Gradient-Variation Online Adaptivity for Accelerated Optimization with Hölder Smoothness (spotlight)

Yuheng Zhao[1], Yu-Hu Yan[1], Kfir Yehuda Levy[2], Peng Zhao[1] [1] Nanjing University, China [2] Technion, Haifa, Israel

Our Paper (ArXiv) Homepage (Yuheng Zhao)

TL;DR

- ***Accelerated optimization*** *can be understood by* ***gradient-variation online learning***
- *We achieve* ***universality****: automatically adapting to an unknown level of (Hölder) smoothness*
  - ✓ *Gradient-variation online learning*
  - ✓ *Convex OPT & strongly convex OPT*

Acceleration in Optimization (OPT)

Smoothness ***is crucial for*** ***acceleration*** ***in optimization.***

$$\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x}), \quad \textit{with (stochastic) gradient oracle } \mathbf{g}(\cdot).$$

Goal: ***convergence rate*** *given T oracle queries* $\mathrm{GAP}_T \triangleq \mathcal{L}(\bar{\mathbf{x}}_T) - \mathcal{L}(\mathbf{x}^\star) \le \epsilon(T)$.

Optimal rates (acceleration; $\kappa \triangleq L/\lambda$)

Setting	Non-smooth (Lipschitz)	L-Smooth
Convex	$\mathcal{O}\left(1/\sqrt{T}\right)$	$\mathcal{O}\left(1/T^2\right)$
λ-Strongly Convex	$\mathcal{O}\left(1/(\lambda T)\right)$	$\mathcal{O}\left(\exp\left(-T/\sqrt{\kappa}\right)\right)$

Reformulating OPT as Online Learning

Online-to-Batch (O2B) Conversion: *leveraging the rich adaptivity in online learning to enhance optimization.*
Online Learning: *online updates based on online functions.*

☐ O2B: with an online algorithm and weights $\{\alpha_t\}_{t=1}^T$

1. *Passing to oracle online algorithm's submission* $\mathbf{x}_t$
2. *Construct online function* $f_t(\mathbf{x}) \triangleq \alpha_t \langle \mathbf{g}(\mathbf{x}_t), \mathbf{x}\rangle$

☐ O2B guarantees: ($\bar{\mathbf{x}}_t \triangleq \frac{1}{\alpha_{1:t}}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, $\alpha_{1:t} \triangleq \sum_{s=1}^t \alpha_s$)

$$\mathrm{GAP}_T \le \frac{1}{\alpha_{1:T}} \sum_{t\in[T]} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{x}^\star)\right) \triangleq \frac{1}{\alpha_{1:T}} \mathrm{REG}_T^{\boldsymbol{\alpha}}.$$

☐ Online Convex Optimization (OCO)

At each round $t = 1, 2, \ldots, T$:
- The learner submits $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$;
- The environments decide a convex loss function f_t;
- The learner suffers $f_t(\mathbf{x}_t)$ and receives gradient information.

Goal: minimize ***regret*** $\mathrm{REG}_T \triangleq \sum_{t\in[T]} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t\in[T]} f_t(\mathbf{x})$

Acceleration via Stabilized O2B Conversion and Gradient-Variation Online Adaptivity

Gradient-variation *regret helps maintain stability, and can be well controlled with a more* ***stabilized*** *O2B conversion.*

Leading to ***constant regret*** *even with more heavily weighted online gradients, thereby achieving acceleration.*

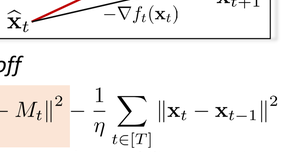
☐ Gradient-variation regret for smooth online functions:

$$\mathrm{REG}_T \le \mathcal{O}\left(\sqrt{V_T}\right), \textit{ where } V_T \triangleq \sum_{t=1}^T \sup_{\mathbf{x}\in\mathcal{X}} \|\nabla f_t(\mathbf{x}) - \nabla f_{t-1}(\mathbf{x})\|^2.$$

smaller than T in easier case!

Algorithm: ***Optimistic Online Gradient Descent (O-OGD)*** [Chiang et al., 2012]:

$$\mathbf{x}_t = \Pi_{\mathcal{X}}\left[\widehat{\mathbf{x}}_t - \eta M_t\right], \qquad \widehat{\mathbf{x}}_{t+1} = \Pi_{\mathcal{X}}\left[\widehat{\mathbf{x}}_t - \eta \nabla f_t(\mathbf{x}_t)\right],$$



Analysis: *a "bias-variance" trade-off*

$$\mathrm{REG}_T \lesssim \frac{1}{\eta} + \eta \sum_{t\in[T]} \|\nabla f_t(\mathbf{x}_t) - M_t\|^2 - \frac{1}{\eta}\sum_{t\in[T]} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$$

"bias" — *"variance": can be converted into gradient variations* — *stability negativity*

☐ Control the gradient variation with a stabilized O2B: [Cutkosky, 2019]

Key insight: ***query gradient on moving averages!***

1. *Passing to oracle the average* $\bar{\mathbf{x}}_t \triangleq \frac{1}{\alpha_{1:t}}\sum_{s=1}^t \alpha_s \mathbf{x}_s$
2. *Construct online function* $f_t(\mathbf{x}) \triangleq \alpha_t \langle \mathbf{g}(\bar{\mathbf{x}}_t), \mathbf{x}\rangle$

Guarantee: $\mathrm{GAP}_T \le \frac{1}{\alpha_{1:T}} \sum_{t\in[T]} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{x}^\star)\right)$

