Lecture 9. Optimistic Online Mirror Descent

Advanced Optimization (Fall 2023)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

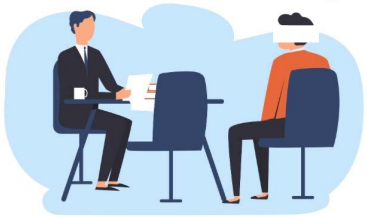
Beyond the Worst-Case Analysis

- 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…

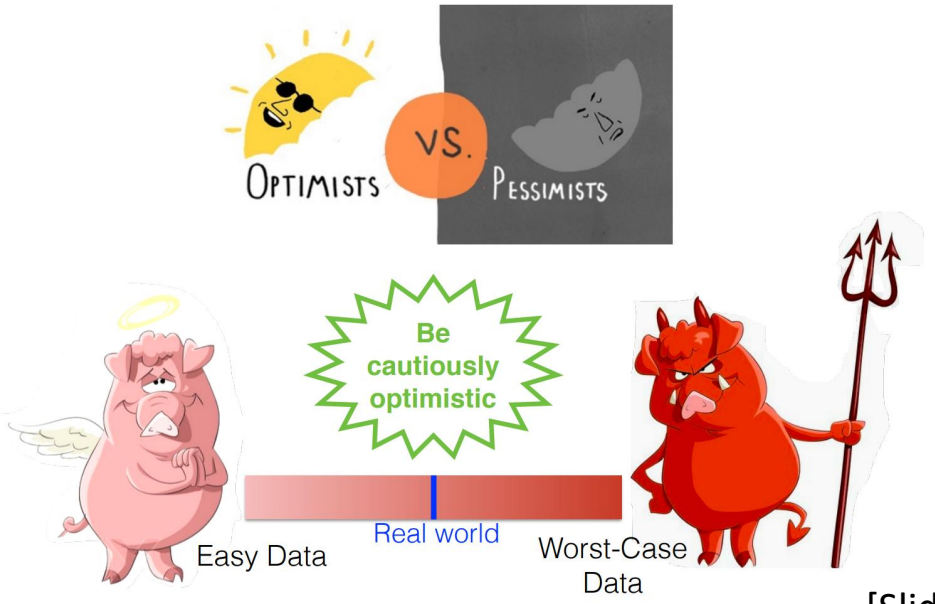
⟹ We are after some more ***problem-dependent*** guarantees.

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]

Small-Loss Bounds for PEA

Theorem 2. *Suppose that* $\forall t \in [T]$ *and* $i \in [N], 0 \le \ell_{t,i} \le 1$, *then Hedge with learning rate* $\eta \in (0,1)$ *guarantees*

$$\mathrm{Regret}_T \le \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle,$$

which can further imply

$$\mathrm{Regret}_T \le \frac{1}{1-\eta}\left(\frac{\ln N}{\eta} + \eta L_{T,i^\star}\right) = \mathcal{O}\left(\sqrt{L_{T,i^\star} \log N} + \log N\right),$$

by setting $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{\ln N}{L_{T,i^\star}}}\right\}$.

- When $L_{T,i^\star} = \mathcal{O}(T)$, it can recover the ***minimax*** $\mathcal{O}(\sqrt{T \log N})$ guarantee.
- When $L_{T,i^\star} = \mathcal{O}(1)$, the regret bound is $\mathcal{O}(\log N)$, which is independent of T!

Small-Loss Bounds for PEA

- Addressing the unpleasant dependence on $L_{T,i^\star}$ via ***self-confident tuning***.

Theorem 4. *Suppose that* $\forall t \in [T]$ *and* $i \in [N], 0 \le \boldsymbol{\ell}_{t,i} \le 1$, *then Hedge with adaptive learning rate* $\eta_t = \sqrt{\frac{\ln N}{L_t + 1}}$ *guarantees*

$$\begin{aligned}\mathrm{Regret}_T &\le 8\sqrt{(L_{T,i^\star} + 1)\ln N} + 3\ln N \\ &= \mathcal{O}\left(\sqrt{L_{T,i^\star}\log N} + \log N\right),\end{aligned}$$

where $L_t = \sum_{s=1}^{t} \langle \boldsymbol{p}_s, \boldsymbol{\ell}_s \rangle$ *is cumulative loss the learner suffered at time* t.

