Lecture 8. Optimistic Online Mirror Descent

Advanced Optimization (Fall 2024)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Optimistic OMD
- Applications
  - Small-Loss bound
  - Gradient-Variance bound
  - Gradient-Variation bound

Part 1. Optimistic OMD

- Optimistic Online Learning Setting

- Optimistic OMD Framework

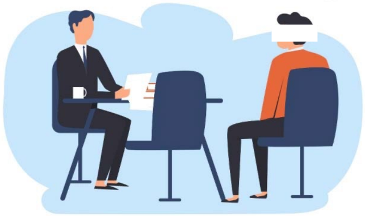
Beyond the Worst-Case Analysis

- All above regret guarantees hold against the worst case
  - Matching the ***minimax optimality***
  - The environment is ***fully adversarial***

oblivious adversary

examination

adaptive adversary



interview

- However, in practice:
  - We are not always interested in the ***worst-case scenario***
  - Environments can exhibit ***specific patterns***: gradual change, periodicity…

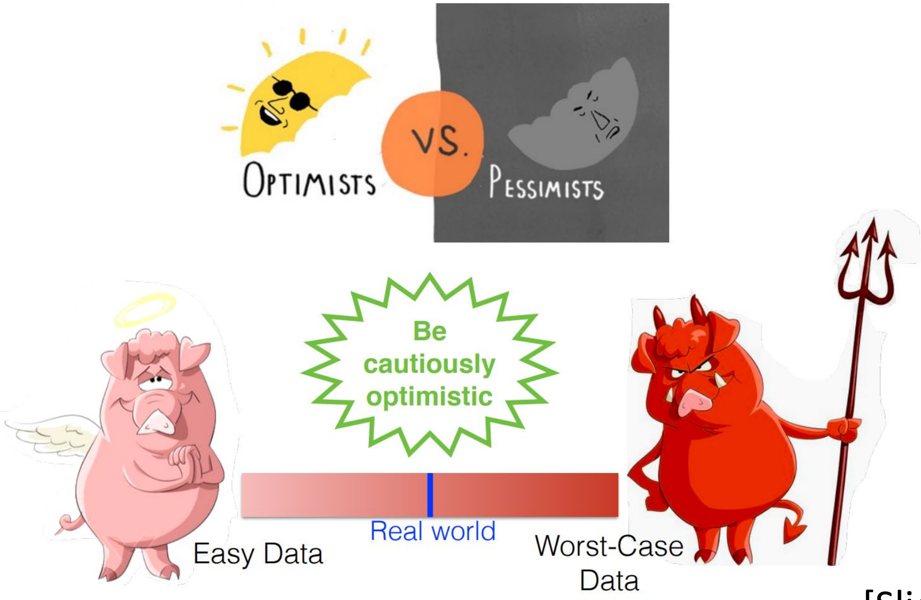
⇨ We are after ***problem-dependent*** guarantees.

Beyond the Worst-Case Analysis

- Beyond the worst-case analysis, achieving more adaptive results.

  (1) ***adaptivity***: achieving better guarantees in easy problem instances;

  (2) ***robustness***: maintaining the same worst-case guarantee.



[Slides from Dylan Foster, Adaptive Online Learning @NIPS'15 workshop]

Towards a Unified Framework

- Previous small-loss bounds seem to be ad-hoc designed.

- Is there a ***unified framework*** to get problem-dependent bounds?