☐ Then we see the stabilization effect of gradient variations

Online gradient variation: Let $\tilde{\mathbf{x}}_t \triangleq \frac{1}{\alpha_{1:t}}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1}\right)$

$$\|\nabla f_t(\mathbf{x}_t) - M_t\|^2 = \alpha_t^2 \|\nabla \mathcal{L}(\bar{\mathbf{x}}_t) - \nabla \mathcal{L}(\tilde{\mathbf{x}}_t))\|^2$$
$$\le L^2 \alpha_t^2 \|\bar{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 \le L^2 \alpha_t^2 \frac{\alpha_t^2}{\alpha_{1:t}^2} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$$

cancelled out — ***stabilization***

O2B conversion:

$$\mathrm{GAP}_T \le \frac{\mathrm{REG}_T^{\boldsymbol{\alpha}}}{\alpha_{1:T}}, \Rightarrow \alpha_t = t,\ \alpha_{1:T} = \Omega(T^2),\ \mathrm{REG}_T^{\boldsymbol{\alpha}} \le \mathcal{O}(1) \Rightarrow \textbf{\textit{acceleration!}}$$

Key Reference:
[1] Yu Nesterov, Universal gradient methods for convex optimization problems, MP'15.
[2] Ashok Cutkosky, Anytime online-to-batch, optimism and acceleration, ICML'19.
[3] Kfir Levy, Online to offline conversions, universality and adaptive minibatch sizes, NIPS'17.

Universal Optimization

> [Nesterov, 2015] *They do not need to know in advance the actual level of (Hölder) smoothness of the objective function. ..... At the same time, for each particular problem class they automatically ensure the best possible rate of convergence.*

Setting	Convergence Rate
Convex	$\mathcal{O}(\frac{L}{T^2} + \frac{\sigma}{\sqrt{T}})$ for L-smooth; $\mathcal{O}(\frac{1}{\sqrt{T}})$ for Lipschitz [Kavis et al., 2019]
	$\mathcal{O}\left(\frac{L_\nu}{T^{(1+3\nu)/2}} + \frac{\sigma}{\sqrt{T}}\right)$ for (L_ν, ν)-Hölder smooth [Rodomanov et al., 2024]
	$\mathcal{O}\left(\frac{L_\nu}{T^{(1+3\nu)/2}} + \frac{\sigma}{\sqrt{T}}\right)$ for (L_ν, ν)-Hölder smooth **[Ours]**
λ-Strongly Convex	$\mathcal{O}\left(\exp(-\frac{T}{\kappa}) \cdot \frac{T}{\kappa}\right)$ for L-smooth and Lipschitz; $\widetilde{\mathcal{O}}(\frac{1}{\lambda T})$ for Lipschitz [Levy, 2017]
	$\mathcal{O}\left(\exp(-\frac{T}{6\sqrt{\kappa}})\right)$ for L-smooth and Lipschitz; $\widetilde{\mathcal{O}}(\frac{1}{\lambda T})$ for Lipschitz **[Ours]**

We first achieve ***universality in OL****, then* ***apply it to OPT***

Contribution I: Universality with Hölder Smoothness in OCO

Our regrets interpolate between the optimal guarantees in smooth and non-smooth regimes

Convex $\mathrm{REG}_T \le \mathcal{O}\left(\sqrt{V_T} + L_\nu T^{\frac{1-\nu}{2}}\right)$ — smooth ($\nu = 1$) $\mathcal{O}\left(\sqrt{V_T}\right)$; non-smooth ($\nu = 0$) $\mathcal{O}\left(\sqrt{T}\right)$

λ-S.C. $\mathrm{REG}_T \le \mathcal{O}\left(\frac{1}{\lambda}\log V_T + \frac{1}{\lambda}L_\nu^2 (\log T)^{\frac{1-\nu}{1+\nu}}\right)$ — smooth ($\nu = 1$) $\mathcal{O}\left(\frac{1}{\lambda}\log V_T\right)$; non-smooth ($\nu = 0$) $\mathcal{O}\left(\frac{1}{\lambda}\log T\right)$

☐ **Algorithm:** *O-OGD with non-increasing step sizes*
☐ **Analysis:** a "virtual clipping" discussion, e.g., [Kavis et al., 2019] — large ss.: small accumulated variance; small ss.: enough for a cancellation.

Contribution II: Universal Stochastic Convex OPT

Our gradient-variation online universality exhibits great usefulness, when applied to OPT via O2B

Stochastic Convex $\mathrm{GAP}_T \le \mathcal{O}\left(\frac{L_\nu}{T^{(1+3\nu)/2}} + \frac{\sigma}{\sqrt{T}}\right)$ — smooth ($\nu = 1$) $\mathcal{O}\left(1/T^2\right)$; non-smooth ($\nu = 0$) $\mathcal{O}\left(1/\sqrt{T}\right)$

(stochastic variance σ)

Contribution III: Universal Deterministic Strongly Convex OPT

For the first time, *we provide a universal method that*
- *achieves accelerated convergence in the smooth regime*
- *maintaining near-optimal convergence in the non-smooth one*

solving ***open problem*** *since* [Levy, 2017]

Deterministic λ-S.C. $\mathrm{GAP}_T \le \mathcal{O}\left(\frac{1}{\lambda}\min\left\{\exp\left(\frac{-T}{6\sqrt{\kappa}}\right), \frac{\log T}{T}\right\}\right)$

☐ ***Achieving universality in strongly convex case is more challenging****, and we address this by integrating a detection-based guess-and-check procedure*

Contribution IV: Parameter-free Deterministic Smooth and S.C. OPT

With only the Oracle budget *T* as an input $\mathrm{GAP}_T \le \mathcal{O}\left(\frac{1}{\lambda}\exp\left(\frac{-T}{(1+4\sqrt{2\kappa})\lceil 2\log_2 T\rceil}\right)\right)$

☐ *Apply the above algorithm and search the curvature*