LAMDA
Learning And Mining from DatA

A *Simple* and *Optimal* Approach for Universal Online Learning with Gradient Variations

Joint work with Peng Zhao and Zhi-Hua Zhou

Presented by Yu-Hu Yan

2025.12.26

Problem Setup

Online Convex Optimization (OCO)

At each round $t = 1, 2, \ldots, T$:

- the learner submits $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$
- at the same time, environments decide a convex loss function $f_t : \mathcal{X} \mapsto \mathbb{R}$
- the learner suffers $f_t(\mathbf{x}_t)$ and receives gradient information of the loss function

Regret: Online prediction as good as the best offline model

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

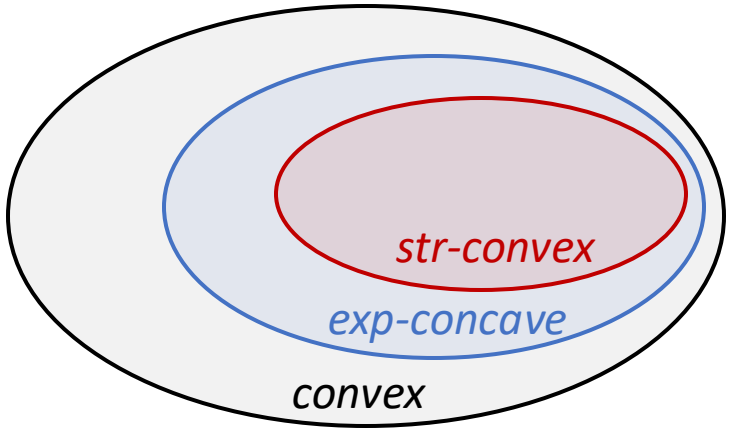
cumulative loss of best offline model

cumulative loss of the online model

Problem Setup

Curvatures in OCO

$$\text{REG}_T \triangleq \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$$



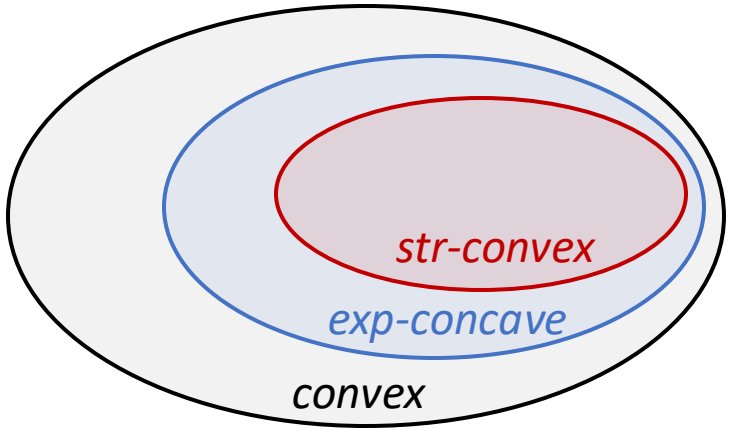
Online learning usually considers three kinds of curvatures:

- ***convex***: $f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
- ***λ-strongly convex***: $f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - \frac{\lambda}{2}\|\mathbf{x} - \mathbf{y}\|^2$.
- ***α-exp-concave***: $f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - \frac{\alpha}{2}\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle^2$.

Problem Setup

☐ Curvatures in OCO

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



In OCO, the type of ***functional curvature*** plays an important role in the best attainable regret bounds.

Function type	Algorithm	Regret
convex	Online Gradient Descent with $\eta_t \approx \frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
λ-strongly convex	Online Gradient Descent with $\eta_t = \frac{1}{\lambda t}$	$\mathcal{O}(\frac{1}{\lambda} \cdot \log T)$
α-exp-concave	Online Newton Step with α	$\mathcal{O}(\frac{1}{\alpha} \cdot d \log T)$

Problem Setup

☐ Curvatures in OCO

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

In OCO, the type of ***functional curvature*** plays an important role in the best attainable regret bounds.

Classical algorithms are only suitable for ***one specific curvature type***.

What if the ***curvature type is unknown***?

In this talk, we focus on ***universal online learning***, where the curvature is unknown.

