Introduction to NeurIPS’25 (Spotlight) Paper:

Gradient-Variation Online Adaptivity for Accelerated Optimization with Hölder Smoothness

Yuheng Zhao, Yu-Hu Yan, Kfir Yehuda Levy, Peng Zhao

Advanced OPT 2025.12.26

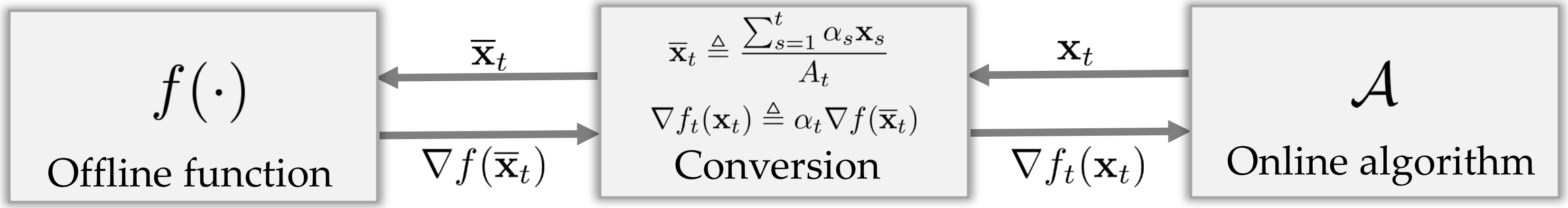
南京大學 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

TECHNION Israel Institute of Technology

Preview

- Reducing **offline opt.** as online opt. via (stabilized) online-to-batch conversion



(AOpt lec4&9)

- Convergence rate bounded by: $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \leq \frac{\mathrm{Reg}_T^{\boldsymbol{\alpha}}(\mathbf{x}^\star)}{A_T}.$ *weighted regret* / *sum of weights*

  - G-Lipschitz case: $\mathcal{O}\left(\frac{GD}{\sqrt{T}}\right)$ — *by OGD:* $\mathcal{O}(GD\sqrt{T})$ regret with $\alpha_t = 1, A_T = T$
  - L-Smooth case: $\mathcal{O}\left(\frac{LD^2}{T^2}\right)$ — *by **O**OGD:* $\mathcal{O}(LD^2)$ regret with $\alpha_t = t, A_T \approx T^2$
  - **Unknown case…**

    e.g., an interpolation between smoothness and non-smoothness

Target: *ONE (online) algorithm, adapt to an **unknown** level of smoothness*

This is called "universality" in offline optimization [Nesterov, 2015]

Preview

- We aim at **universality** by reducing **offline opt.** as **online opt.**

➢ G-Lipschitz case: $\mathcal{O}\left(\frac{GD}{\sqrt{T}}\right)$

➢ L-Smooth case: $\mathcal{O}\left(\frac{LD^2}{T^2}\right)$

by OGD: $\mathcal{O}(GD\sqrt{T})$ regret with $\alpha_t = 1, A_T = T$

by ***O****OGD:* $\mathcal{O}(LD^2)$ regret with $\alpha_t = t, A_T \approx T^2$

➢ **Unknown case…**

e.g., an interpolation between smoothness and non-smoothness

Target: ***ONE*** *(online) algorithm, adapt to an* ***unknown*** *level of smoothness*

This is called "universality" in offline optimization [Nesterov, 2015]

- A function class the algorithm will adapt to: **Hölder Smoothness**

➢ (L_ν, ν)-Hölder Smooth: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_\nu \|\mathbf{x} - \mathbf{y}\|^\nu$ ⇨ G-Lipschitz: $L_\nu = 2G, \nu = 0$; L-Smooth: $L_\nu = L, \nu = 1$

$L_\nu > 0, \nu \in [0, 1]$

➢ **Universal optimal rate:** $\mathcal{O}\left(\frac{L_\nu D^{1+\nu}}{T^{\frac{1+3\nu}{2}}}\right)$

by OCO: *What's the online algorithm and what's the regret bound?*

Contents

Review of AOpt-lec9: Acceleration

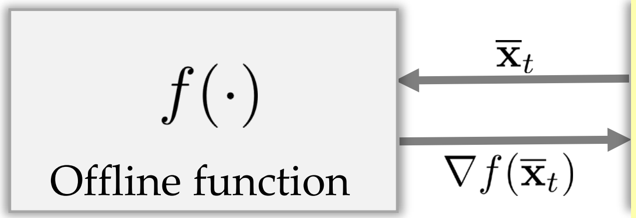
- Recall that ***accelerated*** rates can be achieved for smooth convex optimization using Nesterov's Accelerated GD, and also using ***OOGD w. O2B conversion***
- Stabilized Online-to-Batch Conversion [Cutkosky, 2019]