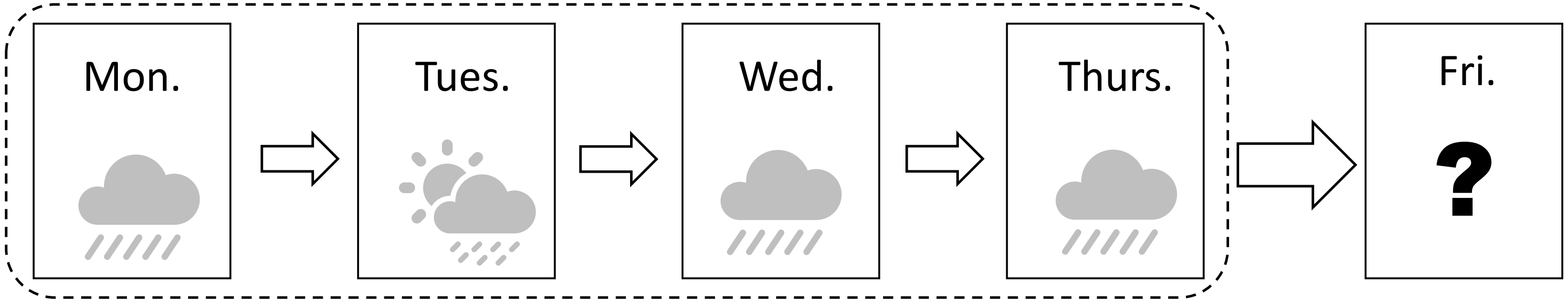
- **A reflection**: If we want to achieve adaptivity to the niceness of the environments, what does a "nice" environment mean?

$\Rightarrow$ **The environment is *"predictable"* !**

Optimistic Online Learning

- **Intuition**: what if the environment is ***"predictable"*** ?

⇨ We can to some extent ***"guess"*** the next move.



If it is within the same season
and no extreme weather

Guess: *It still seems to rain on Friday?*

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 then updates the model.

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.

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

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

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

Advanced Optimization (Fall 2024) Lecture 6. Online Mirror Descent 45

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*

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

Advanced Optimization (Fall 2024) Lecture 6. Online Mirror Descent 46

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

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

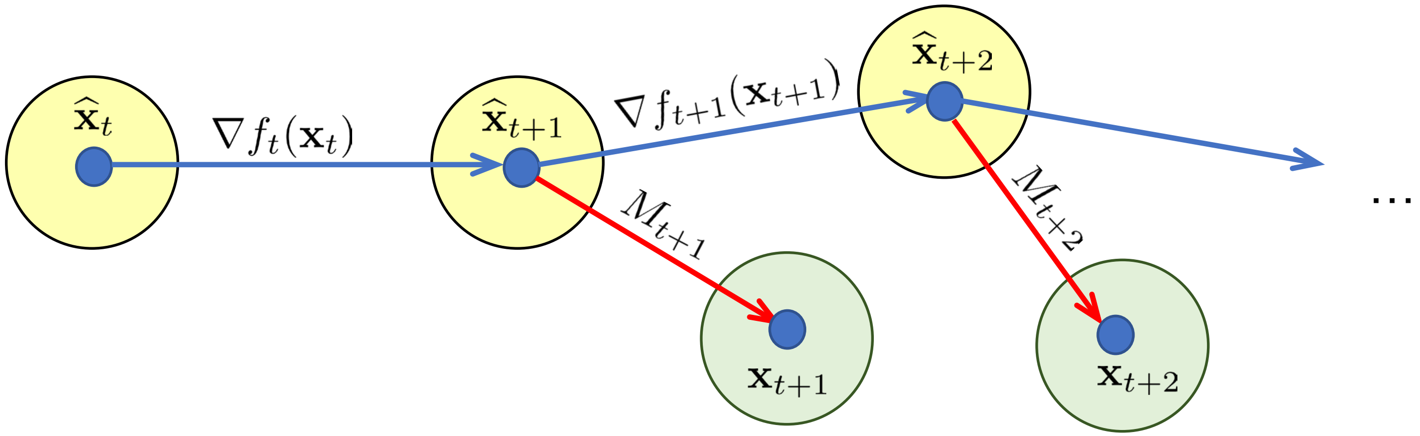
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\eta_{t+1} \left\langle M_{t+1}, \mathbf{x}\right\rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_{t+1})\right\}$$

where $M_{t+1} \in \mathbb{R}^d$ is the optimistic vector at each round.

Understanding Optimistic OMD

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

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_{t+1} \langle {\color{red}M_{t+1}}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_{t+1}) \right\}$$



Optimistic OMD: Regret Analysis

Theorem 4 (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:*

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

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

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

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.

