Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning
Online Learning Challenges Everywhere

act

update

observe

...
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)
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

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

Learner plays distribution $w^t = (w^t_1, \ldots, w^t_K)$ on experts

Adversary reveals expert losses $\ell^t = (\ell^t_1, \ldots, \ell^t_K) \in [0, 1]^K$

Learner incurs loss $w^t \ell^t$

The goal is to have small regret $R_k^T := T \sum_{t=1}^T w^t \ell^t - T \sum_{t=1}^T \ell^t_k$ with respect to every expert $k$. 
Fundamental learning model: Hedge setting

- $K$ experts

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

- Learner incurs loss $w_t^T \ell_t$
Fundamental learning model: Hedge setting

- $K$ experts

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

- Learner incurs loss $\mathbf{w}_t^T \mathbf{\ell}_t$

- The goal is to have small regret

\[
R^k_T := \sum_{t=1}^T \mathbf{w}_t^T \mathbf{\ell}_t - \sum_{t=1}^T \ell^k_t
\]

with respect to every expert $k$. 
The **Hedge** algorithm with **learning rate** $\eta$

$$w^k_{t+1} := \frac{e^{-\eta L^k_t}}{\sum_k e^{-\eta L^k_t}}$$  

where  

$$L^k_t = \sum_{s=1}^t \ell^k_s,$$

upon proper tuning of $\eta$ ensures [Freund and Schapire, 1997]  

$$R^k_T \prec \sqrt{T \ln K}$$  

for each expert $k$

which is tight for adversarial (worst-case) losses.
Squint [Koolen and Van Erven, 2015]

Notation  For each expert $k$:

$$r_t^k = w_t^T \ell_t - \ell_t^k$$  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 $k$:

\[
\begin{align*}
    r^k_t &= w^T_t \ell_t - \ell^k_t && \text{Instantaneous regret} \\
    R^k_T &= \sum_{t=1}^T r^k_t && \text{Cumulative regret} \\
    V^k_T &= \sum_{t=1}^T (r^k_t)^2 && \text{Uncentered variance of the excess loss}
\end{align*}
\]

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

\[
w^k_{T+1} \propto \pi(k) \int_0^{1/2} \exp \left( \eta R^k_T - \eta^2 V^k_T \right) \, d\eta
\]

Constant time per expert per round.
Squint [Koolen and Van Erven, 2015]

**Notation**  For each expert $k$:

- $r_t^k = w_t^T \ell_t - \ell_t^k$  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

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

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

Constant time per expert per round.

Squint ensures

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

Beats worst-case regret when $V_T^k = o(\sqrt{T})$. 
In round $t = 1, 2, \ldots$
- Learner predicts $w_t$ (from unit ball)
- Encounter convex loss function $f_t(u) : \mathbb{R}^d \rightarrow \mathbb{R}$

Learner
- observes gradient $g_t := \nabla f_t(w_t)$ (from unit ball)
- incurs loss $f_t(w_t)$
In round $t = 1, 2, \ldots$
- Learner predicts $w_t$ (from unit ball)
- Encounter convex loss function $f_t(u) : \mathbb{R}^d \to \mathbb{R}$

Learner
- observes gradient $g_t := \nabla f_t(w_t)$ (from unit ball)
- incurs loss $f_t(w_t)$

The goal is to have small regret

$$R_T^u := \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u)$$

with respect to every point $u$. 
Online gradient descent with learning rate $\eta$ [Zinkevich, 2003]

$$w_{t+1} = w_t - \eta g_t$$

recall $g_t = \nabla f_t(w_t)$. 
Classical Result

**Online gradient descent** with learning rate $\eta$ [Zinkevich, 2003]

$$w_{t+1} = w_t - \eta g_t$$

recall $g_t = \nabla f_t(w_t)$.