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a convex function with a convex and compact set $\mathcal{X}$. Then, for the following output with weighted average (regardless of how the $\{\mathbf{x}_t\}_{t=1}^T$ are generated): $\overline{\mathbf{x}}_t = \frac{1}{A_t}\sum_{s=1}^t \alpha_s \mathbf{x}_s$, with $A_t \triangleq \sum_{s=1}^t \alpha_s$ and $\alpha_t > 0$, we have the following online-to-batch conversion:*

$$f(\overline{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^T \langle \alpha_t \nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle}{A_T} \triangleq \frac{\mathrm{Reg}_T^{\boldsymbol{\alpha}}(\mathbf{x}^\star)}{A_T}.$$

weighted regret

sum of weights



Set weights $\alpha_t = t$ for all $t \in [T]$, then $A_T = \Theta(T^2)$.

We aim to use online algorithm ensuring $\mathcal{O}(1)$ regret.

Optimistic OGD with a suitable optimism design!

Review of AOpt-lec9: Acceleration

- Recall that ***accelerated*** rates can be achieved for smooth convex optimization using Nesterov's Accelerated GD, and also using ***OOGD w. O2B conversion***
- We achieve an *O(1)* regret using ***Optimistic OGD***

Optimistic online learning:

$$\nabla f_t(\mathbf{x}_t) = \alpha_t \nabla f(\bar{\mathbf{x}}_t), \quad M_t = \alpha_t \nabla f(\widetilde{\mathbf{x}}_t)$$

(with $\widetilde{\mathbf{x}}_t$ to be determined)

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

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

$$\Longrightarrow \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \frac{D^2}{2\eta} + \eta \sum_{t=1}^T \|\alpha_t \nabla f(\bar{\mathbf{x}}_t) - \alpha_t \nabla f(\widetilde{\mathbf{x}}_t)\|_2^2 - \frac{1}{4\eta}\sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$$

$$\text{(L-smoothness)} \quad \le \frac{D^2}{2\eta} + \eta \sum_{t=1}^T \alpha_t^2 L^2 \|\bar{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 - \frac{1}{4\eta}\sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$$

optimism design: approximate $\bar{\mathbf{x}}_t$ as possible as we can

Review of AOpt-lec9: Acceleration

- Recall that ***accelerated*** rates can be achieved for smooth convex optimization using Nesterov's Accelerated GD, and also using ***OOGD w. O2B conversion***
- We achieve an *O(1)* regret using ***Optimistic OGD***

Optimism design:

by def $\bar{\mathbf{x}}_t \triangleq \frac{1}{A_t}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_t\right)$,

we set $\widetilde{\mathbf{x}}_t \triangleq \frac{1}{A_t}\left(\sum_{s=1}^{t-1} \alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1}\right)$

$\Rightarrow$ $\overline{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{\alpha_t}{A_t}(\mathbf{x}_t - \mathbf{x}_{t-1})$

Stabilization effect via stabilized O2B cooperating with a suitable optimism

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \frac{\alpha_t^4 L^2}{A_t^2} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 - \frac{1}{4\eta} \sum_{t=1}^{T} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2$$

ensure that $\left(\frac{\eta \alpha_t^4 L^2}{A_t^2} - \frac{1}{4\eta}\right) \leq 0$ with $\alpha_t = t \implies \eta \leq \frac{1}{4L}$

$\Rightarrow$ Therefore, by setting $\eta = \frac{1}{4L}$, we have $\mathrm{Reg}_T^{\boldsymbol{\alpha}} \leq 2LD^2 = \mathcal{O}(1)$.

But not universal: *Require a prior knowledge of* ***smoothness parameter***

Universal Gradient-variation OL

- Motivation: Accelerated OPT via gradient-variation online learning

  $\Rightarrow$ Universal OPT via ***universal gradient-variation online learning***?

  i.e., adapts to an unknown level of smoothness

- We have learned gradient-variation OL in AOpt lec8:

Not universal!

Theorem 4 (Gradient Variation Regret Bound). *Assume that* $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ *and* $\eta_t = \min\{\frac{1}{4L}, \frac{D}{\sqrt{1+\widetilde{V}_{t-1}}}\}$ *and* $M_t =$ … *any comparator* $\mathbf{u} \in \mathcal{X}$ *is* … $_{-1})\|_2^2$ *is the empirical estimates of* V_t.

Proof. Finally, putting three terms together yields

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

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

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

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

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

Advanced Optimization (Fall 2025) Lecture 8. Adaptive Online Optimization 101

In fact, we do not need to do this clipping with L!

Universal Gradient-variation OL

- Motivation: Accelerated OPT via gradient-variation online learning

  ⇨ Universal OPT via ***universal gradient-variation online learning***?

  i.e., adapts to an unknown level of smoothness
- In OOGD, we do not need to do clipping with L!