Problem Setup

Universal Online Learning

$$\mathrm{REG}_T(\mathcal{A}, \{f_t\}_{t=1}^T) \triangleq \sum_{t=1}^T f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^T f_t(\mathbf{x})$$

In this talk, we focus on ***universal online learning***, where the curvature is unknown.

Universal Regret Minimization

$$\mathrm{REG}_T(\mathcal{A}, \{f_t\}_{t=1}^T) = \begin{cases} \mathrm{REG}_T(\mathcal{A}_{\mathrm{sc}}, \mathcal{F}_{\mathrm{sc}}^{\lambda}), & \text{when } \{f_t\}_{t=1}^T \text{ belongs to } \mathcal{F}_{\mathrm{sc}}^{\lambda}, \\ \mathrm{REG}_T(\mathcal{A}_{\mathrm{ec}}, \mathcal{F}_{\mathrm{ec}}^{\alpha}), & \text{when } \{f_t\}_{t=1}^T \text{ belongs to } \mathcal{F}_{\mathrm{ec}}^{\alpha}, \\ \mathrm{REG}_T(\mathcal{A}_{\mathrm{c}}, \mathcal{F}_{\mathrm{c}}), & \text{when } \{f_t\}_{t=1}^T \text{ belongs to } \mathcal{F}_{\mathrm{c}}, \end{cases}$$

Problem Setup

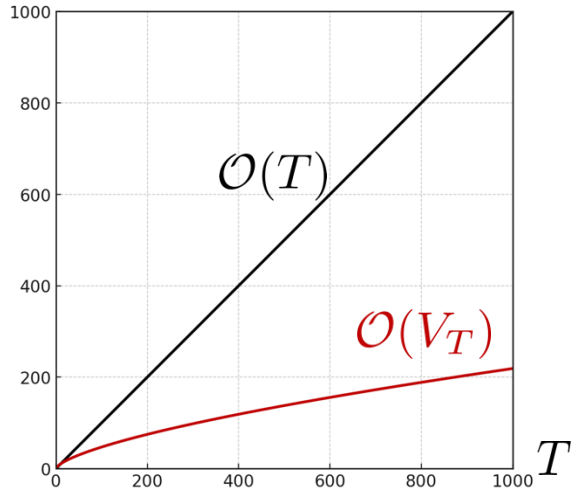
☐ Problem-dependent regret

- Regret measured by T only considers the ***worst-case*** scenarios.
- Can we exploit the ***niceness*** of environments for improved results?

Gradient variation:

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

cumulative variations in gradients, reflecting the difficulty of online problems



The regret bounds can be strengthened to $\mathcal{O}(\frac{1}{\lambda}\log V_T)$, $\mathcal{O}(\frac{d}{\alpha}\log V_T)$, and $\mathcal{O}(\sqrt{V_T})$.

Problem Setup

Gradient variation:

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

cumulative variations in gradients, reflecting the difficulty of online problems.

☐ Why do we study gradient variation?

(i) Gradient variation implies other problem-dependent quantities directly in analysis.

e.g.,

Small-loss term:

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

cumulative loss of the best model

Gradient-variance term:

$$W_T \triangleq \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} \left\| \nabla f_t(\mathbf{x}) - \frac{1}{T}\sum_{s=1}^{T} \nabla f_s(\mathbf{x}) \right\|^2$$

variance of gradients

(ii) Gradient variation can ***bridge stochastic and adversarial online optimization***.

[Sachs et al., Between stochastic and adversarial online convex optimization: Improved regret bounds via smoothness, NeurIPS 2022]

(iii) Gradient variation in achieving ***fast rates in games***.

[Syrgkanis et al., Fast convergence of regularized learning in games, NIPS 2015 (Best Paper Award)]

(iv) Gradient variation in ***accelerated convex smooth optimization***.

Main Result Overview

Our Results

Theorem 1. *Under standard assumptions (boundedness and smoothness), our algorithm*

- *achieves* $\mathcal{O}(\log V_T)$ *regret for strongly convex functions;*
- *achieves* $\mathcal{O}(d \log V_T)$ *regret for exp-concave functions;*
- *achieves* $\mathcal{O}(\sqrt{V_T})$ *regret for convex functions.*

$V_T = \sum_t \sup_{\mathbf{x} \in \mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2$ *is the gradient variation.*

A ***single*** algorithm with simultaneously ***optimal gradient-variation regret*** bounds for convex/exp-concave/strongly convex functions.