Proof

- The key is to have a proper regret decomposition.

- Due to the two-step updates, we need to incorporate optimism and intermediate decision in regret analysis.

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

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

$$\Longrightarrow \quad f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u}\rangle \ \text{(convexity)}$$

$$= \underbrace{\langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1}\rangle}_{\texttt{term (a)}} + \underbrace{\langle M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1}\rangle}_{\texttt{term (b)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{u}\rangle}_{\texttt{term (c)}}$$

Proof

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (a)}} + \underbrace{\langle M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (b)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{u} \rangle}_{\texttt{term (c)}}$

For term (a), we use the ***stability lemma***.

Lemma 2 (Stability Lemma). *Consider the following updates:*

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

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

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

$$\begin{aligned} \texttt{term (a)} &= \langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle \\ &\leq \|\nabla f_t(\mathbf{x}_t) - M_t\|_\star \|\mathbf{x}_t - \widehat{\mathbf{x}}_{t+1}\| \leq \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_\star^2 \end{aligned}$$

Proof

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (a)}} + \underbrace{\langle M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (b)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{u} \rangle}_{\texttt{term (c)}}$

For term (b), we adopt the ***Bregman Proximal inequality***.

> **Lemma 3** (Bregman Proximal Inequality). *Consider convex optimization problem with the following update form*
>
> $$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \{ \langle \mathbf{g}_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \}.$$
>
> *Then, it satisfies the following inequality for any* $\mathbf{u} \in \mathcal{X}$*:*
>
> $$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Thus, according to update rule: $\mathbf{x}_t = \arg\min_{\mathbf{x} \in \mathcal{X}} \{ \eta_t \langle M_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_t) \}$

$$\texttt{term (b)} = \langle M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle \leq \frac{1}{\eta_t} \Big(\mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \widehat{\mathbf{x}}_t) - \mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{x}_t, \widehat{\mathbf{x}}_t) \Big)$$

Proof

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (a)}} + \underbrace{\langle M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (b)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{u} \rangle}_{\texttt{term (c)}}$

For term (c), we also adopt the ***Bregman Proximal inequality***.

Lemma 3 (Bregman Proximal Inequality). *Consider convex optimization problem with the following update form*

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

Then, it satisfies the following inequality for any $\mathbf{u} \in \mathcal{X}$*:*

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

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

$$\texttt{term (c)} = \langle \nabla f_t(\mathbf{x}_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{u} \rangle \leq \frac{1}{\eta_t} \Big(\mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_t) - \mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_{t+1}) - \mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \widehat{\mathbf{x}}_t) \Big)$$

Proof

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t) - M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (a)}} + \underbrace{\langle M_t, \mathbf{x}_t - \widehat{\mathbf{x}}_{t+1} \rangle}_{\texttt{term (b)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \widehat{\mathbf{x}}_{t+1} - \mathbf{u} \rangle}_{\texttt{term (c)}}$

Put the three terms together, we can finish the proof.

$$\texttt{term (a)} \leq \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_\star^2$$

$$\texttt{term (b)} \leq \frac{1}{\eta_t}\Big(\cancel{\mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \widehat{\mathbf{x}}_t)} - \mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{x}_t, \widehat{\mathbf{x}}_t)\Big)$$

$$\texttt{term (c)} \leq \frac{1}{\eta_t}\Big(\mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_t) - \mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_{t+1}) - \cancel{\mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \widehat{\mathbf{x}}_t)}\Big)$$

$$\Longrightarrow f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_\star^2 + \frac{1}{\eta_t}\Big(\mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_t) - \mathcal{D}_\psi(\mathbf{u}, \widehat{\mathbf{x}}_{t+1})\Big) - \frac{1}{\eta_t}\Big(\mathcal{D}_\psi(\widehat{\mathbf{x}}_{t+1}, \mathbf{x}_t) + \mathcal{D}_\psi(\mathbf{x}_t, \widehat{\mathbf{x}}_t)\Big) \qquad \square$$

