Optimistic Online-to-Batch Conversions for Accelerated Convergence and Universality

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

Convex Smooth Optimization (CO)

TL; DR

1. **Online learning** is essential in stochastic optimization.
2. Online learning + **Online-to-batch** solves convex optimization problem.
3. **Optimism** is essential for acceleration in smooth optimization.
4. Optimism is **unnecessary** in online learning algorithm.

Problem Setup: Convex Optimization

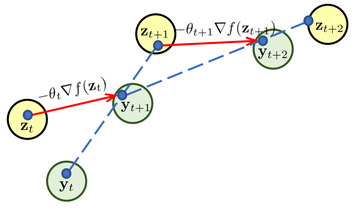
$$\min \quad f(\mathbf{x}) \qquad \text{s.t.} \quad \mathbf{x} \in \mathcal{X}$$

- $f(\cdot)$ is convex and L-smooth.

A fundamental problem in optimization and machine learning.

Optimal method: **Nesterov's Accelerated Gradient (NAG)**

$$\mathbf{y}_{t+1} = \mathbf{z}_t - \theta_t \nabla f(\mathbf{z}_t)$$
$$\mathbf{z}_{t+1} = \mathbf{y}_{t+1} + \beta_{t+1}(\mathbf{y}_{t+1} - \mathbf{y}_t)$$



	Convex	Strongly Convex
Lipschitz	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$	$\mathcal{O}\left(\frac{1}{T}\right)$
Smooth	$\mathcal{O}\left(\frac{1}{T^2}\right)$	$\mathcal{O}\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right)\right)$

Online-to-Batch Conversion & Online Learning

*Recent studies solve **CO** via Online-to-Batch (**O2B**) conversion and Online Learning (**OL**)*

☐ Why Online Learning (OL) is essential in CO?

OL is essential in **CO** mainly in the ***stochastic gradient*** setup.

- *Stochastic* optimization: there is a noisy gradient oracle $\mathbf{g}(\cdot)$

$$\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}] = \nabla f(\mathbf{x}), \quad \mathbb{E}[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \mid \mathbf{x}] \le \sigma^2$$

Example: **E**mpirical **R**isk **M**inimization (**ERM**)

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{x}; z_i)$$

⇩ *Mini-batching* (SGD, Adam, etc.) leads to **SO**

$$f_t(\mathbf{x}) = \frac{1}{|\mathcal{S}_t|}\sum_{z_i \in \mathcal{S}_t} \ell(\mathbf{x}; z_i)$$

- **Big data:** facing millions of samples (N is very large).
- Computing full gradient is almost ***impossible*** (e.g., limited memory of GPU facilities).

Note that $\mathcal{S}_t$ is still sampled from a fixed distribution (over all N samples)

- The optimization objective f_t varies across iterations (due to changing mini-batches).
- This renders the optimization process ***interactive*** — the algorithm continuously adapts in response to a new batch of samples.

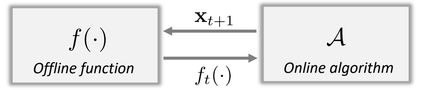
Online Learning: data comes as a ***stream***, and model ***online updates***

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*** $\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})$ *The learner's excess loss compared to the best fixed comparator in hindsight.*

☐ How Online Learning (OL) benefits CO?



Algorithm 1 Online-to-Batch (O2B) Conversion

Input: noisy gradient oracle $\mathbf{g}(\cdot)$, online learning algorithm $\mathcal{A}_{\text{OL}}$

1: **for** $t = 1, \ldots, T$ **do**
2: Obtain noisy gradient $\mathbf{g}(\mathbf{x}_t)$
3: Pass loss function $f_t(\cdot) \triangleq \langle \alpha_t \mathbf{g}(\mathbf{x}_t), \cdot \rangle$ to $\mathcal{A}_{\text{OL}}$
4: Receive next point $\mathbf{x}_{t+1}$ from $\mathcal{A}_{\text{OL}}$
5: **end for**
6: **return** $\bar{\mathbf{x}}_T = \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t \mathbf{x}_t$

regret of online learning

$$f\left(\frac{\sum_{t=1}^{T}\alpha_t \mathbf{x}_t}{\alpha_{1:T}}\right) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^{T}\langle \alpha_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star\rangle}{\alpha_{1:T}}$$

Optimal for **convex + Lipschitz** case:
$\alpha_t = 1 \Rightarrow f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \mathcal{O}(1/\sqrt{T})$

Optimal for **strongly convex + Lipschitz** case:
$\alpha_t = t \Rightarrow f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \mathcal{O}(1/T)$

Vanilla **O2B**-based methods ***cannot*** achieve the ***optimal*** accelerated convergence for the ***smooth*** objective.