Key Analysis in Self-confident Tuning

Proof. From the potential-based proof, we already know that

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - L_{T,i^\star} \leq \sqrt{(L_{T-1}+1)\ln N} + \sum_{t=1}^{T} \eta_{t-1} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$$

$$\leq \sqrt{(L_{T-1}+1)\ln N} + \sum_{t=1}^{T} \frac{\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle}{\sqrt{\sum_{s=1}^{t-1} \langle \boldsymbol{p}_s, \boldsymbol{\ell}_s \rangle + 1}} \quad \left(\eta_{t-1} = \sqrt{\tfrac{\ln N}{L_{t-1}+1}}\right) \quad \left(L_{t-1} = \textstyle\sum_{s=1}^{t-1} \langle \boldsymbol{p}_s, \boldsymbol{\ell}_s \rangle\right)$$

How to bound this term?

⇨ *This is actually a common structure to handle.*

Small-Loss Bound for PEA: Proof

Proof. From previous potential-based proof, we already known that

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - L_{T,i^\star} \leq \sqrt{(L_{T-1}+1)\ln N} + \sum_{t=1}^{T} \frac{\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle}{\sqrt{\sum_{s=1}^{t-1} \langle \boldsymbol{p}_s, \boldsymbol{\ell}_s \rangle + 1}}$$

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

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

$\Rightarrow$

$$L_T - L_{T,i^\star} \leq \sqrt{(L_{T-1}+1)\ln N} + 4\sqrt{1+L_T} + 1 \qquad (\ell_i \leq 1, \forall i \in [N])$$

$$\leq \sqrt{(L_T+1)\ln N} + 4\sqrt{1+L_T} + 1$$

Small-Loss Bounds for OCO

Definition 4 (Small Loss)**.** The small-loss quantity of the OCO problem (online function $f_t : \mathcal{X} \mapsto \mathbb{R}$) is defined as

$$F_T = \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$

One essential property for small-loss bound for OCO: ***self-bounding property***.

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

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

Achieving Small-Loss Bound

- We show that under the ***self-bounding condition***, OGD can yield the desired small-loss regret bound.

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

Theorem 6 (Small-loss Bound). *Assume that f_t is L-smooth and non-negative for all $t \in [T]$, when setting $\eta_t = \frac{D}{\sqrt{1+G_t}}$, the regret of OGD to any comparator $\mathbf{u} \in \mathcal{X}$ is bounded as*

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

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

Proof

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

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

$(\eta_1 \triangleq 1)$ $\qquad$ $(G_t = \sum_{s=1}^{t} \|\nabla f_s(\mathbf{x}_s)\|_2^2)$

Lemma 1. *Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. Then*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{1 + \sum_{s=1}^{t} a_s}} \leq 2 \sqrt{1 + \sum_{t=1}^{T} a_t}$$

Several Remarks

- Remark 1: about the non-negative assumption

  When the online functions are non-negative, it is possible to redefine the small-loss quantity by incorporating each-round minimal function value.

- Remark 2: about the smoothness assumption

  Smoothness is necessary to obtain small-loss regret bound by the first-order method (can be proved by the online-to-batch conversion and existing lower bounds for deterministic optimization).

- Remark 3: take care of the way dealing with variance term

  In OGD here we use Lemma 1, while in PEA Hedge for PEA we use Lemma 2.

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**: Adaptive to the niceness of the environment. What does a "nice" environment actually mean?

⇨ The environment is ***"predictable"***.