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.,

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

$$\Longrightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \boxed{\eta \sum_{t=1}^T \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2} \quad \text{(quality of guess)}$$

$$+ \boxed{\frac{1}{2\eta}\sum_{t=1}^T \left(\|\mathbf{u} - \widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u} - \widehat{\mathbf{x}}_{t+1}\|_2^2\right)} \quad \text{(telescoping term)}$$

$$- \boxed{\frac{1}{2\eta}\sum_{t=1}^T \left(\|\widehat{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2\right)} \quad \text{(negative term)}$$

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.,:

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

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

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

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

→ *self-confident tuning*

Optimistic OMD: Regret Analysis

Theorem 4 (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:*

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

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

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

- For problem-independent bounds, negative terms of OMD is usually dropped;
- For problem-dependent bounds, ***negative term*** could be extremely crucial.

Part 2. Applications

- Small-Loss Bound

- Gradient-Variance Bound

- Gradient-Variation Bound

Part 2. Applications

- 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.,:

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

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

Small-Loss Bound

- 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*

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

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

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

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

$\square$

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 6 (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*

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

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.

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 \qquad \text{(quality of guess, term(a))}$$

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

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

For term (a),

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

(self-confident tuning lemma)

$$\le D\sqrt{1 + 2L\sum_{t=1}^{T} f_t(\mathbf{x}_t)} + G^2$$

(self-bounding property)

Small-Loss Bound

Proof.

$$
\begin{aligned}
\texttt{term (b)} &= \sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|\mathbf{u}-\widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u}-\widehat{\mathbf{x}}_{t+1}\|_2^2\right) \\
&\leq \frac{1}{2\eta_T}\sum_{t=1}^{T}\left(\|\mathbf{u}-\widehat{\mathbf{x}}_t\|_2^2 - \|\mathbf{u}-\widehat{\mathbf{x}}_{t+1}\|_2^2\right) \quad (\{\eta_1,\ldots,\eta_T\} \text{ decreasing step size}) \\
&\leq \frac{1}{2\eta_T}\|\mathbf{u}-\widehat{\mathbf{x}}_1\|_2^2 \quad \text{(telescoping)} \\
&\leq \frac{D}{2}\sqrt{1+2L\sum_{t=1}^{T} f_t(\mathbf{x}_t)} + \frac{D}{2} \quad \left(\text{by def of } \eta_T = \tfrac{D}{\sqrt{1+G_T}} \text{ and domain boundedness}\right)
\end{aligned}
$$

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

(converting trick)

$\square$

Part 2. Applications

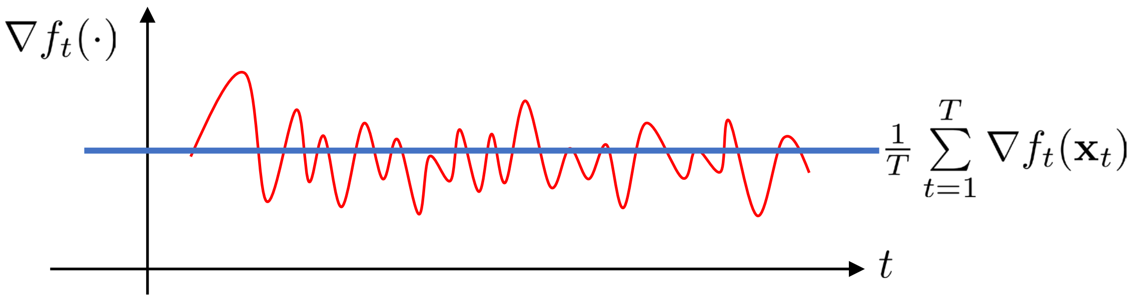
- Small-Loss Bound
- Gradient-Variance Bound
- Gradient-Variation Bound

Gradient-Variance Bound

Definition 3 (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 3 (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 $\boldsymbol{\mu}_T \triangleq \arg\min_{\boldsymbol{\mu}} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}\|_2^2 = \frac{1}{T}\sum_{t=1}^{T} \nabla f_t(\mathbf{x}_t)$.

