Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning

Wouter M. Koolen Peter Grünwald Tim van Erven

Goal

We want to make **efficient** online learning algorithms that **adapt** automatically to the complexity of the environment.

- Worst-case rates in adversarial environments (safe and robust)
- Fast rates in favorable stochastic environments (practice)

Key Observation

Modern adaptive algorithms bound **regret** in terms of **variance**

Friendliness of stochastic environments commonly quantified by condition relating **variance** to **regret**

Learning Model: Online Convex Optimization

In round $t = 1, 2, \ldots$

- Learner chooses $\boldsymbol{w}_t \in \mathcal{U} \subseteq \mathbb{R}^d$
- Environment selects convex loss function $\ell_t : \mathcal{U} \to \mathbb{R}$
- Learner incurs loss $\ell_t(\boldsymbol{w}_t)$ and observes gradient $\nabla \ell_t(\boldsymbol{w}_t)$

Goal: small **regret** $R_T^{\boldsymbol{u}}$ (or upper bound $\tilde{R}_T^{\boldsymbol{u}}$) w.r.t. **every** point $\boldsymbol{u}$

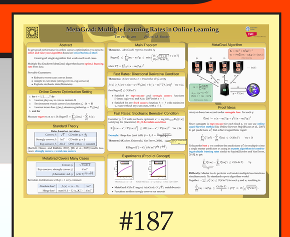
$$R_T^{\boldsymbol{u}} := \sum_{t=1}^T \left(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})\right), \qquad \tilde{R}_T^{\boldsymbol{u}} := \sum_{t=1}^T (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \nabla \ell_t(\boldsymbol{w}_t).$$

Second-order Regret Guarantees

$$\tilde{R}_T^f \leq \sqrt{V_T^f K_T^f} \quad \text{for all } f \in \mathcal{F} \tag{1}$$

Beats worst-case regret when $V_T^{f^*} = o(T)$ and $K_T^{f^*}$ small.

Two Examples



	Squint	MetaGrad
	Koolen and Van Erven [2015]	Van Erven and Koolen [2016]
Setting	Hedge Setting	Online Convex Optimization
$\mathcal{F}$	expert $k \in \{1, 2, \ldots\}$	$\boldsymbol{u} \in \mathcal{U}$
Loss	Linear $\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t$	Convex $\ell_t(\boldsymbol{w}_t)$
Cmplx.	$K_T^k = -\ln \pi(k)$	$K_T^{\boldsymbol{u}} = d \ln T$
Variance	$V_T^k = \sum_{t=1}^T (\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \ell_t^k)^2$	$V_T^{\boldsymbol{u}} = \sum_{t=1}^T \left((\boldsymbol{w}_t - \boldsymbol{u})^\intercal \nabla \ell_t(\boldsymbol{w}_t)\right)^2$
Time/rd.	$O(1)$ per expert	$O(d^2 \ln T)$ plus projection

First Step

Consider losses $\boldsymbol{\ell} \stackrel{\text{iid}}{\sim} \mathbb{P}$ with stochastic best expert $k^* = \arg\min_k \mathbb{E}[\ell^k]$ and **gap** $\min_{k \neq k^*} \mathbb{E}[\ell^k - \ell^{k^*}] > 0$. Then second-order bound (1) implies **constant regret** $\mathbb{E}[R_T^{k^*}] = O(1)$ [Gaillard et al., 2014].

Friendly Stochastic Environments

The **Bernstein condition** [Bartlett and Mendelson, 2006] says that **variance** of excess loss is small near stochastic optimum.

Bernstein condition key to fast rates in statistical learning.