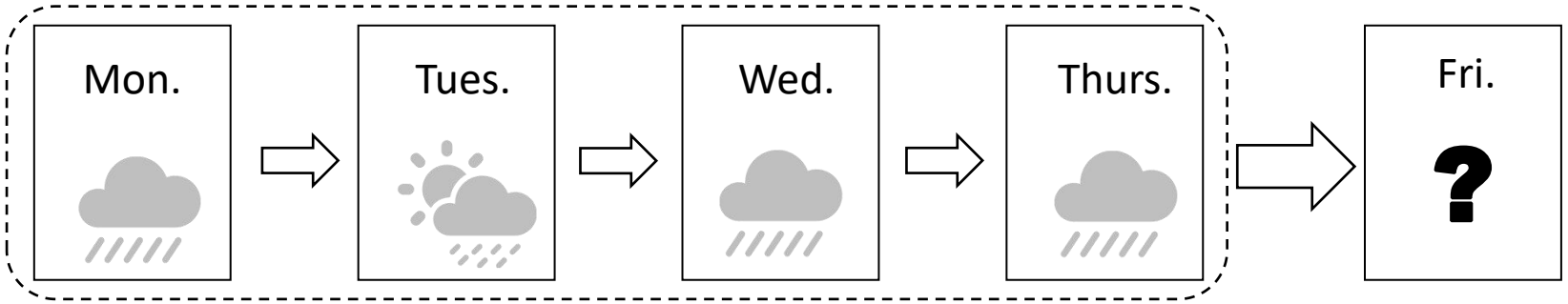
Outline

- Optimistic Online Mirror Descent
  - A Unified Framework
  - Small-Loss bound
  - Gradient-Variance bound
  - Gradient-Variation bound

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?*

Our Previous Efforts

- Review OMD update:

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

- This framework provides a unified framework for the algorithmic design and regret analysis for the worst-case scenarios.

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

Optimistic Online Mirror Descent

- We introduce a sequence of ***optimistic vector*** $\{M_t\}_{t=1}^T$ serving as the available predictable information of ***future gradients***.

Optimistic Online Mirror Descent

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

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

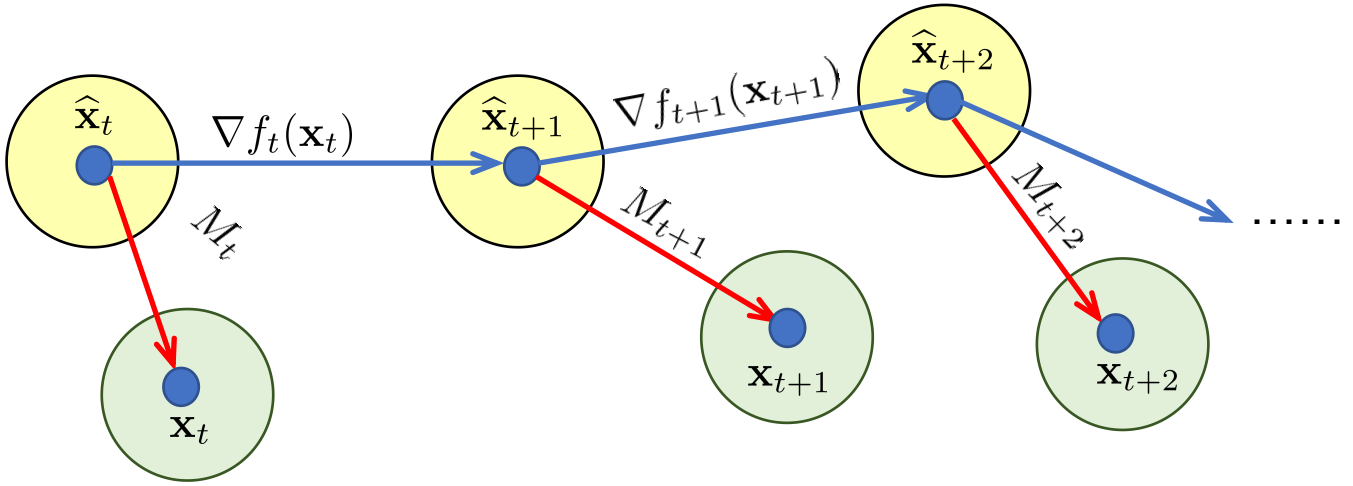
where $M_t \in \mathbb{R}^d$ is the optimistic vector at each round.

essentially two-step mirror-descent updates

Understand Optimistic OMD

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

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



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:*

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

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

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

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.

