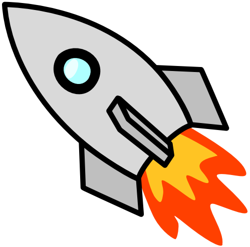
Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning



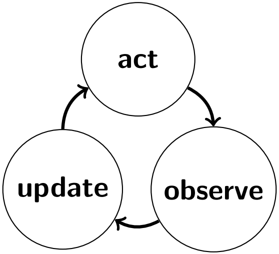
Wouter M. Koolen Peter Grünwald Tim van Erven

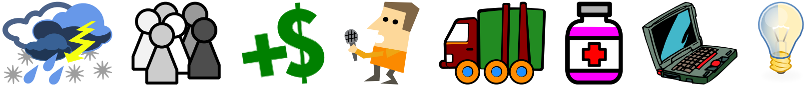
Centrum Wiskunde & Informatica

Universiteit Leiden

MIT, Thursday 30th June, 2016

Online Learning Challenges Everywhere





...

Easy Data

We desire 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)

Easy Data

We desire 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)

This talk

- Review second-order individual sequence bounds (Squint, MetaGrad)
- Review stochastic luckiness criteria (Gap, Tsybakov, Massart, Bernstein)
- Result: second-order algorithms exploit stochastic luckiness

Fundamental learning model: Hedge setting

- K experts

   ...

Fundamental learning model: Hedge setting

- K experts

$\dots$

- In round $t = 1, 2, \dots$
  - Learner plays distribution $\boldsymbol{w}_t = (w_t^1, \dots, w_t^K)$ on experts
  - Adversary reveals expert losses $\boldsymbol{\ell}_t = (\ell_t^1, \dots, \ell_t^K) \in [0, 1]^K$

  - Learner incurs loss $\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t$

Fundamental learning model: Hedge setting

- K experts

$\dots$

- In round $t = 1, 2, \dots$
  - Learner plays distribution $\boldsymbol{w}_t = (w_t^1, \dots, w_t^K)$ on experts
  - Adversary reveals expert losses $\boldsymbol{\ell}_t = (\ell_t^1, \dots, \ell_t^K) \in [0,1]^K$

  - Learner incurs loss $\boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t$

- The goal is to have small **regret**

$$R_T^k \;:=\; \underbrace{\sum_{t=1}^{T} \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t}_{\text{Learner}} - \underbrace{\sum_{t=1}^{T} \ell_t^k}_{\text{Expert } k}$$

with respect to every expert k.

Classical Result

The **Hedge** algorithm with **learning rate** η

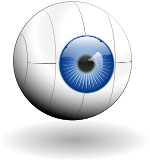
$$w_{t+1}^k \;:=\; \frac{e^{-\eta L_t^k}}{\sum_k e^{-\eta L_t^k}} \qquad \text{where} \qquad L_t^k \;=\; \sum_{s=1}^{t} \ell_s^k,$$

upon proper tuning of η ensures [Freund and Schapire, 1997]

$$R_T^k \;\prec\; \sqrt{T \ln K} \qquad \text{for each expert } k$$

which is tight for adversarial (worst-case) losses.

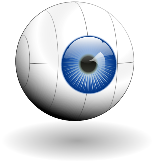
Squint [Koolen and Van Erven, 2015]



Notation For each expert k:

$r_t^k = \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \ell_t^k$	Instantaneous regret
$R_T^k = \sum_{t=1}^T r_t^k$	Cumulative regret
$V_T^k = \sum_{t=1}^T (r_t^k)^2$	Uncentered variance of the excess loss

Squint [Koolen and Van Erven, 2015]



Notation For each expert k:

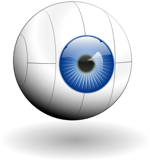
$$
\begin{array}{ll}
r_t^k = \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \ell_t^k & \text{Instantaneous regret} \\
R_T^k = \sum_{t=1}^T r_t^k & \text{Cumulative regret} \\
V_T^k = \sum_{t=1}^T (r_t^k)^2 & \text{Uncentered variance of the excess loss}
\end{array}
$$

Fix prior π on experts. After $T \geq 0$ rounds, Squint plays

$$
w_{T+1}^k \;\propto\; \pi(k) \int_0^{1/2} \exp\left(\eta R_T^k - \eta^2 V_T^k\right) \, \mathrm{d}\eta
$$

Constant time per expert per round.