Main Contribution: Optimistic O2B Conversion

Key message I: ***Optimism*** *is essential for acceleration in smooth case.*

Previous Progress: **Stabilized (Anytime)* O2B Conversion** [Cutkosky, ICML'19]

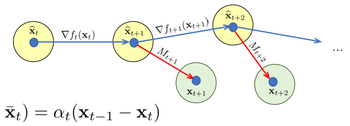
1: **for** $t = 1, \ldots, T$ **do**
2: Submit $\bar{\mathbf{x}}_t = \frac{\sum_{i=1}^{t}\alpha_i \mathbf{x}_i}{\alpha_{1:t}}$ $(\alpha_{1:t} \triangleq \sum_{i=1}^{t}\alpha_i)$
3: Receive feedback $\mathbf{g}_t = \nabla f(\bar{\mathbf{x}}_t)$
4: Send $\alpha_t \mathbf{g}_t$ to $\mathcal{A}$ and obtain $\mathbf{x}_{t+1}$
5: **end for**

The only algorithmic difference from vanilla O2B.

$$f\left(\frac{\sum_{t=1}^{T}\alpha_t \mathbf{x}_t}{\alpha_{1:T}}\right) - f(\mathbf{x}^\star) \le \frac{\sum_{t=1}^{T}\langle \alpha_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^\star\rangle}{\alpha_{1:T}}$$

Stabilized conversion + **optimistic OL** = optimal rate $\mathcal{O}(1/T^2)$

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\eta_t \langle \nabla \ell_{t-1}(\mathbf{x}_{t-1}), \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_t)\},$$
$$\widehat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\eta_t \langle \nabla \ell_t(\mathbf{x}_t), \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \widehat{\mathbf{x}}_t)\}.$$



Key Equations: $A_{t-1}(\bar{\mathbf{x}}_{t-1} - \bar{\mathbf{x}}_t) = \alpha_t(\bar{\mathbf{x}}_t - \mathbf{x}_t), A_t(\bar{\mathbf{x}}_t - \tilde{\mathbf{x}}_t) = \alpha_t(\mathbf{x}_{t-1} - \mathbf{x}_t)$

Key message II: *Optimism is* ***unnecessary*** *in* ***OL*** *algorithm.* [This work]

Algorithm 2 Optimistic Online-to-Batch Conversion

Input: Noisy gradient oracle $\mathbf{g}(\cdot)$, OL algorithm $\mathcal{A}_{\text{OL}}$, weights $\{\alpha_t\}_{t=1}^{T}$ with $\alpha_t > 0$.

1: **Initialize:** $\mathbf{x}_1 = \mathbf{x}_0 \in \mathcal{X}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: Query $\mathbf{g}(\tilde{\mathbf{x}}_{t+1})$ where $\tilde{\mathbf{x}}_{t+1} = \frac{1}{A_{t+1}}\left(\sum_{s=1}^{t}\alpha_s \mathbf{x}_s + \alpha_{t+1}\mathbf{x}_t\right)$
4: Define $f_{t+1}(\mathbf{x}) \triangleq \langle \alpha_{t+1}\mathbf{g}(\tilde{\mathbf{x}}_{t+1}), \mathbf{x}\rangle$ as the $(t+1)$-th round online function for $\mathcal{A}_{\text{OL}}$
5: Get $\mathbf{x}_{t+1}$ from $\mathcal{A}_{\text{OL}}(\mathbf{x}_1, \{f_s(\cdot)\}_{s=1}^{t+1})$
6: **end for**

** We call it "Stabilized" O2B to emphasize its stability property.*

Main Theorem: **Key Equation:** $A_{t-1}(\tilde{\mathbf{x}}_{t-1} - \tilde{\mathbf{x}}_t) = \alpha_t(\tilde{\mathbf{x}}_t - \mathbf{x}_{t-1})$

Theorem 1 (Main Result). *If the objective function $f(\cdot)$ is convex, then we have*

$$A_T\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star)\right] \le \sum_{t=1}^{T}\langle \alpha_t \nabla f(\tilde{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle + \sum_{t=1}^{T}\alpha_t\langle \nabla f(\bar{\mathbf{x}}_t) - \nabla f(\tilde{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_{t-1}\rangle,$$

where $\tilde{\mathbf{x}}_t \triangleq \frac{1}{A_t}\left(\sum_{s=1}^{t-1}\alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_{t-1}\right)$ and $\bar{\mathbf{x}}_t \triangleq \frac{1}{A_t}\left(\sum_{s=1}^{t-1}\alpha_s \mathbf{x}_s + \alpha_t \mathbf{x}_t\right)$.