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 3 (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 5 (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 $M_t = \widetilde{\boldsymbol{\mu}}_{t-1}$, the regret of Optimistic OMD to any comparator $\mathbf{u} \in \mathcal{X}$ is bounded as*

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

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),

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

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

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

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

Gradient-Variance Bound

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

Lemma 2. *Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. Then*

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

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

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

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

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

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

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

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

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

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.

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

Gradient-Variance Bound

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

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

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

Gradient-Variance Bound

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

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

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

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

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

Gradient-Variance Bound

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

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

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

$$\leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right) D^2 + \frac{1}{2\eta_1} D^2 \quad (\eta_t \leq \eta_{t-1} \text{ and } \|\mathbf{x} - \mathbf{y}\|_2 \leq D, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X})$$

$$\leq \frac{D^2}{2\eta_T} + \frac{1}{2\eta_1} D^2 \leq \frac{D}{2}\sqrt{1 + \mathrm{Var}_T} + \frac{D}{2} \quad \left(\frac{1}{\eta_T} = \frac{\sqrt{1+\widetilde{\mathrm{Var}}_{T-1}}}{D} \leq \frac{\sqrt{1+\mathrm{Var}_T}}{D}\right)$$

Gradient-Variance Bound

Proof. Finally, putting three terms together achieves

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

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

$$\texttt{term (c)} \geq 0$$

$$\Longrightarrow \quad \begin{aligned} \mathrm{Regret}_T &= \texttt{term (a)} + \texttt{term (b)} - \texttt{term (c)} \\ &\leq 9D\sqrt{1 + \mathrm{Var}_T + 4G^2(1 + \ln T)} + 39DG^2 + G^2 = \widetilde{\mathcal{O}}\Big(\sqrt{1 + \mathrm{Var}_T}\Big). \end{aligned}$$

Part 2. Applications

- 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

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

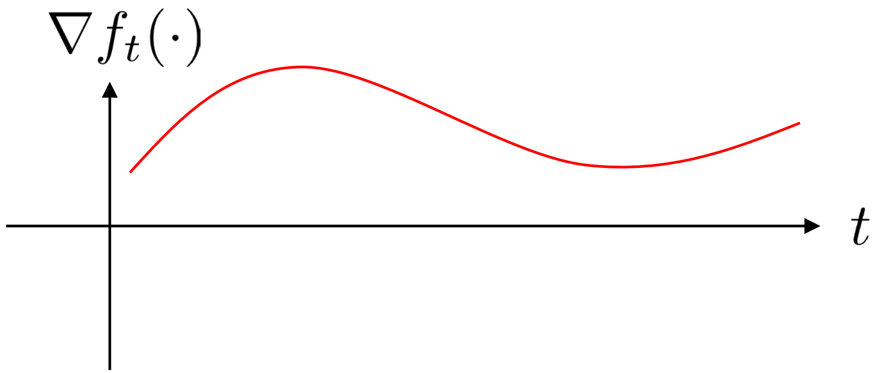
Gradient variation characterizes online functions' ***shifting intensity***.

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

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

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

Optimistic OMD for Gradient-Variation Bound

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

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

Gradient-Variation Bound

Theorem 4 (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*

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

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.

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

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

Proof

Proof. For term 1,

$$
\begin{aligned}
\sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 &\leq \sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\right\|_2^2 + G^2 \qquad (\eta_1 \triangleq 1) \\
&\leq 2\sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}_{t-1})\right\|_2^2 + 2\sum_{t=2}^{T} \eta_t \left\|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\right\|_2^2 + G^2 \\
&\leq 2\sum_{t=2}^{T} \eta_t L^2 \underbrace{\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|_2^2}_{(L\text{-smooth})} + 2D\sum_{t=2}^{T} \frac{\left\|\nabla f_t(\mathbf{x}_{t-1}) - \nabla f_{t-1}(\mathbf{x}_{t-1})\right\|_2^2}{\sqrt{1 + \sum_{s=2}^{t-1} \left\|\nabla f_s(\mathbf{x}_{s-1}) - \nabla f_{s-1}(\mathbf{x}_{s-1})\right\|_2^2}} + G^2
\end{aligned}
$$

Proof

Proof. For term (a),

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

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

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

(L-smooth)

Lemma 2. *Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. Then*

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

Proof

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

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

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

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

Proof

Proof. For the term (c), we have

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

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

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

Does this term look familiar?