Squint [Koolen and Van Erven, 2015]



Notation For each expert k:

$$
\begin{array}{ll}
r_t^k = \boldsymbol{w}_t^\intercal \boldsymbol{\ell}_t - \ell_t^k & \text{Instantaneous regret} \\
R_T^k = \sum_{t=1}^T r_t^k & \text{Cumulative regret} \\
V_T^k = \sum_{t=1}^T (r_t^k)^2 & \text{Uncentered variance of the excess loss}
\end{array}
$$

Fix prior π on experts. After $T \geq 0$ rounds, Squint plays

$$w_{T+1}^k \ \propto \ \pi(k) \int_0^{1/2} \exp\left(\eta R_T^k - \eta^2 V_T^k\right) \, \mathrm{d}\eta$$

Constant time per expert per round.

Squint ensures

$$R_T^k \ \prec \ \sqrt{V_T^k\left(-\ln \pi(k) + \ln \ln T\right)} \qquad \text{for each expert } k.$$

Beats worst-case regret when $V_T^k = o(\sqrt{T})$.

Fundamental Learning Model: Online Convex Optimization

- In round $t = 1, 2, \ldots$
  - Learner predicts $\boldsymbol{w}_t$ (from unit ball)
  - Encounter convex loss function $f_t(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R}$

  - Learner
    - observes gradient $\boldsymbol{g}_t \coloneqq \nabla f_t(\boldsymbol{w}_t)$ (from unit ball)
    - incurs loss $f_t(\boldsymbol{w}_t)$

Fundamental Learning Model: Online Convex Optimization

- In round $t = 1, 2, \ldots$
  - Learner predicts $\boldsymbol{w}_t$ (from unit ball)
  - Encounter convex loss function $f_t(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R}$


  - Learner
    - observes gradient $\boldsymbol{g}_t := \nabla f_t(\boldsymbol{w}_t)$ (from unit ball)
    - incurs loss $f_t(\boldsymbol{w}_t)$

- The goal is to have small **regret**

$$R_T^{\boldsymbol{u}} := \underbrace{\sum_{t=1}^T f_t(\boldsymbol{w}_t)}_{\text{Learner}} - \underbrace{\sum_{t=1}^T f_t(\boldsymbol{u})}_{\text{Point } \boldsymbol{u}}$$

with respect to **every** point $\boldsymbol{u}$.

Classical Result

Online gradient descent with learning rate η [Zinkevich, 2003]

$$\boldsymbol{w}_{t+1} \; = \; \boldsymbol{w}_t - \eta \boldsymbol{g}_t$$

recall $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{w}_t)$.

Classical Result

Online gradient descent with learning rate η [Zinkevich, 2003]

$$\boldsymbol{w}_{t+1} \;=\; \boldsymbol{w}_t - \eta \boldsymbol{g}_t$$

recall $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{w}_t)$.

After T rounds, properly tuned OGD guarantees

$$R_T^{\boldsymbol{u}} \;\le\; O\left(\sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2}\right) \;=\; O(\sqrt{T}) \qquad \text{for all } \boldsymbol{u} \text{ with } \|\boldsymbol{u}\| \le 1,$$

which is tight for adversarial losses.

MetaGrad [Koolen and Van Erven, 2016]

MetaGrad learns the learning rate η by aggregating $\ln T$ instances of Online Newton Step.

MetaGrad guarantees:

$$R_T^{\boldsymbol{u}} \;\leq\; O\left(\sqrt{V_T^{\boldsymbol{u}}\, d \ln T}\right) \quad \text{where} \quad V_T^{\boldsymbol{u}} \;:=\; \sum_{t=1}^{T} \big((\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t\big)^2$$

Run-time $O(d^2 \ln T)$ per round. (Sketching, diagonal version, . . .)

Improves OGD, for by Cauchy-Schwarz:

$$\big((\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t\big)^2 \;\leq\; \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 \|\boldsymbol{g}_t\|^2$$

Recap

We saw two algorithms with bounds of the form

$$R_T^{\boldsymbol{k}} \prec \sqrt{V_T^{\boldsymbol{k}} K_T^{\boldsymbol{k}}}$$

and

$$R_T^{\boldsymbol{u}} \prec \sqrt{V_T^{\boldsymbol{u}} K_T^{\boldsymbol{u}}}$$

But when/how can we guarantee that either V_T is small?

First step

Experts with **gap**. There are k^* and $\alpha > 0$ such that $\forall k \neq k^*$

$$\alpha \;\leq\; \mathbb{E}\left[\ell^{k} - \ell^{k^*}\right]$$

[Gaillard et al., 2014] show that any algorithm with second-order bound

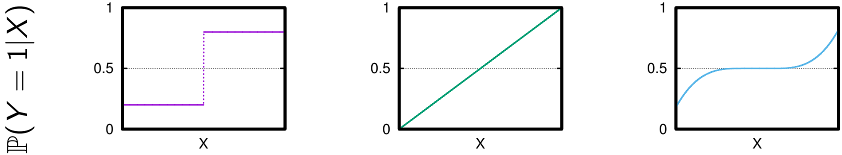
$$R_T^{k^*} \;\leq\; \sqrt{V_T^{k^*} K_T^{k^*}}.$$

satisfies $\mathbb{E}[R_T^{k^*}] = O(1)$.