Main Result Overview

☐ Comparison with previous works

Table 1: Comparison with existing results. The second column shows the regret bounds for strongly convex, exp-concave, and convex functions, following the $\mathcal{O}(\cdot)$-notation. "# Gradient" is the number of gradient queries in each round, where "1" represents exactly one gradient query. "# Base" stands for the number of base learners.

Works	**Regret Bounds**			**Efficiency**	
	Strongly Convex	Exp-concave	Convex	# Gradient	# Base
van Erven and Koolen [2016]	$d \log T$	$d \log T$	$\sqrt{T}$	1	$\log T$

The ***first*** universal result from Tim van Erven.

Main Result Overview

☐ Comparison with previous works

Table 1: Comparison with existing results. The second column shows the regret bounds for strongly convex, exp-concave, and convex functions, following the $\mathcal{O}(\cdot)$-notation. "# Gradient" is the number of gradient queries in each round, where "1" represents exactly one gradient query. "# Base" stands for the number of base learners.

Works	**Regret Bounds**			**Efficiency**	
	Strongly Convex	Exp-concave	Convex	# Gradient	# Base
van Erven and Koolen [2016]	$d \log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Wang et al. [2019]	$\log T$	$d \log T$	$\sqrt{T}$	1	$\log T$

Improved by [Wang et al., UAI 2019] for strongly convex functions.

Main Result Overview

☐ Comparison with previous works

Table 1: Comparison with existing results. The second column shows the regret bounds for strongly convex, exp-concave, and convex functions, following the $\mathcal{O}(\cdot)$-notation. "# Gradient" is the number of gradient queries in each round, where "1" represents exactly one gradient query. "# Base" stands for the number of base learners.

Works	**Regret Bounds**			**Efficiency**	
	Strongly Convex	Exp-concave	Convex	# Gradient	# Base
van Erven and Koolen [2016]	$d \log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Wang et al. [2019]	$\log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Zhang et al. [2022]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\sqrt{F_T}$	$\log T$	$\log T$

The first ***problem-dependent*** regret in universal OCO.

Main Result Overview

☐ Comparison with previous works

Table 1: Comparison with existing results. The second column shows the regret bounds for strongly convex, exp-concave, and convex functions, following the $\mathcal{O}(\cdot)$-notation. "# Gradient" is the number of gradient queries in each round, where "1" represents exactly one gradient query. "# Base" stands for the number of base learners.

Works	**Regret Bounds**			**Efficiency**	
	Strongly Convex	Exp-concave	Convex	# Gradient	# Base
van Erven and Koolen [2016]	$d \log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Wang et al. [2019]	$\log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Zhang et al. [2022]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\sqrt{F_T}$	$\log T$	$\log T$
Yan et al. [2023]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\min\{\sqrt{V_T \log V_T}, \sqrt{F_T \log F_T}\}$	1	$(\log T)^2$

Our ***improved*** gradient-variation bound via a ***multi-layer online ensemble*** approach.

Main Result Overview

☐ Comparison with previous works

Table 1: Comparison with existing results. The second column shows the regret bounds for strongly convex, exp-concave, and convex functions, following the $\mathcal{O}(\cdot)$-notation. "# Gradient" is the number of gradient queries in each round, where "1" represents exactly one gradient query. "# Base" stands for the number of base learners.

Works	Regret Bounds			Efficiency	
	Strongly Convex	Exp-concave	Convex	# Gradient	# Base
van Erven and Koolen [2016]	$d \log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Wang et al. [2019]	$\log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Zhang et al. [2022]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\sqrt{F_T}$	$\log T$	$\log T$
Yan et al. [2023]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\min\{\sqrt{V_T \log V_T}, \sqrt{F_T \log F_T}\}$	1	$(\log T)^2$

An open problem in [Yan et al., NeurIPS 2023]:

Is it possible to achieve the ***optimal*** universal gradient-variation regret, with an ***efficient*** approach (i.e., one gradient query and $\mathcal{O}(\log T)$ base learners)?