Algorithm: OOGD with step size

$$\eta_t = \frac{D}{\sqrt{\sum_{s=1}^{t-1}\|\nabla f_s(\mathbf{x}_s) - M_s\|^2}}$$

⇨ Kavis et al. [2019]

*We can perform **"virtual clipping"** in analysis*

Consider the previous analysis: cancel when $\eta_t \leq \frac{1}{L}$

Proof. Finally, putting three terms together yields

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

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

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

Since step size is **non-increasing**, there ***exist*** some τ:

- $t > \tau$: step size is small enough for cancellation

$$\forall t > \tau, \quad \eta_t \leq \frac{1}{L}$$

- $t \leq \tau$: step size is still large, but…

$$\eta_\tau = \frac{D}{\sqrt{\sum_{s=1}^{\tau-1}\|\nabla f_s(\mathbf{x}_s) - M_s\|^2}} \geq \frac{1}{L}$$

$$\Rightarrow \sqrt{\sum_{s=1}^{\tau-1}\|\nabla f_s(\mathbf{x}_s) - M_s\|^2} \leq LD$$

Sum of empirical GV is small!

Universal Gradient-variation OL

- Motivation: Accelerated OPT via gradient-variation online learning

  ⇨ Universal OPT via ***universal gradient-variation online learning***?

  i.e., adapts to an unknown level of smoothness
- In OOGD, we do not need to do clipping with L!

Algorithm: OOGD with step size $\eta_t = \frac{D}{\sqrt{\sum_{s=1}^{t-1}\|\nabla f_s(\mathbf{x}_s) - M_s\|^2}}$

⇨ *We can perform **"virtual clipping"** in analysis* Kavis et al. [2019]

Since step size is **non-increasing**, there ***exist*** some τ:

- $t > \tau$: step size is small enough for cancellation

$$\forall t > \tau, \quad \eta_t \le \frac{1}{L}$$

- $t \le \tau$: step size is still large, but…

$$\sqrt{\sum_{s=1}^{\tau-1}\|\nabla f_s(\mathbf{x}_s) - M_s\|^2} \le LD$$

$$Reg_T \lesssim \sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t) - M_t\|_2^2 + \frac{D^2}{\eta_{T+1}} - \sum_{t=1}^{T}\frac{1}{2\eta_t}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2$$

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

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

$$LD^2 \qquad D\sqrt{V_T}$$

Similar analysis for Hölder Smoothness

Universal Gradient-variation OL

- Motivation: Accelerated OPT via gradient-variation online learning

  ⇨ Universal OPT via ***universal gradient-variation online learning***?

  i.e., adapts to an unknown level of smoothness
- Universal GV regret under Hölder smoothness

Key technique: regarding Hölder smoothness as ***smoothness with corruption*** [Devolder et al., 2014]

Lemma 1. *Suppose the function f is (L_ν, ν)-Hölder smooth. Then, for any $\delta > 0$, denoting by $L = \delta^{\frac{\nu-1}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$, it holds that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \le L^2 \|\mathbf{x} - \mathbf{y}\|^2 + 4L\delta. \tag{8}$$

Exist only in analysis (for final tuning)

The rest is the same as before… *(and a virtual clipping really helps because this L only exists in analysis)*

Algorithm: OOGD with step size $\eta_t = \frac{D}{\sqrt{\sum_{s=1}^{t-1} \|\nabla f_s(\mathbf{x}_s) - M_s\|^2}}$

$$Reg_T \lesssim D\sqrt{V_T} + LD^2 + D\sqrt{L\delta T}$$

additional corruption

$$\overset{\text{(tuning } \delta)}{=} \mathcal{O}\left(D\sqrt{V_T} + L_\nu D^{1+\nu} T^{\frac{1-\nu}{2}}\right)$$

- G-Lip.: $L_\nu = 2G, \nu = 0, \mathcal{O}(GD\sqrt{T})$
- L-smo.: $L_\nu = L, \nu = 1, \mathcal{O}(D\sqrt{V_T})$

We can apply this universality to offline optimization!

Our Results

Thanks!
Q&A

- Motivation: Accelerated OPT via gradient-variation online learning

  ⇨ Universal OPT via ***universal gradient-variation online learning***?

  i.e., adapts to an unknown level of smoothness

- Gradient-variation regret under Hölder smoothness
- Implications to offline OPT
- Take aways:
  - ✓ Accelerated optimization can be understood by gradient-variation OL
  - ✓ We can achieve universality in OL, then **apply it to OPT**
  - ✓ **More online adaptivity might be useful for OPT** (and maybe not only for universality)

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

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

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

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

(stochastic variance σ)

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

Reference

[1] Yurii 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.

[4] Ali Kavis, Kfir Y. Levy, Francis R. Bach, and Volkan Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. NeurIPS'19.

[5] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. MP'14.

[6] Advanced Optimization Lecture8&9. 2025.