Inspiration: Tsybakov margin condition for classification

Classification: $Y \in \{0, 1\}$.

$$\mathbb{P}\Big(\big|\mathbb{P}(Y = 1|X) - 1/2\big| \leq t\Big) \;\leq\; ct^{\alpha}$$



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

Confusing case: predictors with equal risk but opposite predictions.

Stochastic Luckiness Conditions

IID versions

- **Massart** condition, For $B > 0$ and $\forall k$:

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

- **Bernstein** condition. For $B > 0$, $\beta \in [0, 1]$ and $\forall k$:

$$\mathbb{E}\left[(\ell^k - \ell^{k^*})^2\right] \;\leq\; B\,\mathbb{E}\left[\ell^k - \ell^{k^*}\right]^\beta$$

Fast Rates using Massart

Applying the individual-sequence bound to k^* gives, in expectation,

$$\mathbb{E}\left[R_T^{k^*}\right] \prec \mathbb{E}\left[\sqrt{V_T^{k^*} K_T^{k^*}}\right] \overset{\text{Jensen}}{\leq} \sqrt{\mathbb{E}\left[V_T^{k^*}\right] K_T^{k^*}}$$

Fast Rates using Massart

Applying the individual-sequence bound to k^* gives, in expectation,

$$\mathbb{E}\left[R_T^{k^*}\right] \prec \mathbb{E}\left[\sqrt{V_T^{k^*} K_T^{k^*}}\right] \overset{\text{Jensen}}{\leq} \sqrt{\mathbb{E}\left[V_T^{k^*}\right] K_T^{k^*}}$$

Then

$$\begin{aligned}
\mathbb{E}\left[V_T^{k^*}\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\left(\sum_k w_t^k \ell_t^k - \ell_t^{k^*}\right)^2\right] \\
&\overset{\text{Jensen}}{\leq} \sum_{t=1}^{T} \mathbb{E} \sum_k w_t^k \, \mathbb{E}\left[\left(\ell_t^k - \ell_t^{k^*}\right)^2\right] \\
&\overset{\text{Massart}}{\leq} \sum_{t=1}^{T} \mathbb{E} \sum_k w_t^k B \, \mathbb{E}\left[\ell_t^k - \ell_t^{k^*}\right] = B \, \mathbb{E}\left[R_T^{k^*}\right]
\end{aligned}$$

Fast Rates using Massart

Applying the individual-sequence bound to k^* gives, in expectation,

$$\mathbb{E}\left[R_T^{k^*}\right] \prec \mathbb{E}\left[\sqrt{V_T^{k^*} K_T^{k^*}}\right] \overset{\text{Jensen}}{\leq} \sqrt{\mathbb{E}\left[V_T^{k^*}\right] K_T^{k^*}}$$

Then

$$\begin{aligned}
\mathbb{E}\left[V_T^{k^*}\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\left(\sum_k w_t^k \ell_t^k - \ell_t^{k^*}\right)^2\right] \\
&\overset{\text{Jensen}}{\leq} \sum_{t=1}^{T} \mathbb{E} \sum_k w_t^k \, \mathbb{E}\left[\left(\ell_t^k - \ell_t^{k^*}\right)^2\right] \\
&\overset{\text{Massart}}{\leq} \sum_{t=1}^{T} \mathbb{E} \sum_k w_t^k B \, \mathbb{E}\left[\ell_t^k - \ell_t^{k^*}\right] = B \, \mathbb{E}\left[R_T^{k^*}\right]
\end{aligned}$$

and so $\mathbb{E}[R_T^{k^*}] \prec \sqrt{B \, \mathbb{E}[R_T^{k^*}] K_T^{k^*}}$, hence $\mathbb{E}[R_T^{k^*}] \prec B K_T^{k^*} = O(1)$.

Bernstein for OCO

For experts we looked at

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

Bernstein for OCO

For experts we looked at

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

For stochastic OCO (with $f \sim \mathbb{P}$) we ask

$$\mathbb{E}\left[\langle \boldsymbol{w} - \boldsymbol{u}^*, \nabla f(\boldsymbol{w})\rangle^2\right] \;\le\; B\,\mathbb{E}\left[\langle \boldsymbol{w} - \boldsymbol{u}^*, \nabla f(\boldsymbol{w})\rangle\right]^\beta \quad \forall \boldsymbol{w}.$$