Fix $B > 0$ and $\kappa \in [0, 1]$. We say

- $\boldsymbol{\ell} \sim \mathbb{P}$ are (B, κ)-Bernstein for stochastic experts if

$$\mathbb{E}\left[(\ell^k - \ell^{k^*})^2\right] \leq B\, \mathbb{E}\left[\ell^k - \ell^{k^*}\right]^\kappa \quad \forall k.$$

- $\ell \sim \mathbb{P}$ are (linearized) (B, κ)-Bernstein for stochastic OCO if

$$\mathbb{E}\left[\left((\boldsymbol{w} - \boldsymbol{u}^*)^\intercal \nabla \ell(\boldsymbol{w})\right)^2\right] \leq B\, \mathbb{E}\left[(\boldsymbol{w} - \boldsymbol{u}^*)^\intercal \nabla \ell(\boldsymbol{w})\right]^\kappa \quad \forall \boldsymbol{w}.$$

See paper for extensions beyond iid.

Main Theorem

*In any stochastic setting satisfying the (B, κ)-Bernstein condition, a second-order regret bound (1) implies **fast rates** both **in expectation**:*

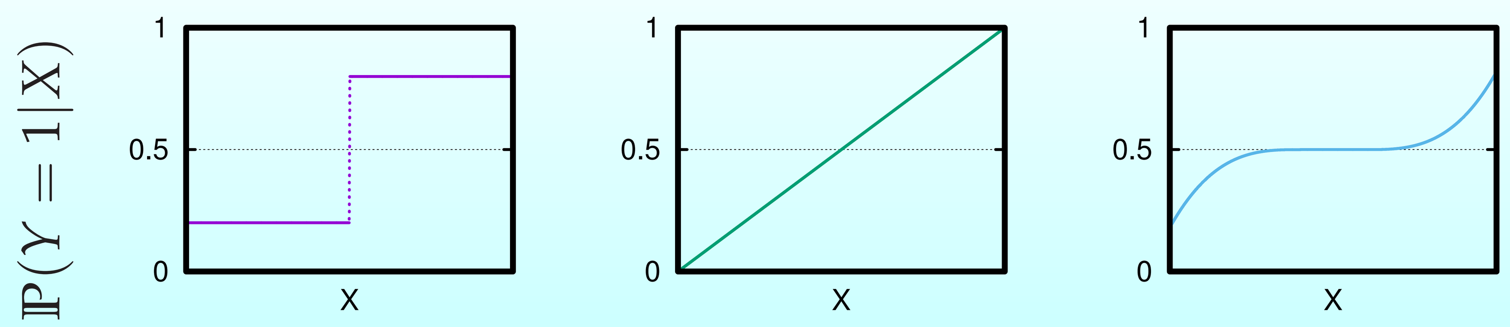
$$\mathbb{E}[R_T^{f^*}] = O\left(K_T^{\frac{1}{2-\kappa}} T^{\frac{1-\kappa}{2-\kappa}}\right),$$

*and **with high probability**: for any $\delta > 0$, with probability at least $1 - \delta$,*

$$R_T^{f^*} = O\left((K_T - \ln \delta)^{\frac{1}{2-\kappa}} T^{\frac{1-\kappa}{2-\kappa}}\right).$$

Inspiration: Tsybakov Margin Condition

Classification: $Y \in \{0, 1\}$. $\quad \mathbb{P}\left(\left|\mathbb{P}(Y = 1 | X) - 1/2\right| \leq t\right) \leq c t^\alpha$



easy	moderate	hard
$\alpha = \infty$	$\alpha = 1$	$\alpha = 0$

Confusing case: predictors with **equal risk** but **opposite predictions**.

Hinge Loss Example

Unregularized hinge loss on unit ball.