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.

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

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

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

Proof

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

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

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

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

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

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \le \|\mathbf{g} - \mathbf{g}'\|_\star .$$

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

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*

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

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

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

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

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

Proof

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

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

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

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

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

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

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.,

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

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

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

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

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

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.,:

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

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

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

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

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

→ ***self-confident tuning***

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:*

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

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

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

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

Optimistic OMD: Applications

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

Optimistic OMD: 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.,:

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

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

Small-Loss Bound

- 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):

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

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

□

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 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*

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

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.

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

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

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

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

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

(converting trick)

□

Optimistic OMD: Applications

- Small-Loss Bound

- Gradient-Variance Bound

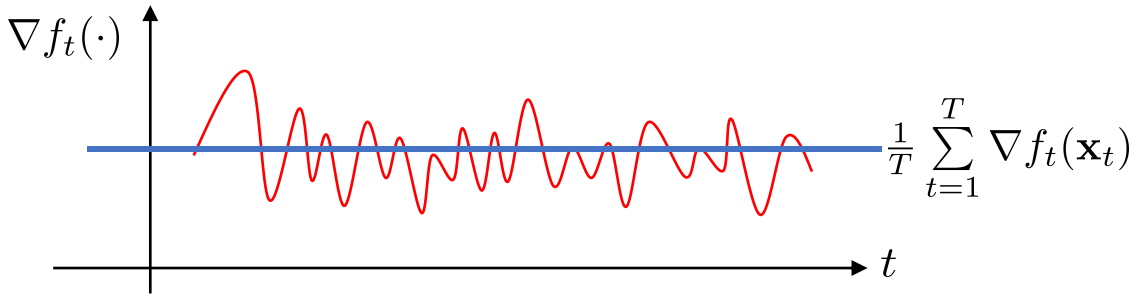
- Gradient-Variation Bound

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*

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

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),

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

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

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

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

$$(\eta_t\leq 1,\forall t\in[T])$$

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

Gradient-Variance Bound

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

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

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

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

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

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

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

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

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

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

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:

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

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:

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

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

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

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

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

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

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

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

Optimistic OMD: 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

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

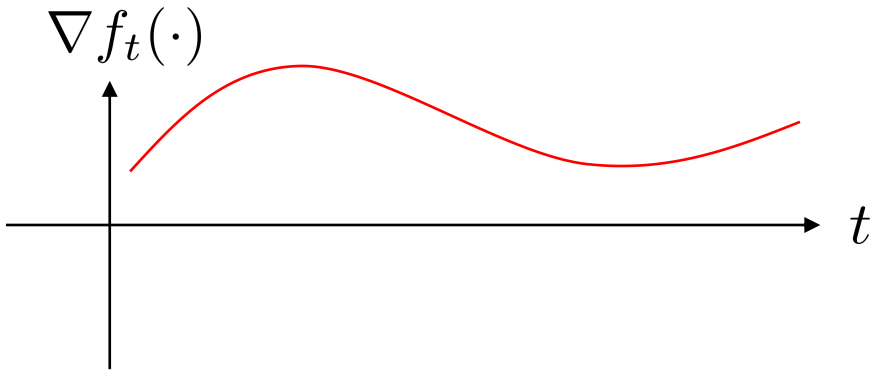
Gradient variation characterizes online functions' ***shifting intensity***.

- ***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

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

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

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

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

(negative term)

Proof

Proof. For term 1,

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

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

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

(L-smooth)

Proof

Proof. For term (a),

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

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

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

(L-smooth)

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

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

Proof

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

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

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

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

Proof

Proof. For the term (c), we have

$$
\begin{aligned}
\texttt{term (c)} &= \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\widehat{\mathbf{x}}_{t+1} - \mathbf{x}_t\|_2^2 + \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2 \right) \\
&\geq \sum_{t=2}^{T} \frac{1}{2\eta_t} \left(\|\widehat{\mathbf{x}}_t - \mathbf{x}_{t-1}\|_2^2 + \|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 \right) \qquad \left(\tfrac{1}{\eta_t} \geq \tfrac{1}{\eta_{t-1}}\right) \\
&\geq \sum_{t=2}^{T} \frac{1}{4\eta_t} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \qquad (a^2 + b^2 \geq (a+b)^2/2)
\end{aligned}
$$

Does this term look familiar?