Proof

Proof. We then analysis term (b),

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

Proof

Proof. Finally, putting three terms together yields

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

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

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

$$\Longrightarrow \quad \text{Regret}_T = \texttt{term (a)} + \texttt{term (b)} - \texttt{term (c)}$$

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

A Summary of 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	$M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$	$\approx \frac{D}{\sqrt{1+\widetilde{V}_{t-1}}}$	$\mathcal{O}\left(\sqrt{1+V_T}\right)$

Gradient-Variation Algorithm: Implications

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

$$\boxed{\sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|_2^2} \leq \boxed{3 \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t) - \boldsymbol{\mu}_t\|_2^2} \qquad (\leq 3\,\mathrm{Var}_T)$$

$$(\approx V_T)$$

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

$$+ \boxed{3 \sum_{t=1}^{T} \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\|_2^2}$$

$(\boldsymbol{\mu}_t = \frac{(t-1)\boldsymbol{\mu}_{t-1} + \nabla f_t(\mathbf{x}_t)}{t})$

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

$(\leq 3 \cdot \frac{\pi^2}{6})$

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

Gradient-Variation Algorithm: Implications

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

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

$(\approx V_T)$

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

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

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)

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.

Offline Scenario

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

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

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

  When the function is convex and ***smooth***, we can use this ***gradient-variation*** algorithm to obtain an averaged model with error bound as

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

Offline Scenario

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

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

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

  When the function is convex and ***Lipschitz***, we can use this ***problem-independent*** algorithm to obtain an averaged model with error bound as

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

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

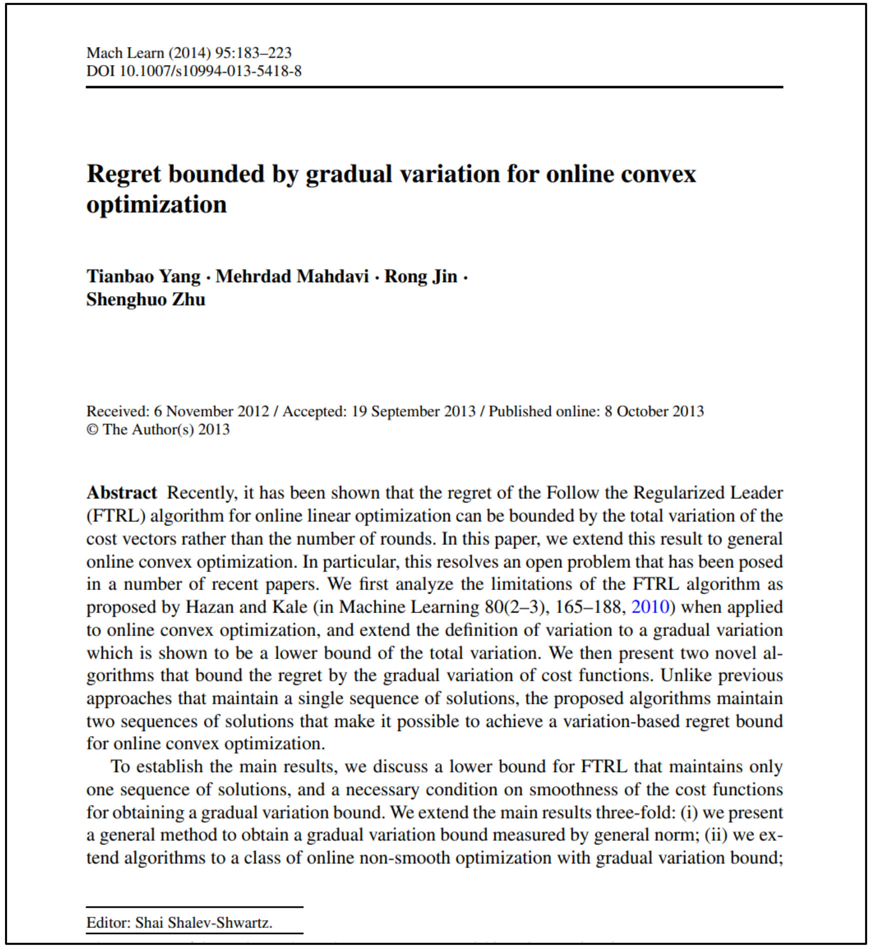
We study the online convex optimization problem, in which an online algorithm has to make repeated decisions with convex loss functions and hopes to achieve a small regret. We consider a natural restriction of this problem in which the loss functions have a small deviation, measured by the sum of the 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