After $T$ rounds, properly tuned OGD guarantees

$$R^u_T \leq O \left( \sqrt{\sum_{t=1}^{T} \|g_t\|^2} \right) = O(\sqrt{T}) \quad \text{for all } u \text{ with } \|u\| \leq 1,$$

which is tight for adversarial losses.
MetaGrad [Koolen and Van Erven, 2016]

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

MetaGrad guarantees:

$$R_T^u \leq O\left(\sqrt{V_T^u d \ln T}\right)$$

where

$$V_T^u := \sum_{t=1}^{T}((w_t - u)^T g_t)^2$$

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

Improves OGD, for by Cauchy-Schwarz:

$$((w_t - u)^T g_t)^2 \leq \|w_t - u\|^2 \|g_t\|^2$$
Recap

We saw two algorithms with bounds of the form

\[ R^k_T \prec \sqrt{V^k_T K^k_T} \]

and

\[ R^u_T \prec \sqrt{V^u_T K^u_T} \]

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

$$R_{k^*}^{k^*} \leq \sqrt{V_{k^*} K_{k^*}}.$$ 

satisfies $\mathbb{E}[R_{k^*}^{k^*}] = O(1).$
Classification: \( Y \in \{0, 1\} \).

\[
P\left(\left| P(Y = 1 \mid X) - 1/2 \right| \leq t \right) \leq ct^\alpha
\]

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,

$$ E \left[ R_{k^*}^T \right] < E \left[ \sqrt{V_{k^*}^T K_{k^*}^T} \right] \overset{\text{Jensen}}{\leq} \sqrt{E \left[ V_{k^*}^T \right]} K_{k^*}^T $$

and so

$$ E \left[ R_{k^*}^T \right] < B E \left[ R_{k^*}^T \right] K_{k^*}^T, \text{ hence } E \left[ R_{k^*}^T \right] \leq B K_{k^*}^T = O(1). $$
Fast Rates using Massart

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

$$
\mathbb{E} \left[ R_{T}^{k^*} \right] \leq \mathbb{E} \left[ \sqrt{\mathbb{V}_{T}^{k^*} K_{T}^{k^*}} \right] \leq \sqrt{\mathbb{E} \left[ \mathbb{V}_{T}^{k^*} \right] K_{T}^{k^*}}
$$

Then

$$
\mathbb{E} \left[ \mathbb{V}_{T}^{k^*} \right] = \sum_{t=1}^{T} \mathbb{E} \left[ \left( \sum_{k} w_{t}^{k} \ell_{t}^{k} - \ell_{t}^{k^*} \right)^2 \right]
$$

$$
\leq \sum_{t=1}^{T} \mathbb{E} \sum_{k} w_{t}^{k} \mathbb{E} \left[ (\ell_{t}^{k} - \ell_{t}^{k^*})^2 \right]
$$

$$
\leq \sum_{t=1}^{T} \mathbb{E} \sum_{k} w_{t}^{k} B \mathbb{E} \left[ \ell_{t}^{k} - \ell_{t}^{k^*} \right] = B \mathbb{E} \left[ R_{T}^{k^*} \right]
$$
Fast Rates using Massart

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

$$
\mathbb{E}
\left[
R_{T}^{k^*}
\right]
\prec
\mathbb{E}
\left[
\sqrt{V_{T}^{k^*} K_{T}^{k^*}}
\right]
\leq
\sqrt{\mathbb{E}
\left[
V_{T}^{k^*}
\right]K_{T}^{k^*}}
$$

Then

$$
\mathbb{E}
\left[
V_{T}^{k^*}
\right]
= 
\sum_{t=1}^{T}
\mathbb{E}
\left[
\left(
\sum_{k} w_{t}^{k} \ell_{t}^{k} - \ell_{t}^{k^*}
\right)^{2}
\right]
\leq 
\sum_{t=1}^{T}
\mathbb{E}
\sum_{k} w_{t}^{k}
\mathbb{E}
\left[
\left(
\ell_{t}^{k} - \ell_{t}^{k^*}
\right)^{2}
\right]
\leq 
\sum_{t=1}^{T}
\mathbb{E}
\sum_{k} w_{t}^{k} B
\mathbb{E}
\left[
\ell_{t}^{k} - \ell_{t}^{k^*}
\right]
= 
B \mathbb{E}
\left[
R_{T}^{k^*}
\right]
$$