Proof

Proof. We then analysis term (b),

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

Proof

Proof. Finally, putting three terms together yields

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

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

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

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

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

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 $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$):

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

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

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

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

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

⇨ 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 $M_t = \nabla f_{t-1}(\mathbf{x}_{t-1})$):

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

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

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

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:

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

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

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

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

Offline Scenario

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

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

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

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

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

Variation-type Bounds: History Bits

Extracting Certainty from Uncertainty: Regret Bounded by Variation in Costs

Elad Hazan
IBM Almaden
650 Harry Rd, San Jose, CA 95120
hazan@us.ibm.com

Satyen Kale
Microsoft Research
1 Microsoft Way, Redmond, WA 98052
sakale@microsoft.com

Abstract

Prediction from expert advice is a fundamental problem in machine learning. A major pillar of the field is the existence of learning algorithms whose average loss approaches that of the best expert in hindsight (in other words, whose average regret approaches zero). Traditionally the regret of online algorithms was bounded in terms of the number of prediction rounds.

Cesa-Bianchi, Mansour and Stoltz [4] posed the question whether it is be possible to bound the regret of an online algorithm by the *variation* of the observed costs. In this paper we resolve this question, and prove such bounds in the fully adversarial setting, in two important online learning scenarios: prediction from expert advice, and online linear optimization.

1 Introduction

A cornerstone of modern machine learning are algorithms for prediction from expert advice. The seminal work of Littlestone and Warmuth [12], Vovk [13] and Freund and Schapire [6] gave algorithms which, under fully adversarial cost sequences, attain average cost approaching that of the best expert in hindsight.

To be more precise, consider a prediction setting in which an online learner has access to n experts. Iteratively, the learner may chose the advice of any expert deterministically or randomly. After choosing a course of action, an adversary reveals the cost of following the advice of the different experts, from which the expected cost of the online learner is derived. The classic results mentioned above give algorithms which sequentially produce randomized decisions, such that the difference between the (expected) cost of the algorithm and the best expert in hindsight grows like $O(\sqrt{T \log n})$, where T is the number of prediction iterations. This extra additive cost is known as the regret of the online learning algorithm.

However, *a priori* it is not clear why online learning algorithms should have high regret (growing with the number of iterations) in an unchanging environment. As an extreme example, consider a setting in which there are only two experts. Suppose that the first expert always incurs cost 1, whereas the second expert always incurs cost $\frac{1}{2}$. One would expect to "figure out" this pattern quickly, and focus on the second expert, thus incurring a total cost that is at most $\frac{1}{2}$ plus at most a constant extra cost (irrespective of the number of rounds T), thus having only constant regret. However, any straightforward application of previously known analyses of expert learning algorithms only gives a regret bound of $\Theta(\sqrt{T})$ in this simple case (or very simple variations of it).

More generally, the natural bound on the regret of a "good" learning algorithm should depend on *variation* in the sequence of costs, rather than purely on the number of iterations. If the cost sequence has low variation, we expect our algorithm to be able to perform better.

This intuition has a direct analog in the stochastic setting: here, the sequence of experts' costs are independently sampled from a distribution. In this situation, a natural bound on the rate of convergence to the optimal expert is controlled by the variance of the distribution (low variance should imply faster convergence). This was formalized by Cesa-Bianchi, Mansour and Stoltz [4], who assert that *"proving such a rate in the fully adversarial setting would be a fundamental result"*.

