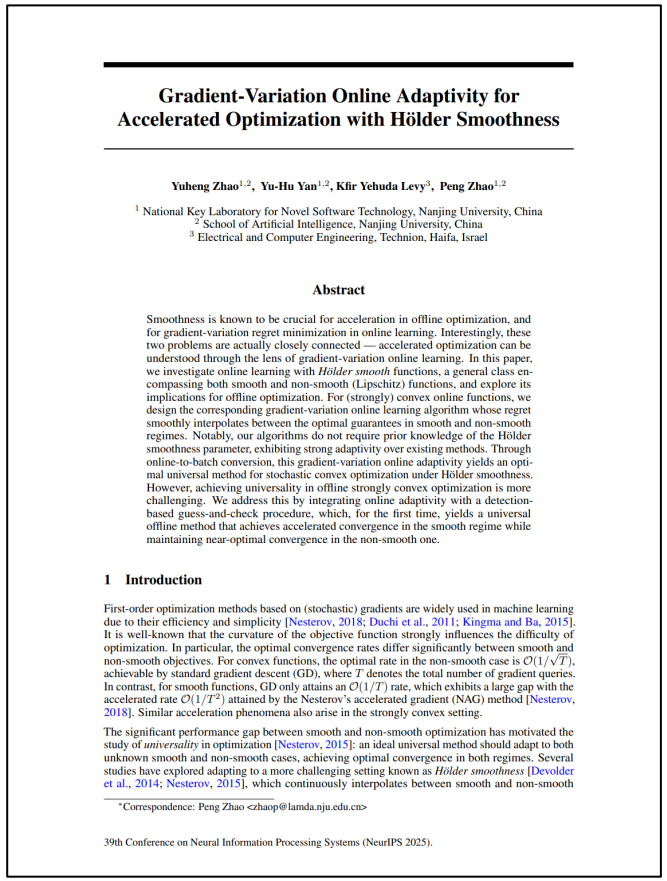
The stochastically extended adversarial (SEA) model, introduced by Sachs et al. (2022), serves as an interpolation between stochastic and adversarial online convex optimization. Under the smoothness condition on expected loss functions, it is shown that the expected *static regret* of optimistic follow-the-regularized-leader (FTRL) depends on the cumulative stochastic variance $\sigma_{1:T}^2$ and the cumulative adversarial variation $\Sigma_{1:T}^2$ for convex functions. Sachs et al. (2022) also provide a regret bound based on the maximal stochastic variance $\sigma_{\max}^2$ and the maximal adversarial variation $\Sigma_{\max}^2$ for strongly convex functions. Inspired by their work, we investigate the theoretical guarantees of optimistic online mirror descent (OMD) for the SEA model with smooth expected loss functions. For convex and smooth functions, we obtain the *same* $\mathcal{O}(\sqrt{\sigma_{1:T}^2 + \Sigma_{1:T}^2})$ regret bound, but with a relaxation of the convexity requirement from individual functions to expected functions. For strongly convex and smooth functions, we establish an $\mathcal{O}\left(\frac{1}{\lambda}\left(\sigma_{\max}^2 + \Sigma_{\max}^2\right)\log\left(\left(\sigma_{1:T}^2 + \Sigma_{1:T}^2\right)/\left(\sigma_{\max}^2 + \Sigma_{\max}^2\right)\right)\right)$ bound, *better* than their $\mathcal{O}((\sigma_{\max}^2 + \Sigma_{\max}^2)\log T)$ result. For exp-concave and smooth functions, our approach yields a *new* $\mathcal{O}(d\log(\sigma_{1:T}^2 + \Sigma_{1:T}^2))$ bound. Moreover, we introduce the first expected *dynamic regret* guarantee for the SEA model with convex and smooth expected functions, which is more favorable than static regret bounds in non-stationary environments. Furthermore, we expand our investigation to scenarios with non-smooth expected loss functions and propose novel algorithms built upon optimistic OMD with an implicit update, successfully attaining both static and dynamic regret guarantees.

1 Introduction

Online convex optimization (OCO) is a fundamental framework for online learning and has been applied in a variety of real-world applications such as spam filtering and portfolio

*. Corresponding author.

Optimistic Online Mirror Descent for Bridging Stochastic and Adversarial Online Convex Optimization. JMLR 2024.



Gradient-Variation Online Adaptivity for Accelerated Optimization with Hölder Smoothness

Yuheng Zhao[1,2], **Yu-Hu Yan**[1,2], **Kfir Yehuda Levy**[3], **Peng Zhao**[1,2]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] School of Artificial Intelligence, Nanjing University, China
[3] Electrical and Computer Engineering, Technion, Haifa, Israel

Abstract

Smoothness is known to be crucial for acceleration in offline optimization, and for gradient-variation regret minimization in online learning. Interestingly, these two problems are actually closely connected — accelerated optimization can be understood through the lens of gradient-variation online learning. In this paper, we investigate online learning with *Hölder smooth* functions, a general class encompassing both smooth and non-smooth (Lipschitz) functions, and explore its implications for offline optimization. For (strongly) convex online functions, we design the corresponding gradient-variation online learning algorithm whose regret smoothly interpolates between the optimal guarantees in smooth and non-smooth regimes. Notably, our algorithms do not require prior knowledge of the Hölder smoothness parameter, exhibiting strong adaptivity over existing methods. Through online-to-batch conversion, this gradient-variation online adaptivity yields an optimal universal method for stochastic convex optimization under Hölder smoothness. However, achieving universality in offline strongly convex optimization is more challenging. We address this by integrating online adaptivity with a detection-based guess-and-check procedure, which, for the first time, yields a universal offline method that achieves accelerated convergence in the smooth regime while maintaining near-optimal convergence in the non-smooth one.