- ***Look-ahead OL regret:*** Since $\tilde{\mathbf{x}}_{t+1} = \frac{1}{A_{t+1}}\left(\sum_{s=1}^{t}\alpha_s \mathbf{x}_s + \alpha_{t+1}\mathbf{x}_t\right)$, the $(t+1)$-th online function $f_{t+1}(\mathbf{x}) \triangleq \langle \alpha_{t+1}\mathbf{g}(\tilde{\mathbf{x}}_{t+1}), \mathbf{x}\rangle$ can be obtained in the t-th round.

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \eta_t \alpha_{t+1}\nabla f(\tilde{\mathbf{x}}_{t+1})\right] \Rightarrow \text{BLUE-TERM} \le \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2\eta} - \frac{1}{2\eta}\sum_{t=1}^{T}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$$

(simple OGD)

- ***Optimistic via O2B:*** Essential optimistic term for acceleration.

$$\text{RED-TERM} \le L\sum_{t=1}^{T}\alpha_t\|\bar{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = L\sum_{t=1}^{T}\frac{\alpha_t^2}{A_t}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

- ***Overall:***

$$\text{L.H.S.} \le \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2\eta} + \sum_{t=1}^{T}\left(\frac{L\alpha_t^2}{A_t} - \frac{1}{2\eta}\right)\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \le 2L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Previous solution:	***Stabilized O2B***	+	***Optimistic OL***	$\Big\}\ \mathcal{O}\left(\frac{1}{T^2}\right)$
Our solution:	***Optimistic O2B***	+	***OGD***	

Strongly Convexity & Universality

Strongly Convex Setup:

Theorem 2. *If the objective function $f(\cdot)$ is λ-strongly convex, then we have*

$$A_T\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star)\right] \le \sum_{t=1}^{T}\alpha_t\left[h_t(\mathbf{x}_t) - h_t(\mathbf{x}^\star)\right] + \sum_{t=1}^{T}\alpha_t\langle \nabla \hat{f}(\bar{\mathbf{x}}_t) - \nabla \hat{f}(\tilde{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_{t-1}\rangle,$$

where $\hat{f}(\cdot) \triangleq f(\cdot) - \frac{\lambda}{2}\|\cdot\|^2$ and $h_t(\cdot) \triangleq \langle \nabla \hat{f}(\tilde{\mathbf{x}}_t), \cdot\rangle + \frac{\lambda}{2}\|\cdot\|^2$ is a λ-strongly convex surrogate.

Algorithm: With $\alpha_1 = 1$ and $\alpha_t = \frac{1}{4\sqrt{\kappa}}A_{t-1}$ for $t \ge 2$,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{\lambda A_t}\left(\nabla h_t(\mathbf{x}_t) - M_t + M_{t+1}\right),\ M_t = \begin{cases}\alpha_1 \nabla \hat{f}(\tilde{\mathbf{x}}_1) + \alpha_1 \mathbf{x}_1, & t = 1\\ \alpha_t \nabla \hat{f}(\tilde{\mathbf{x}}_t) + \alpha_t \mathbf{x}_{t-1}, & t \ge 2\end{cases} \Rightarrow \mathcal{O}\left(\exp\left(-\frac{T}{\sqrt{\kappa}}\right)\right)$$

Interesting observation: ***Distributed*** *optimism between* ***OL*** *algorithm and* ***O2B*** *conversion.*

Smoothness-Universal Setup: **Key Equation:** $A_{t-1}(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{t-1}) = \alpha_t(\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_t) + \alpha_{t-1}(\mathbf{x}_{t-1} - \mathbf{x}_{t-2})$

Theorem 3. *If the objective $f(\cdot)$ is convex, when $\alpha_1 = 1$ and $\alpha_T = 0$, the final term of $A_T\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star)\right]$ can be bounded by*

$$\sum_{t=1}^{T}\langle \alpha_t \nabla f(\tilde{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}^\star\rangle + \sum_{t=1}^{T-1}\alpha_t\langle \nabla f(\tilde{\mathbf{x}}_{t+1}) - \nabla f(\tilde{\mathbf{x}}_t), \mathbf{x}_t - \mathbf{x}_{t-1}\rangle - \sum_{t=2}^{T}A_{t-1}\mathcal{D}_f(\tilde{\mathbf{x}}_{t-1}, \tilde{\mathbf{x}}_t),$$

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

Algorithm: With $\alpha_t = t$ for $t \in [T-1]$, $\alpha_T = 0$,

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \eta_t \alpha_{t+1}\nabla f(\tilde{\mathbf{x}}_{t+1})\right] \qquad \eta_t = \frac{D}{\sqrt{\sum_{s=1}^{t}\alpha_s^2\|\nabla f(\tilde{\mathbf{x}}_{s+1}) - \nabla f(\tilde{\mathbf{x}}_s)\|^2}}$$

one-gradient per round

Result: $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \left\{\mathcal{O}(LD^2/T^2) \text{ when smooth, and } \mathcal{O}(GD/\sqrt{T}) \text{ when non-smooth}\right\}$.