Online Learning in Complex Environments:
The Need for a Data-dependent Theory

Tim van Erven

University of Amsterdam

Game AI Workshop @ Maastricht

March 30, 2023

Online Convex Optimization

Parameters $\boldsymbol{\theta}$ take values in a convex domain $\Theta \subset \mathbb{R}^d$

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner estimates $\boldsymbol{\theta}_t \in \Theta$
3: Nature reveals convex loss function $f_t : \Theta \to \mathbb{R}$
4: **end for**

Goal: Predict almost as well as the best possible parameters $\boldsymbol{\theta}^*$:

$$\text{Regret}_T(\boldsymbol{\theta}^*) = \sum_{t=1}^{T} f_t(\boldsymbol{\theta}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{\theta}^*)$$

Online Convex Optimization

Parameters $\boldsymbol{\theta}$ take values in a convex domain $\Theta \subset \mathbb{R}^d$

1: **for** $t = 1, 2, \ldots, T$ **do**
2: Learner estimates $\boldsymbol{\theta}_t \in \Theta$
3: Nature reveals convex loss function $f_t : \Theta \to \mathbb{R}$
4: **end for**

Goal: Predict almost as well as the best possible parameters $\boldsymbol{\theta}^*$:

$$\text{Regret}_T(\boldsymbol{\theta}^*) = \sum_{t=1}^{T} f_t(\boldsymbol{\theta}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{\theta}^*)$$

Viewed as a **zero-sum game** against Nature:

$$V = \min_{\boldsymbol{\theta}_1} \max_{f_1} \min_{\boldsymbol{\theta}_2} \max_{f_2} \cdots \min_{\boldsymbol{\theta}_T} \max_{f_T} \max_{\boldsymbol{\theta}^* \in \Theta} \text{Regret}_T(\boldsymbol{\theta}^*)$$

Example: Electricity Forecasting

Typically, functions f_t determined by data:

- Every day t an electricity company needs to predict how much electricity Y_t is needed the next day
- Given feature vector $\boldsymbol{X}_t \in \mathbb{R}^d$, predict $\hat{Y}_t = \boldsymbol{X}_t^\intercal \boldsymbol{\theta}_t$ with a linear model
- Next day: observe Y_t
- Measure loss by $f_t(\boldsymbol{\theta}_t) = (Y_t - \hat{Y}_t)^2$ and improve parameter estimates: $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1}$

Online Gradient Descent

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_{t+1} &= \boldsymbol{\theta}_t - \eta_t \nabla f_t(\boldsymbol{\theta}_t) \\ \boldsymbol{\theta}_{t+1} &= \min_{\boldsymbol{\theta} \in \Theta} \|\tilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}\|\end{aligned}$$

Theorem (Zinkevich, 2003)

Suppose Θ compact with diameter at most D, and $\|\nabla f_t(\boldsymbol{\theta}_t)\| \leq G$. Then online gradient descent with $\eta_t = \frac{D}{G\sqrt{t}}$ guarantees

$$\text{Regret}_T(\boldsymbol{\theta}^*) \leq \frac{3}{2} GD\sqrt{T}$$

for **any** *choices of Nature.*

Without further assumptions, this is optimal (up to a constant factor).

OGD is Optimal, but is it Good?

Theorem (Lower Bound)

For any learning algorithm, there exists an OCO task with $\text{diam}(\Theta) \leq D$ *and* $\|\nabla f_t(\boldsymbol{\theta}_t)\| \leq G$ *such that*

$$\text{Regret}_T(\boldsymbol{\theta}^*) \geq cGD\sqrt{T}$$

for some absolute constant $c > 0$.