Examples where Bernstein applies, 1/2

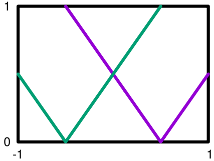
- Unregularized hinge loss on unit ball.
  - Data $(\boldsymbol{x}_t, y_t) \sim \mathbb{P}$ i.i.d.
  - Hinge loss $f_t(\boldsymbol{u}) = \max\{0, 1 - y_t \boldsymbol{x}_t^\intercal \boldsymbol{u}\}$.
  - Mean $\boldsymbol{\mu} = \mathbb{E}[y\boldsymbol{x}]$ and second moment $\boldsymbol{D} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\intercal]$.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

Examples where Bernstein applies, 2/2

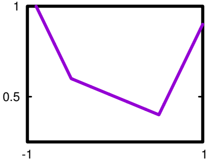
- Absolute loss:

$$f_t(u) \;=\; |u - x_t|$$

where $x_t = \pm\frac{1}{2}$ i.i.d. with probability 0.4 and 0.6.



Individual functions



Long-term average

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

Main result

Theorem

In any stochastic setting satisfying the (B, β)-Bernstein Condition, the guarantees for Squint and for MetaGrad

$$R_T^\theta \;\leq\; \sqrt{V_T^\theta K_T^\theta} \quad \textit{for all } \theta \in \Theta$$

imply fast rates for the respective algorithms both in expectation and with high probability. That is,

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

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

$$R_T^{\theta^*} \;=\; O\left((K_T - \ln \delta)^{\frac{1}{2-\beta}}\, T^{\frac{1-\beta}{2-\beta}}\right).$$

High probability, sketch 1/4

Fix $x^\theta \in [-1,1]$ and $\theta \in \Theta$. Bernstein

$$\mathbb{E}\left[(x^\theta)^2\right] \;\leq\; B\,\mathbb{E}\left[x^\theta\right]^\beta \qquad \text{for all } \theta \in \Theta$$

implies the Central condition [Van Erven et al., 2015]

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{-\eta x^\theta}\right] \;\leq\; O(\eta^{\frac{1}{1-\beta}}) \quad \text{for all } \eta \geq 0$$

High probability, sketch 2/4

We show

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{-\eta x^\theta}\right] \quad \le \quad O(\eta^{\frac{1}{1-\beta}}) \quad \text{for all } \eta \ge 0$$

implies (for $c \approx \frac{1}{2}$)

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{c\eta^2 (x^\theta)^2 - \eta x^\theta}\right] \quad \le \quad O(\eta^{\frac{1}{1-\beta}}) \quad \text{for all } \eta \ge 0$$

Telescope to

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{\sum_{t=1}^T c\eta^2 (x^\theta)^2 - \eta x^\theta}\right] \quad \le \quad O(T\eta^{\frac{1}{1-\beta}}) \quad \text{for all } \eta \ge 0$$

High probability, sketch 3/4

Combining

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{c\eta^2 V_T^\theta - \eta R_T^\theta}\right] \;\leq\; O\big(T\eta^{\frac{1}{1-\beta}}\big) \quad \text{for all } \eta \geq 0$$

with the individual sequence regret bound

$$R_T^\theta \;\leq\; 2\sqrt{V_T^\theta K_T^\theta} \;=\; \inf_\eta \left\{\eta V_T^\theta + \frac{K_T^\theta}{\eta}\right\}$$

so that

$$2\eta R_T^\theta \;\leq\; \frac{\eta^2}{2} V_T^\theta + 8 K_T^\theta$$

gives (using $c \approx 1/2$)

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{\eta R_T^\theta - 8K_T^\theta}\right] \;\leq\; O\big(T\eta^{\frac{1}{1-\beta}}\big) \quad \text{for all } \eta \geq 0$$

High probability, sketch 4/4

By Markov

$$\frac{1}{\eta} \ln \mathbb{E}\left[e^{\eta R_T^\theta - 8K_T^\theta}\right] \;\leq\; O(T\eta^{\frac{1}{1-\beta}}) \quad \text{for all } \eta \geq 0$$

implies with high probability

$$\eta R_T^\theta \;\leq\; 8K_T^\theta + T\eta^{\frac{1}{1-\beta}}$$

and optimally tuning η results in

$$R_T^\theta \;\leq\; O\left(K_T^{\frac{1}{2-\beta}}\, T^{\frac{1-\beta}{2-\beta}}\right).$$

Conclusion

We showed that Squint and MetaGrad (online learning algorithms with second-order bounds) adapt to Bernstein stochastic luckiness.

The results extend

- Non-iid. Only need the Bernstein condition **conditionally**. There are k^*, $B > 0$ and $\beta \in [0, 1]$ such that

$$\mathbb{E}\left[(\ell_t^k - \ell_t^{k^*})^2 \middle| \text{past}\right] \;\leq\; B\,\mathbb{E}\left[\ell_t^k - \ell_t^{k^*} \middle| \text{past}\right]^\beta \qquad \forall k \forall t.$$

  E.g. algorithmic information theory setting.

Thank you!