1 Introduction

First-order optimization methods based on (stochastic) gradients are widely used in machine learning due to their efficiency and simplicity [Nesterov, 2018; Duchi et al., 2011; Kingma and Ba, 2015]. It is well-known that the curvature of the objective function strongly influences the difficulty of optimization. In particular, the optimal convergence rates differ significantly between smooth and non-smooth objectives. For convex functions, the optimal rate in the non-smooth case is $\mathcal{O}(1/\sqrt{T})$, achievable by standard gradient descent (GD), where T denotes the total number of gradient queries. In contrast, for smooth functions, GD only attains an $\mathcal{O}(1/T)$ rate, which exhibits a large gap with the accelerated rate $\mathcal{O}(1/T^2)$ attained by the Nesterov's accelerated gradient (NAG) method [Nesterov, 2018]. Similar acceleration phenomena also arise in the strongly convex setting.

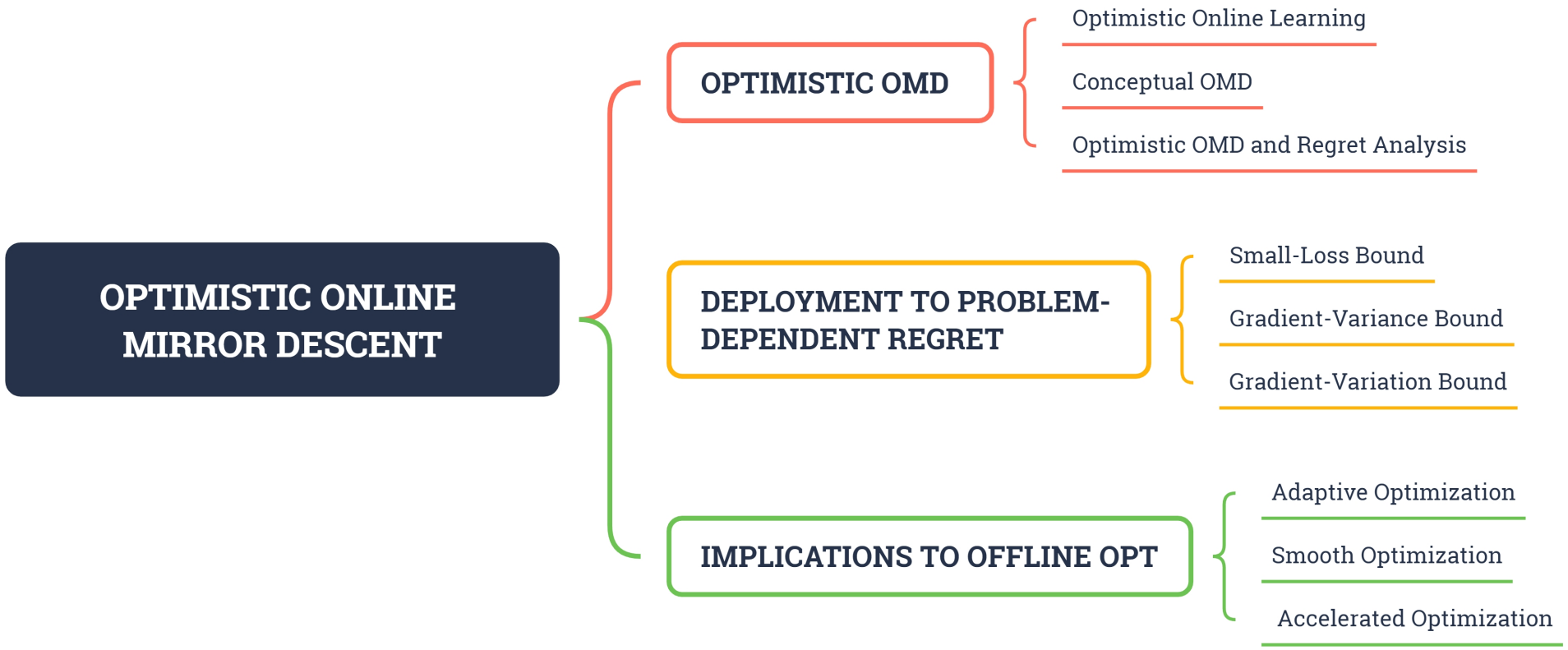
The significant performance gap between smooth and non-smooth optimization has motivated the study of *universality* in optimization [Nesterov, 2015]: an ideal universal method should adapt to both unknown smooth and non-smooth cases, achieving optimal convergence in both regimes. Several studies have explored adapting to a more challenging setting known as *Hölder smoothness* [Devolder et al., 2014; Nesterov, 2015], which continuously interpolates between smooth and non-smooth

*Correspondence: Peng Zhao <zhaop@lamda.nju.edu.cn>

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Gradient-Variation Online Adaptivity for Accelerated Optimization with Hölder Smoothness. NeurIPS 2025.

Summary



Q & A

Thanks!