and so 

$$
\mathbb{E}[R_{T}^{k^*}] \prec \sqrt{B \mathbb{E}[R_{T}^{k^*}] K_{T}^{k^*}}
$$

hence

$$
\mathbb{E}[R_{T}^{k^*}] \prec BK_{T}^{k^*} = O(1).
$$
Bernstein for OCO

For experts we looked at

$$\mathbb{E} \left[ (\ell^k - \ell^k)^2 \right] \leq B \mathbb{E} \left[ \ell^k - \ell^k \right]^\beta \quad \forall k.$$
Bernstein for OCO

For experts we looked at

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

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

$$\mathbb{E} \left[ \langle w - u^*, \nabla f(w) \rangle^2 \right] \leq B \mathbb{E} \left[ \langle w - u^*, \nabla f(w) \rangle \right]^\beta \quad \forall w.$$
Examples where Bernstein applies, 1/2

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

  ![Graphs of individual functions and long-term average](image-url)

  **Individual functions**

  **Long-term average**

  Bernstein with \( \beta = 1 \) and \( B = 5 \).
Main result

Theorem

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

\[
R^\theta_T \leq \sqrt{V^\theta_T K^\theta_T} \quad \text{for all } \theta \in \Theta
\]

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

\[
\mathbb{E}[R^\theta_T] = O\left(K_T^{\frac{1}{2 - \beta}} T^{\frac{1 - \beta}{2 - \beta}}\right),
\]

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

\[
R^\theta_T = O\left((K_T - \ln \delta)^{\frac{1}{2 - \beta}} T^{\frac{1 - \beta}{2 - \beta}}\right).
\]
Fix $x^\theta \in [-1, 1]$ and $\theta \in \Theta$. Bernstein

$$\mathbb{E} \left[ (x^\theta)^2 \right] \leq B \mathbb{E} \left[ x^\theta \right]^\beta$$

for all $\theta \in \Theta$

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

$$\frac{1}{\eta} \ln \mathbb{E} \left[ e^{-\eta x^\theta} \right] \leq O(\eta^{\frac{1}{1-\beta}})$$

for all $\eta \geq 0$
We show

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

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

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

Telescope to

$$\frac{1}{\eta} \ln \mathbb{E} \left[ e^{\sum_{t=1}^{T} c\eta^2(x^\theta)^2 - \eta x^\theta} \right] \leq O(T\eta^{\frac{1}{1-\beta}}) \quad \text{for all } \eta \geq 0$$
Combining
\[ \frac{1}{\eta} \ln \mathbb{E} \left[ e^{\eta^2 V^\theta_T - \eta R^\theta_T} \right] \leq O(T \eta^{1-\beta}) \text{ for all } \eta \geq 0 \]
with the individual sequence regret bound
\[ R^\theta_T \leq 2 \sqrt{V^\theta_T K^\theta_T} = \inf \left\{ \eta V^\theta_T + \frac{K^\theta_T}{\eta} \right\} \]
so that
\[ 2\eta R^\theta_T \leq \frac{\eta^2}{2} V^\theta_T + 8K^\theta_T \]
gives (using \( c \approx 1/2 \))
\[ \frac{1}{\eta} \ln \mathbb{E} \left[ e^{\eta R^\theta_T - 8K^\theta_T} \right] \leq O(T \eta^{1-\beta}) \text{ for all } \eta \geq 0 \]
By Markov

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

implies with high probability

\[
\eta R_T^\theta \leq 8K_T^\theta + T \eta^{1-\beta}
\]

and optimally tuning $\eta$ results in

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

  \[
  \mathbb{E} \left[ (\ell^k_t - \ell^{k*}_t)^2 \right| \text{past} \right] \leq B \mathbb{E} \left[ \ell^k_t - \ell^{k*}_t \right| \text{past} \]^{\beta} \quad \forall k \forall t.
  
  E.g. algorithmic information theory setting.
Thank you!