Main Result Overview

LAMDA
Learning And Mining from DatA

☐ Comparison with previous works

Table 1: Comparison with existing results. The second column shows the regret bounds for strongly convex, exp-concave, and convex functions, following the $\mathcal{O}(\cdot)$-notation. "# Gradient" is the number of gradient queries in each round, where "1" represents exactly one gradient query. "# Base" stands for the number of base learners.

Works	**Regret Bounds**			**Efficiency**	
	Strongly Convex	Exp-concave	Convex	# Gradient	# Base
van Erven and Koolen [2016]	$d \log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Wang et al. [2019]	$\log T$	$d \log T$	$\sqrt{T}$	1	$\log T$
Zhang et al. [2022]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\sqrt{F_T}$	$\log T$	$\log T$
Yan et al. [2023]	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\min\{\sqrt{V_T \log V_T}, \sqrt{F_T \log F_T}\}$	1	$(\log T)^2$
Ours	$\log \min\{V_T, F_T\}$	$d \log \min\{V_T, F_T\}$	$\min\{\sqrt{V_T}, \sqrt{F_T}\}$	1	$\log T$

This work: ***Optimal*** problem-dependent regret with an ***efficient*** approach.

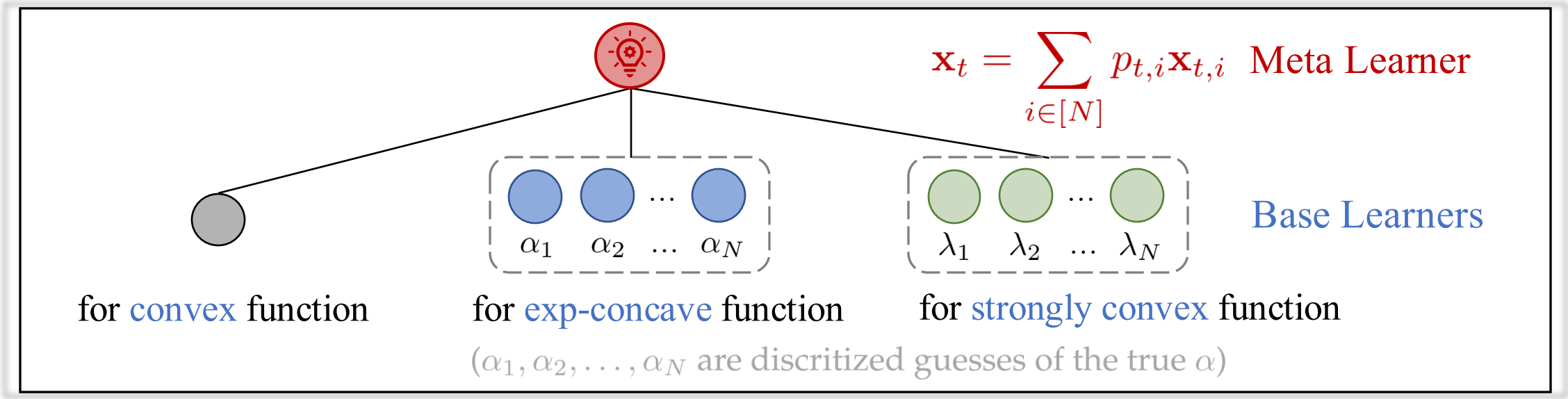
A General Framework

Universal Regret Minimization

$$\mathrm{REG}_T(\mathcal{A}, \{f_t\}_{t=1}^T) = \begin{cases} \mathrm{REG}_T(\mathcal{A}_{\mathrm{sc}}, \mathcal{F}_{\mathrm{sc}}^{\lambda}), & \text{when } \{f_t\}_{t=1}^T \text{ belongs to } \mathcal{F}_{\mathrm{sc}}^{\lambda}, \\ \mathrm{REG}_T(\mathcal{A}_{\mathrm{ec}}, \mathcal{F}_{\mathrm{ec}}^{\alpha}), & \text{when } \{f_t\}_{t=1}^T \text{ belongs to } \mathcal{F}_{\mathrm{ec}}^{\alpha}, \\ \mathrm{REG}_T(\mathcal{A}_{\mathrm{c}}, \mathcal{F}_{\mathrm{c}}), & \text{when } \{f_t\}_{t=1}^T \text{ belongs to } \mathcal{F}_{\mathrm{c}}, \end{cases}$$

Online Ensemble [Zhao-Zhang-Zhang-Zhou, JMLR 2024]

General goal: *To handle the* ***uncertainty*** *of environments.*



- ***Base learners*** guess the curvature (str-convex/exp-concave/cvx).
- ***Meta learner*** tracks the best base learner.