In this paper we prove the first such regret bounds on online learning algorithms in two important scenarios: prediction from expert advice, and the more general framework of online linear optimization. Our algorithms have regret bounded by the variation of the cost sequence, in a manner that is made precise in the following sections. Thus, our bounds are tighter than *all* previous bounds, and hence yield better bounds on the applications of previous bounds (see, for example, the applications in [4]).

1.1 Online linear optimization

Online linear optimization [10] is a general framework for online learning which has received much attention recently. In this framework the decision set is an arbitrary bounded, closed, convex set in Euclidean space $K \subseteq \mathbb{R}^n$ rather than a fixed set of experts, and the costs are determined by adversarially constructed vectors, $f_1, f_2, \ldots \in \mathbb{R}^n$, such that the cost of point $x \in K$ is given by $f_t \cdot x$. The online learner iteratively chooses a point in the convex set $x_t \in K$, and then the cost vector f_t is revealed and the cost $f_t \cdot x_t$ is occurred. The performance of online learning algorithms is measured by the regret, which is defined as the difference in the total cost of the sequence of points chosen by the algorithm, viz.

Extracting Certainty from Uncertainty: Regret Bounded by Variation in Costs. COLT 2008.

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.

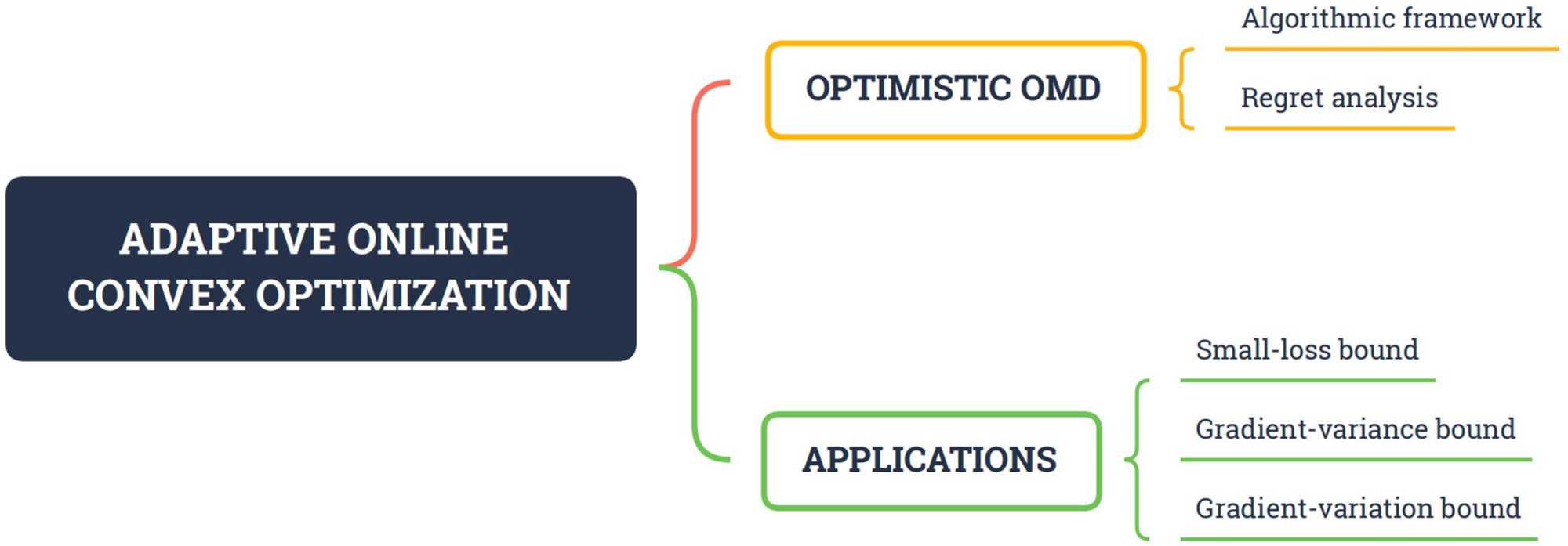
1. Introduction

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

COLT 2012
best student paper award

Online Optimization with Gradual Variations. COLT 2012.

Summary



Q & A

Thanks!