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

Yu-Hu Yan, Peng Zhao, Zhi-Hua Zhou

LAMDA Group, Nanjing University, China

NANJING UNIVERSITY · LAMDA Learning And Mining from DatA · NEURAL INFORMATION PROCESSING SYSTEMS

Online Convex Optimization (OCO)

Online Learning: data comes as a ***stream***, and model ***online updates***

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
- the learner suffers $f_t(\mathbf{x}_t)$ and receives gradient information

Goal: minimize ***regret***

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

cumulative loss of best offline model

cumulative loss of the online model

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

Function type	Algorithm	Regret
λ-strongly convex	OGD with $\eta_t = \frac{1}{\lambda t}$	$\mathcal{O}(\frac{1}{\lambda}\log T)$
α-exp-concave	ONS knowing α	$\mathcal{O}(\frac{d}{\alpha}\log T)$
convex	OGD with $\eta_t \approx \frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$

Burdensome in practice!

Universal OCO with Gradient Variations

*Recent studies consider **two levels of adaptivity**.*

❖ **High-Level:** adaptive to ***unknown*** function curvature

Target: *a single algorithm that is agnostic to function curvature*

$f_t(\cdot)$ can be either convex, exp-concave, or strongly convex, and is ***unknown***.
- convex: $f(\mathbf{x}) - f(\mathbf{y}) \le \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
- α-exp-concave: $f(\mathbf{x}) - f(\mathbf{y}) \le \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle - \alpha/2 \cdot \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle^2$.
- λ-strongly convex: $f(\mathbf{x}) - f(\mathbf{y}) \le \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle - \lambda/2 \cdot \|\mathbf{x}-\mathbf{y}\|^2$.

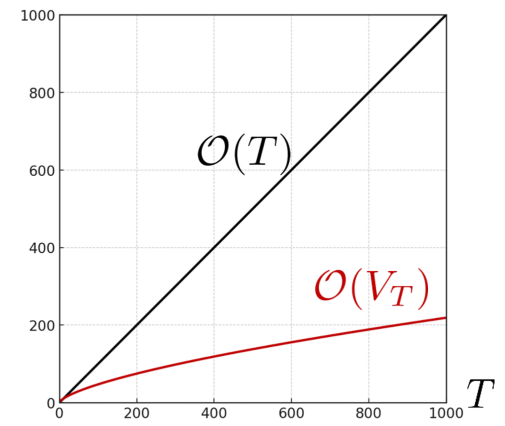
❖ **Low-Level:** adaptive to ***unknown*** niceness of environments

Target: *regret bounds measured by problem-dependent quantities*

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 $\mathcal{O}(\sqrt{V_T})$.

⇨ **Challenge:** ***handling uncertainties of two levels simultaneously***

Recent Progress and Our Contribution

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]	$\log V_T$	$d\log V_T$	$\sqrt{T}$	$\log T$	$\log T$
Yan et al. [2023]	$\log V_T$	$d\log V_T$	$\sqrt{V_T \log V_T}$	1	$(\log T)^2$

\# **Gradient:** number of gradient queries. "1" is the best = as efficient as OGD.

\# **Base:** number of base learners. "$\log T$" is necessary for online ensemble.

*Can we achieve the **optimal universal gradient-variation regret**, with an **efficient** approach (i.e., 1 gradient query and $\mathcal{O}(\log T)$ base learners)?*

Theorem 1. *Under standard assumptions, our two-layer online ensemble algorithm*
- *achieves $\mathcal{O}(\frac{1}{\lambda}\log V_T)$ regret for strongly convex functions;*
- *achieves $\mathcal{O}(\frac{d}{\alpha}\log V_T)$ regret for exp-concave functions;*
- *achieves $\mathcal{O}(\sqrt{V_T})$ regret for convex functions,*

using 1 gradient per round.

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

Key References:

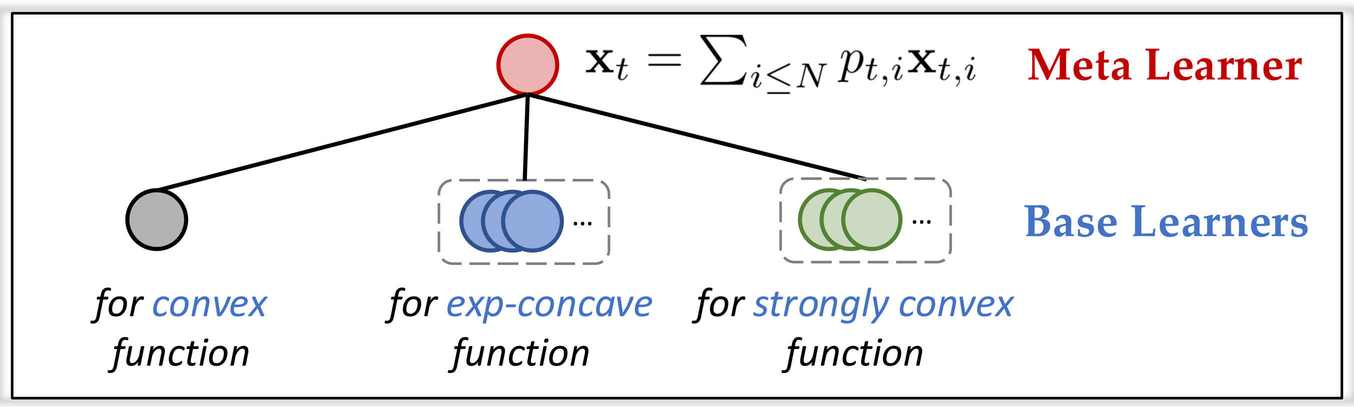
[1] Van Erven-Koolen, MetaGrad: Multiple Learning Rates in Online Learning, NIPS'16