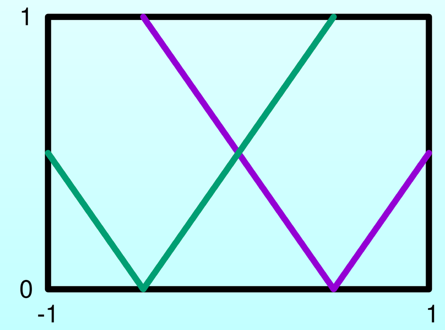
- Data $(\boldsymbol{x}_t, y_t) \sim \mathbb{P}$ i.i.d.
- Hinge loss $\ell_t(\boldsymbol{u}) = \max\{0, 1 - y_t \boldsymbol{x}_t^\intercal \boldsymbol{u}\}$.
- Mean $\boldsymbol{\mu} = \mathbb{E}[y\boldsymbol{x}]$ and second moment $\boldsymbol{D} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\intercal]$.
- Bernstein with $\kappa = 1$ and $B = \frac{2\lambda_{\max}(\boldsymbol{D})}{\|\boldsymbol{\mu}\|}$

Absolute Loss Example

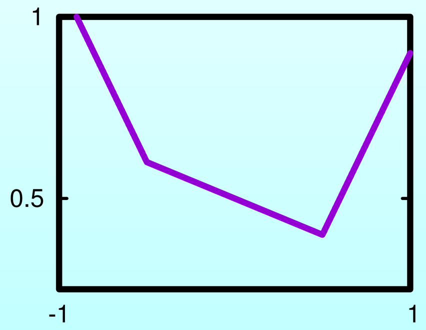
Absolute loss:

$$\ell_t(u) = |u - x_t|$$

where $x_t = \pm\frac{1}{2}$ i.i.d. with probabilities $2/5$ and $3/5$.



Individual loss functions



Long-term average loss

Bernstein with $\kappa = 1$ and $B = 5$.

Proof Ideas (OCO)

In-expectation for $\kappa = 1$: Consider $\ell \sim \mathbb{P}$ with stochastic optimum $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathcal{U}} \mathbb{E}\left[\ell(\boldsymbol{u})\right]$. The second-order regret bound (1) implies

$$\mathbb{E}\left[R_T^{\boldsymbol{u}^*}\right] \leq \mathbb{E}\left[\tilde{R}_T^{\boldsymbol{u}^*}\right] \leq \mathbb{E}\left[\sqrt{V_T^{\boldsymbol{u}^*} K_T^{\boldsymbol{u}^*}}\right] \leq \sqrt{\mathbb{E}\left[V_T^{\boldsymbol{u}^*}\right] K_T^{\boldsymbol{u}^*}}.$$

Let $x_t^{\boldsymbol{u}} := (\boldsymbol{u} - \boldsymbol{u}^*)^\intercal \nabla \ell_t(\boldsymbol{u})$ denote the excess linearzed loss of $\boldsymbol{u}$ in round t. The Bernstein condition for $\kappa = 1$ yields

$$\mathbb{E}\left[V_T^{\boldsymbol{u}^*}\right] = \sum_{t=1}^T \mathbb{E}\left[(x_t^{\boldsymbol{u}_t})^2\right] \leq B \sum_{t=1}^T \mathbb{E}\left[x_t^{\boldsymbol{u}_t}\right] = B\, \mathbb{E}\left[\tilde{R}_T^{\boldsymbol{u}^*}\right].$$

Combining the above two inequalities and solving for $\mathbb{E}\left[\tilde{R}_T^{\boldsymbol{u}^*}\right]$ gives

$$\mathbb{E}\left[R_T^{\boldsymbol{u}^*}\right] \leq B K_T^{\boldsymbol{u}^*}.$$

For $\kappa < 1$: linearize ($z^\kappa = \kappa^\kappa (1-\kappa)^{1-\kappa} \inf_{\epsilon > 0} \left\{\epsilon^{\kappa - 1} z + \epsilon^\kappa\right\}$ for $z \geq 0$) to show

$$c_1 \cdot \epsilon^{1-\kappa}\, \mathbb{E}\left[V_T^{\boldsymbol{u}^*}\right] \leq \mathbb{E}\left[\tilde{R}_T^{\boldsymbol{u}^*}\right] + c_2 \cdot T \cdot \epsilon.$$

High probability: requires **sophisticated martingale argument**.