Basic Idea

- **Regret decomposition:** How to control meta-regret in two layers

$$\mathrm{REG}_T = \left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star})\right] + \left[\sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right]$$

meta regret *base regret*

- **Key idea:** Exploiting the ***second-order regret bound*** on the meta level

[Zhang et al., ICML 2022]

$$\sum_{t=1}^{T}\langle p_t, \ell_t\rangle - \sum_{t=1}^{T}\ell_{t,i} \le \mathcal{O}\left(\sqrt{\sum_{t=1}^{T} r_{t,i}^2}\right)$$

(second-order bound, e.g., Adapt-ML-Prod)

[Gaillard et al, COLT 2014]

$$\Rightarrow \quad \ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i}\rangle \qquad r_{t,i} \triangleq \langle p_t, \ell_t\rangle - \ell_{t,i}$$

$$\sum_{t=1}^{T}\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star}\rangle \lesssim \sqrt{\sum_{t=1}^{T}\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star}\rangle^2}$$

Basic Idea

- **Regret decomposition:** How to control meta-regret in two layers

$$\mathrm{REG}_T = \left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star})\right] + \left[\sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right]$$

meta regret *base regret*

- **Key idea:** Exploiting the ***second-order regret bound*** on the meta level

$$\Rightarrow \quad \ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle \qquad \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle \lesssim \sqrt{\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle^2}$$
$$r_{t,i} \triangleq \langle p_t, \ell_t \rangle - \ell_{t,i}$$

e.g., exp-concave

$$\Rightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) \leq \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle - \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle^2 \leq \mathcal{O}(1)$$

Basic Idea

- **Regret decomposition:** How to control meta-regret in two layers

$$\mathrm{REG}_T = \left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star})\right] + \left[\sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right]$$

meta regret *base regret*

- **Key idea:** Exploiting the ***second-order regret bound*** on the meta level

$$\Rightarrow \quad \ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle \qquad \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle \lesssim \sqrt{\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle^2}$$
$$r_{t,i} \triangleq \langle p_t, \ell_t \rangle - \ell_{t,i}$$

e.g., strongly convex

$$\Rightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) \leq \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle - \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{x}_{t,i^\star}\|^2 \leq \mathcal{O}(1)$$

Basic Idea

- **Regret decomposition:** How to control meta-regret in two layers

$$\mathrm{REG}_T = \left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star})\right] + \left[\sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right]$$

meta regret *base regret*

- **Key idea:** Exploiting the ***second-order regret bound*** on the meta level

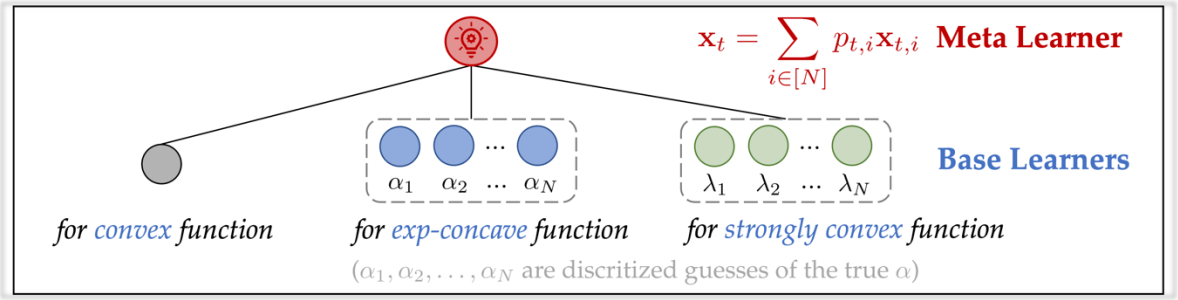
$$\ell_{t,i} \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i} \rangle$$
$$r_{t,i} \triangleq \langle p_t, \ell_t \rangle - \ell_{t,i}$$

$$\Rightarrow \quad \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle \lesssim \sqrt{\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle^2}$$

e.g., convex

$$\Rightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_{t,i^\star}) \leq \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle \lesssim \sqrt{\sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star} \rangle^2}$$