© 2012 C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin & S. Zhu.



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
© The Author(s) 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.

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

COLT 2012 best student paper award

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*

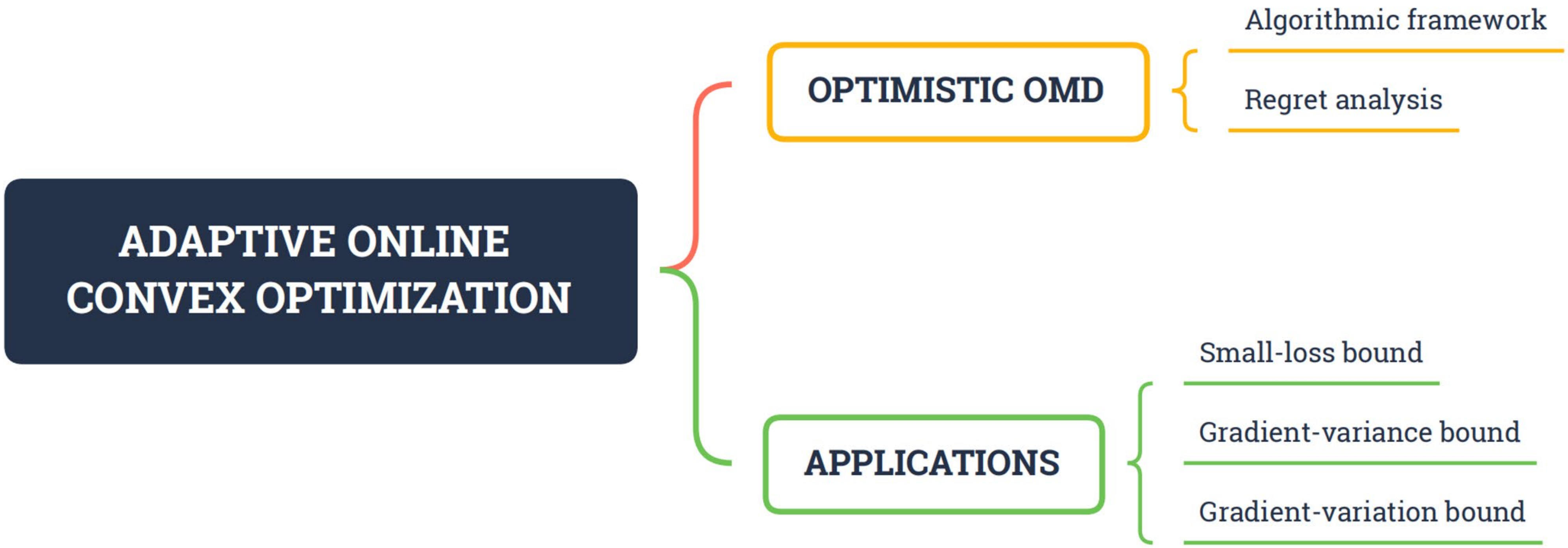
Abstract. We propose a prox-type method with efficiency estimate $O(\epsilon^{-1})$ for approximating saddle points of convex-concave $C^{1,1}$ functions and solutions of variational inequalities with monotone Lipschitz continuous operators. Application examples include matrix games, eigenvalue minimization 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.

Summary



Q & A

Thanks!