Proof:

- $\Theta = [-\frac{D}{2}, \frac{D}{2}]$
- $f_t(\theta) = \theta g_t$ with $\Pr(g_t = -G) = \Pr(g_t = +G) = 1/2$

Then for any algorithm

$$\mathbb{E}\Big[\sum_{t=1}^{T} f_t(\theta_t)\Big] = 0,$$

but

$$\mathbb{E}\Big[\min_{\theta^* \in \Theta} \sum_{t=1}^{T} f_t(\theta^*)\Big] = \frac{D}{2}\,\mathbb{E}\Big[\min\Big\{\sum_{t=1}^{T} g_t, -\sum_{t=1}^{T} g_t\Big\}\Big] \leq -cDG\sqrt{T}.$$

OGD is Optimal, but is it Good?

Theorem (Lower Bound)

For any learning algorithm, there exists an OCO task with $\text{diam}(\Theta) \leq D$ *and* $\|\nabla f_t(\boldsymbol{\theta}_t)\| \leq G$ *such that*

$$\text{Regret}_T(\boldsymbol{\theta}^*) \geq cGD\sqrt{T}$$

for some absolute constant $c > 0$.

Proof:

- $\Theta = [-\frac{D}{2}, \frac{D}{2}]$
- $f_t(\theta) = \theta g_t$ with $\Pr(g_t = -G) = \Pr(g_t = +G) = 1/2$

Hardest case:

- **Linear functions** f_t
- **Gradients** g_t are **pure noise**, with maximal variance
  $\rightarrow$ nothing interesting to learn, so irrelevant for applications

What if there is Less Noise?

Theorem (Sachs, Hadiji, Van Erven, Guzmán, 2023)

There exists an algorithm (optimistic follow-the-regularized-leader) that guarantees the worst-case bound

$$\text{Regret}_T(\boldsymbol{\theta}^*) = O(GD\sqrt{T})$$

and, if the f_t are i.i.d. and $F_t(\boldsymbol{\theta}) = \mathbb{E}[f_t(\boldsymbol{\theta})]$ is L-smooth, then

$$\mathbb{E}[\text{Regret}_T(\boldsymbol{\theta}^*)] = O(\sigma D\sqrt{T} + LD^2)$$

- $\sigma^2 = \max_{\boldsymbol{\theta}\in\Theta} \text{Var}(\nabla f_t(\boldsymbol{\theta}))$
- Exploits stochasticity and smoothness if available, but does **not assume** them
- Previously known for linear losses, i.e. $L = 0$ [Rakhlin, Sridharan, 2013]
- Recovers optimal rates in stochastic acceleration
  (via online-to-batch conversion on scaled losses)

What if there is Less Noise? (Refined Version)

Theorem (Sachs, Hadiji, Van Erven, Guzmán, 2023)

There exists an algorithm (optimistic follow-the-regularized-leader) that guarantees the worst-case bound

$$\text{Regret}_T(\boldsymbol{\theta}^*) = O(GD\sqrt{T})$$

and, if the f_t are stochastic and each $F_t(\boldsymbol{\theta}) = \mathbb{E}[f_t(\boldsymbol{\theta})|\mathcal{F}_{t-1}]$ is L-smooth, then

$$\mathbb{E}[\text{Regret}_T(\boldsymbol{\theta}^*)] = O((\bar{\sigma}_T + \bar{\Sigma}_T)D\sqrt{T} + LD^2)$$

- $\bar{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^T \max_{\boldsymbol{\theta}\in\Theta} \text{Var}(\nabla f_t(\boldsymbol{\theta}))$: average **variance** of the gradients
- $\bar{\Sigma}_T^2 = \frac{1}{T}\sum_{t=1}^T \max_{\boldsymbol{\theta}\in\Theta} \|\nabla F_t(\boldsymbol{\theta}) - \nabla F_{t-1}(\boldsymbol{\theta})\|^2$: average **drift** in the expected gradients
- **Interpolates** between i.i.d. and adversarial settings

Towards a Data-dependent Theory of Online Learning

Applications are not zero-sum games:

1. Worst-case regret witnessed on **fully random data**
   - Not relevant for practice!
2. **Nature is not trying to win**: e.g.
   - Consumers do not adjust electricity consumption to make statistical analysis hard

Can often adapt to some data- or distribution-dependent measure of
easiness of the data
to get much better performance!