Key Techniques



☐ How to obtain gradient-variation regret?

$$\begin{cases} \textbf{What we want: } V_T = \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 \\ \textbf{What we have: } \bar{V}_T = \sum_{t=2}^{T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \end{cases}$$

(in the t-th round, we query the gradient of $\nabla f_t(\mathbf{x}_t)$)

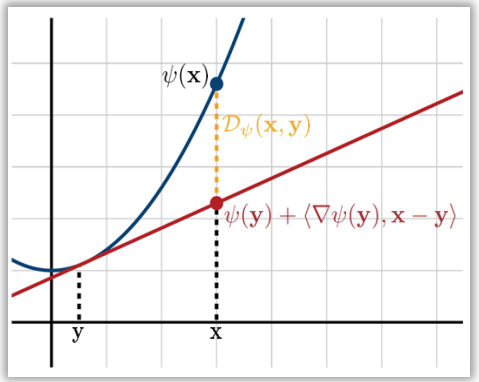
Our Solution:

A ***tighter*** upper bound for ***squared*** gradient change:

Definition 1 (Theorem 2.1.5 of (Nesterov, 2018)). $f(\cdot)$ is L-smooth over $\mathbb{R}^d$ if and only if $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq 2L\mathcal{D}_f(\mathbf{y}, \mathbf{x})$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

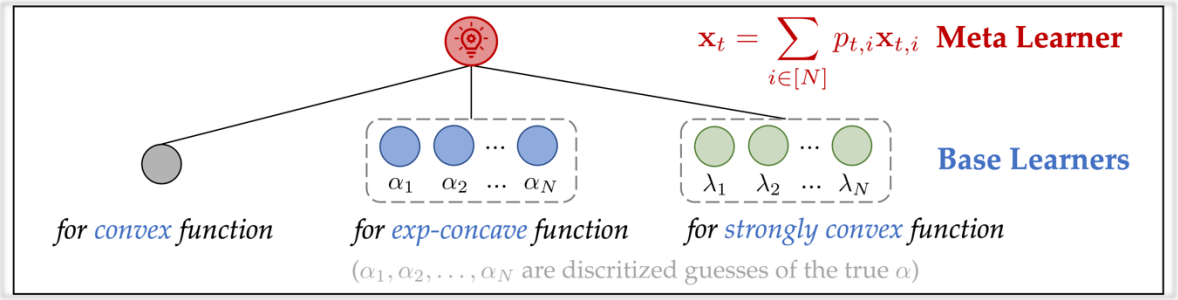
Bregman divergence: $\mathcal{D}_f(\mathbf{x}, \mathbf{y}) \triangleq f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$

tighter than $\|\nabla f_t(\mathbf{x}) - \nabla f_t(\mathbf{y})\|^2 \leq L^2\|\mathbf{x} - \mathbf{y}\|^2$ by the smoothness assumption



Bregman divergence

Key Techniques



☐ How to obtain gradient-variation regret?

$$\begin{cases} \textbf{What we want: } V_T = \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 \\ \textbf{What we have: } \bar{V}_T = \sum_{t=2}^{T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \end{cases}$$

(in the t-th round, we query the gradient of $\nabla f_t(\mathbf{x}_t)$)