[2] Zhang-Wang-Yi-Yang, A Simple yet Universal Strategy for Online Convex Optimization, ICML'22

[3] Yan-Zhao-Zhou, Universal Online Learning with Gradient Variations: A Multi-layer Online Ensemble Approach, NeurIPS'23

A General Online Ensemble Framework

Basic Idea: Online Ensemble [Zhao-Zhang-Zhang-Zhou, JMLR'24]



Ensemble is effective in handling uncertainty, e.g., in dynamic/adaptive regret minimization.

Regret Decomposition: ***meta regret*** + ***base regret***

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

meta regret · *base regret*

($i^\star$: the index of *the best base learner* with the right guess of the curvature type and the closest guess of the curvature coefficient)

- **base regret:** black-box optimization
- **meta regret:** exp-concave functions *(also holds for strongly convex ones)*

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

(second-order bound, e.g., Adapt-ML-Prod) [Gaillard-Stoltz-Van Erven, COLT'14]

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

Technical Contributions

Contribution I: a novel analysis for *empirical gradient variation*

$$V_T \triangleq \sum_{t\le T} \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2 \Leftarrow \bar{V}_T \triangleq \sum_{t\le T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2$$

(one-gradient model)

Our Solution: *with **two key analytical components***

① Part I: a useful smoothness property

Previous: $\bar{V}_T \lesssim \sum_{t\le T} \|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_t)\|^2 + \sum_{t\le T} \|\nabla f_{t-1}(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})\|^2$ [Yan-Zhao-Zhou, NeurIPS'23]

$\lesssim V_T + L^2 \sum_{t\le T} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$ (smoothness: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|$)

Handling algorithmic stability is challenging, leading to suboptimal results and less efficient algorithms.

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

Tighter for squared gradient variation than $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \le L^2\|\mathbf{x}-\mathbf{y}\|^2$ as $\mathcal{D}_f(\mathbf{y},\mathbf{x}) \le \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$.

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

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

Next Step: cancel this term.

② Part II: negative term from linearization

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

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

*Bregman divergence can be seen as **compensation from linearization**.*

Contribution II: analysis for empirical gradient variation on *surrogates*

*For gradient efficiency, we use the **surrogate** functions in* [Yan et al., NeurIPS'23].

Taking λ-strongly convex functions as an example:

$$\sum_{t\le T} f_t(\mathbf{x}_t) - \sum_{t\le T} f_t(\mathbf{x}^\star) \le \sum_{t\le T}\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle - \frac{\lambda_i^\star}{2}\sum_{t\le T}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

(strong convexity and property of the best best learner: $\lambda_i^\star \le \lambda \le 2\lambda_i^\star$)

$$= \sum_{t\le T}\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t,i^\star}\rangle - \frac{\lambda_i^\star}{2}\sum_{t\le T}\|\mathbf{x}_t - \mathbf{x}_{t,i^\star}\|^2 \quad \textit{meta regret as previous}$$

$$+ \left[\sum_{t\le T}\left(\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t,i^\star}\rangle + \frac{\lambda_i^\star}{2}\|\mathbf{x}_{t,i^\star} - \mathbf{x}_t\|^2\right)\right] - \left[\sum_{t\le T}\left(\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}^\star\rangle + \frac{\lambda_i^\star}{2}\|\mathbf{x}^\star - \mathbf{x}_t\|^2\right)\right]$$

base regret on surrogates: $h_{t,i}^{\mathrm{sc}}(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \frac{\lambda_i}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$

⇨ Base learners can update efficiently on surrogates, using ***only one*** gradient $\nabla f_t(\mathbf{x}_t)$.

Running an optimistic algorithm on base regret gives $\mathcal{O}(\frac{1}{\lambda}\log D_T)$, where

$$D_T = \sum_{t\le T}\|\nabla h_{t,i}^{\mathrm{sc}}(\mathbf{x}_{t,i}) - \nabla h_{t-1,i}^{\mathrm{sc}}(\mathbf{x}_{t-1,i})\|^2 = \sum_{t\le T}\|\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1}) + \lambda_i(\mathbf{x}_{t,i} - \mathbf{x}_t) - \lambda_i(\mathbf{x}_{t-1,i} - \mathbf{x}_{t-1})\|^2$$

gradient variation on surrogates · *handle as previous* · ??

Requirement: *we should avoid directly dealing with the algorithmic stability of $\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$.*

Our Solution:

$$\sum_{t\le T}\|(\mathbf{x}_{t,i} - \mathbf{x}_t) - (\mathbf{x}_{t-1,i} - \mathbf{x}_{t-1})\|^2 \lesssim \sum_{t\le T}\|\mathbf{x}_{t,i} - \mathbf{x}_t\|^2 + \sum_{t\le T}\|\mathbf{x}_{t-1,i} - \mathbf{x}_{t-1}\|^2 \le 2\sum_{t\le T}\|\mathbf{x}_{t,i} - \mathbf{x}_t\|^2$$

Key observation: *it is controlled when **aggregated across the whole time horizon**, using the negative term in the meta regret.*