Standard Textbook View of General OCO [Hazan, 2016]

Convex f_t	$\sqrt{T}$	Online Gradient Descent with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex f_t	$\ln T$	Online Gradient Descent wwith $\eta_t \propto \frac{1}{t}$
Exp-concave f_t	$d \ln T$	Online Newton Step with $\eta \propto 1$

Minimax rates based on curvature (bounded domain and gradients)

- **Strongly convex:** second derivative at least $\alpha > 0$, implies exp-concave
- **Exp-concave:** $e^{-\alpha \ell_t}$ concave
  Satisfied by log loss, logistic loss, squared loss, but not hinge loss

Standard Textbook View of General OCO [Hazan, 2016]

Convex f_t	$\sqrt{T}$	Online Gradient Descent with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex f_t	$\ln T$	Online Gradient Descent wwith $\eta_t \propto \frac{1}{t}$
Exp-concave f_t	$d \ln T$	Online Newton Step with $\eta \propto 1$

Minimax rates based on curvature (bounded domain and gradients)

Limitations:

- Different method in each case. (Requires sophisticated users.)
- Theoretical tuning of η_t **very conservative**
- What if curvature varies between rounds?
- In many applications data are **stochastic** (i.i.d.) Should be easier than worst case...

Standard Textbook View of General OCO [Hazan, 2016]

Convex f_t	$\sqrt{T}$	Online Gradient Descent with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex f_t	$\ln T$	Online Gradient Descent wwith $\eta_t \propto \frac{1}{t}$
Exp-concave f_t	$d \ln T$	Online Newton Step with $\eta \propto 1$

Minimax rates based on curvature (bounded domain and gradients)

Limitations:

- Different method in each case. (Requires sophisticated users.)
- Theoretical tuning of η_t **very conservative**
- What if curvature varies between rounds?
- In many applications data are **stochastic** (i.i.d.) Should be easier than worst case...

Need Adaptive Methods!

- Difficulty: All existing methods learn η at too slow rate [HP2005] so **overhead of learning best η ruins potential benefits**

Theorem (Van Erven, Koolen, 2016, Van Erven, Koolen, Van der Hoeven, 2021)

The MetaGrad algorithm guarantees the following **data-dependent** *bound:*

$$\text{Regret}_T(\boldsymbol{\theta}^*) \leq \sum_{t=1}^{T} (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^\intercal \nabla f_t(\boldsymbol{\theta}_t) \preccurlyeq \begin{cases} \sqrt{T \ln \ln T} \\ \\ \sqrt{V_T(\boldsymbol{\theta}^*)\, d \ln T} + d \ln T \end{cases}$$

where

$$V_T(\boldsymbol{\theta}^*) = \sum_{t=1}^{T} ((\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)^\intercal \nabla f_t(\boldsymbol{\theta}_t))^2.$$

Key Feature:

- Pay only $\ln \ln T$ for learning η

Consequences

1. Non-stochastic adaptation:

Convex f_t	$\sqrt{T \ln \ln T}$
Exp-concave f_t	$d \ln T$
Fixed convex $f_t = f$	$d \ln T$

Extension by [Wang, Lu, Zhang, 2020] also achieves $O(\ln T)$ for strongly convex losses

Consequences

1. Non-stochastic adaptation:

Convex f_t	$\sqrt{T \ln \ln T}$
Exp-concave f_t	$d \ln T$
Fixed convex $f_t = f$	$d \ln T$

Extension by [Wang, Lu, Zhang, 2020] also achieves $O(\ln T)$ for strongly convex losses

2. Stochastic without curvature [Koolen, Grünwald, Van Erven, 2016]:

Suppose f_t i.i.d. with stochastic optimum $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_f[f(\boldsymbol{\theta})]$.
Then expected regret $\mathbb{E}[\text{Regret}_T(\boldsymbol{\theta}^*)]$:

Absolute loss* $f_t(\theta) = \lvert\theta - X_t\rvert$ for d=1	$\ln T$
Hinge loss* $\max\{0, 1 - Y_t \langle \boldsymbol{\theta}, \boldsymbol{X}_t \rangle\}$	$d \ln T$
(B, β)-**Bernstein**	$(Bd \ln T)^{1/(2-\beta)}\ T^{(1-\beta)/(2-\beta)}$

*Conditions apply

MetaGrad Experiments [Van Erven, Koolen, Van der Hoeven, 2021]

- 17 benchmark **UCI data sets**: 11 classification, 6 regression
- Tune algorithms according to theory
  (**No secret hyperparameter optimization!**)
- Measure regret ratio between algorithm and Online Gradient Descent

Algorithm	Median Regret Ratio	Computation Time
AdaGrad	3.54	$O(dT)$
Online Gradient Descent	1.00	$O(dT)$
MetaGrad	**0.25**	$O(d^2 T)$

MetaGrad Experiments [Van Erven, Koolen, Van der Hoeven, 2021]

- 17 benchmark **UCI data sets**: 11 classification, 6 regression
- Tune algorithms according to theory
  (**No secret hyperparameter optimization!**)
- Measure regret ratio between algorithm and Online Gradient Descent

Algorithm	Median Regret Ratio	Computation Time
AdaGrad	3.54	$O(dT)$
Online Gradient Descent	1.00	$O(dT)$
MetaGrad	**0.25**	$O(d^2 T)$

But... MetaGrad **slow in high dimensions**. Fast approximations:

Coordinatewise	0.32	$O(dT)$
Sketching($m = 1$)	0.31	$O(mdT)$
Sketching($m = 10$)	0.27	$O(mdT)$
Sketching($m = 50$)	0.25	$O(mdT)$

Many Possible Sources of Easiness in Data

- Statistical: nature is (approximately) stationary
- Curvature: loss function is benign
- Game-theoretic: other players update slowly + smoothness
- Model selection: maybe a simple comparator $\boldsymbol{\theta}^*$ is optimal
- Structured comparator classes
- Smoothed analysis: data have smooth distribution
- Bandits: less exploration needed
- . . . ?

Food for Thought

Can organize by:

- Application domain: games, bandits, full information, market making, optimization, bilateral trade, ...
- Types of adaptivity: types of loss functions, easiness caused by statistical or deterministic regularity, adapting to hyperparameters like G or $\|\boldsymbol{\theta}^*\|$, ...
- Techniques: adaptive learning rates, optimistic gradient estimates, adaptive exploration for bandits, ...

Working group goals:

- List desirable types of adaptivity for various settings
- Organize them
- Prioritize
- Identify expertise, common interests and collaborations

Food for Thought

Can organize by:

- Application domain: games, bandits, full information, market making, optimization, bilateral trade, ...
- Types of adaptivity: types of loss functions, easiness caused by statistical or deterministic regularity, adapting to hyperparameters like G or $\|\boldsymbol{\theta}^*\|$, ...
- Techniques: adaptive learning rates, optimistic gradient estimates, adaptive exploration for bandits, ...

Working group goals:

- List desirable types of adaptivity for various settings
- Organize them
- Prioritize
- Identify expertise, common interests and collaborations

No silver bullet:

- The price of adaptivity: if overhead (computational/regret) too large, may not be worth it
- Some types of adaptivity may be mutually exclusive, e.g. G vs $\|\boldsymbol{\theta}^*\|$, or may be impossible in some settings.

References

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.

W. M. Koolen, P. Grünwald, and T. van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 4457–4465, 2016.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 993–1019, 2013.

T. van Erven and W. M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 3666–3674, 2016.

T. van Erven, W. M. Koolen, and D. van der Hoeven. Metagrad: Adaptation using multiple learning rates in online learning. *Journal of Machine Learning Research*, 22(161):1–61, 2021. URL `http://jmlr.org/papers/v22/20-1444.html`.