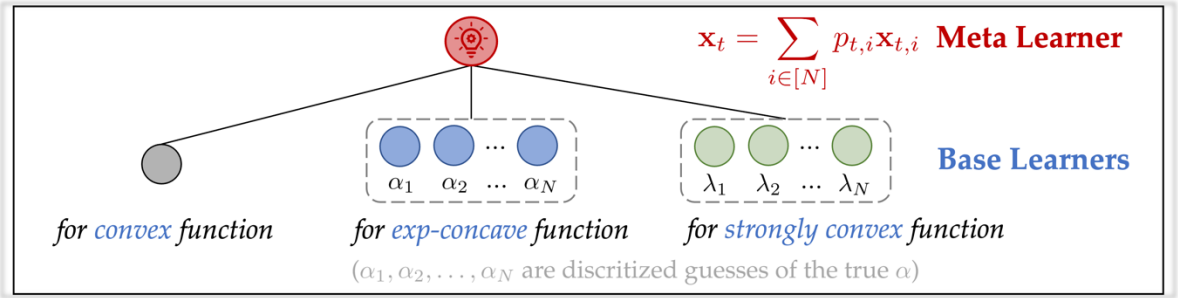
Our Solution:

Definition 1 (Theorem 2.1.5 of (Nesterov, 2018)). $f(\cdot)$ is L-smooth over $\mathbb{R}^d$ if and only if $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq 2L\mathcal{D}_f(\mathbf{y}, \mathbf{x})$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

$$\bar{V}_T \lesssim \sum_{t=2}^{T} \left(\|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}^\star)\|^2 + \|\nabla f_t(\mathbf{x}^\star) - \nabla f_{t-1}(\mathbf{x}^\star)\|^2 + \|\nabla f_{t-1}(\mathbf{x}^\star) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \right)$$

$$\lesssim L\sum_{t=2}^{T} \mathcal{D}_{f_t}(\mathbf{x}^\star, \mathbf{x}_t) + V_T + L\sum_{t=2}^{T} \mathcal{D}_{f_{t-1}}(\mathbf{x}^\star, \mathbf{x}_{t-1}) \leq 2L\sum_{t=1}^{T} \mathcal{D}_{f_t}(\mathbf{x}^\star, \mathbf{x}_t) + V_T,$$

Key Techniques



How to obtain gradient-variation regret?

$$\bar{V}_T \lesssim \sum_{t=2}^{T} \left(\|\nabla f_t(\mathbf{x}_t) - \nabla f_t(\mathbf{x}^\star)\|^2 + \|\nabla f_t(\mathbf{x}^\star) - \nabla f_{t-1}(\mathbf{x}^\star)\|^2 + \|\nabla f_{t-1}(\mathbf{x}^\star) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2 \right)$$

$$\lesssim L\sum_{t=2}^{T} \mathcal{D}_{f_t}(\mathbf{x}^\star, \mathbf{x}_t) + V_T + L\sum_{t=2}^{T} \mathcal{D}_{f_{t-1}}(\mathbf{x}^\star, \mathbf{x}_{t-1}) \leq 2L\sum_{t=1}^{T} \mathcal{D}_{f_t}(\mathbf{x}^\star, \mathbf{x}_t) + V_T,$$

Solution: $$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}^\star) = \sum_{t=1}^{T} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle - \sum_{t=1}^{T} \mathcal{D}_{f_t}(\mathbf{x}^\star, \mathbf{x}_t)$$ (*algorithm-independent!*)

Negative Bregman divergence can be seen as ***compensation for linearization***.

Summary

- **Problem:** universal online learning with gradient variations
- **General framework:** online ensemble with adaptivity
- **General analysis:** meta-base regret decomposition
- **Key techniques:** empirical gradient variation decomposition + negative Bregman divergence from linearization

Universal online learning with gradient variations: A multi-layer online ensemble approach, NeurIPS'23 **(Spotlight)**

A simple and optimal approach for universal online learning with gradient variations, NeurIPS'24

Thanks!

References

❖ **Universal Online Learning:**

- van Erven et al.. Metagrad: multiple learning rates in online learning, NIPS'16
- Wang et al.. Adaptivity and optimality: a universal algorithm for online convex optimization, UAI'19
- Wang et al.. Adapting to smoothness: a more universal algorithm for online convex optimization, AAAI'20
- Zhang et al.. A simple yet universal strategy for online convex optimization, ICML'22

❖ **Collaborative Online Ensemble:**

- Zhao et al.. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization, JMLR

❖ **Our Works:**

- Yan-Zhao-Zhou. Universal online learning with gradual variations: A multi-layer online ensemble approach, NeurIPS'23
- Yan-Zhao-Zhou. A Simple and Optimal Approach for Universal Online Learning with Gradient Variations, NeurIPS'24