Second-order Quantile Methods

Wouter M. Koolen    Tim van Erven

Kyushu University, Monday 3\textsuperscript{rd} October, 2016
Focus on expert setting

Online sequential prediction with expert advice

Core instance of advanced online learning tasks
- Bandits
- Combinatorial & matrix prediction
- Online convex optimization
- Boosting
- ...
Beyond the Worst Case

Two reasons data is often easier in practice:

- Data complexity
  - Stochastic data (gap)
  - Low noise
  - Low variance

- Second-order

- Model complexity
  - Simple model is good
  - Multiple good models

- Quantiles

- Any combination
Beyond the Worst Case

Two reasons data is often *easier* in practice:

<table>
<thead>
<tr>
<th>Data complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>► Stochastic data (gap)</td>
</tr>
<tr>
<td>► Low noise</td>
</tr>
<tr>
<td>► Low variance</td>
</tr>
</tbody>
</table>

Second-order & Quantiles

▶ Any combination
Beyond the Worst Case

Two reasons data is often easier in practice:

<table>
<thead>
<tr>
<th>Data complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Stochastic data (gap)</td>
</tr>
<tr>
<td>▶ Low noise</td>
</tr>
<tr>
<td>▶ Low variance</td>
</tr>
</tbody>
</table>

second-order
Beyond the Worst Case

Two reasons data is often easier in practice:

<table>
<thead>
<tr>
<th>Data complexity</th>
<th>Model complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Stochastic data (gap)</td>
<td>▶ Simple model is good</td>
</tr>
<tr>
<td>▶ Low noise</td>
<td>▶ Multiple good models</td>
</tr>
<tr>
<td>▶ Low variance</td>
<td></td>
</tr>
</tbody>
</table>

second-order
Beyond the Worst Case

Two reasons data is often easier in practice:

<table>
<thead>
<tr>
<th>Data complexity</th>
<th>Model complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stochastic data (gap)</td>
<td>Simple model is good</td>
</tr>
<tr>
<td>Low noise</td>
<td>Multiple good models</td>
</tr>
<tr>
<td>Low variance</td>
<td></td>
</tr>
</tbody>
</table>

- Second-order
- Quantiles
Beyond the Worst Case

Two reasons data is often easier in practice:

Data complexity
- Stochastic data (gap)
- Low noise
- Low variance

Model complexity
- Simple model is good
- Multiple good models

Second-order & Quantiles
- Any combination
All we need is the right learning rate

Existing algorithms
(Hedge, Prod, ...)

with

oracle
learning rate $\eta$

exploit

Sec-ord. & Quant.
All we need is the right learning rate

Existing algorithms (Hedge, Prod, . . .) with oracle learning rate $\eta$

Can we exploit Second-order & Quantiles on-line?
All we need is the right learning rate

Existing algorithms (Hedge, Prod, ...)

with oracle learning rate $\eta$

Can we exploit Second-order & Quantiles on-line?

Can we learn the learning rate?
But everyone struggles with the learning rate

Oracle $\eta$

- not monotonic,
- not smooth over time.

State of the art:

or

But everyone struggles with the learning rate

Oracle $\eta$

- **not** monotonic,
- **not** smooth over time.

State of the art:

Second-order


or

Quantiles

Main Result

Our new algorithm **Squint** learns the learning rate. It offers

- Run-time of Hedge
- Tiny \((\ln \ln T)\) overhead over oracle learning rate.
- Extension to Combinatorial Games
- Extension to Continuous domains (MetaGrad)
Overview

- Fundamental online learning problem
- Review previous guarantees
- New Squint algorithm with improved guarantees
Fundamental model for learning: 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 $w^t \ell^t$

The goal is to have small regret $R_k^T := T \sum_t w^t \ell^t - T \sum_t \ell_k^t$ with respect to every expert $k$.
Fundamental model for learning: 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 $\ell_t = (\ell^1_t, \ldots, \ell^K_t) \in [0, 1]^K$

- Learner incurs loss $\mathbf{w}_t^T \ell_t$
Fundamental model for learning: 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 $w^\top_t \ell_t$

- The goal is to have small regret

$$R^k_T := \sum_{t=1}^T w^\top_t \ell_t - \sum_{t=1}^T \ell^k_t$$

with respect to every expert $k$. 
Classic Hedge Result

The **Hedge** algorithm with **learning rate** $\eta$

$$w_{t+1}^k := \frac{e^{-\eta L_t^k}}{\sum_k e^{-\eta L_t^k}}$$

where

$$L_t^k = \sum_{s=1}^t \ell_s^k,$$

upon proper tuning of $\eta$ ensures [Freund and Schapire, 1997] $R_k T \preceq \sqrt{T \ln K}$ for each expert $k$ which is tight for adversarial (worst-case) losses but underwhelming in practice.

Two broad lines of improvement.

- Second-order bounds
- Quantile bounds
Classic Hedge Result

The **Hedge** algorithm with **learning rate** \( \eta \)

\[
w_{t+1}^k := \frac{e^{-\eta L_t^k}}{\sum_k e^{-\eta L_t^k}} \quad \text{where} \quad L_t^k = \sum_{s=1}^t \ell_s^k,
\]

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

\[
R_k^T \prec \sqrt{T \ln K} \quad \text{for each expert} \ k
\]

which is tight for adversarial (worst-case) losses
Classic Hedge Result

The Hedge algorithm with learning rate $\eta$

$$w_{t+1}^k := \frac{e^{-\eta L_t^k}}{\sum_k e^{-\eta L_t^k}}$$

where

$$L_t^k = \sum_{s=1}^t \ell_k^s,$$

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

$$R_T^k \preceq \sqrt{T \ln K}$$

for each expert $k$

which is tight for adversarial (worst-case) losses

but underwhelming in practice
Classic Hedge Result

The **Hedge** algorithm with **learning rate** $\eta$

$$w_{t+1}^k := \frac{e^{-\eta L_t^k}}{\sum_k e^{-\eta L_t^k}}$$

where

$$L_t^k = \sum_{s=1}^t \ell_s^k,$$

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

$$R_T^k \prec \sqrt{T \ln K}$$

for each expert $k$

which is tight for adversarial (worst-case) losses

but **underwhelming** in practice

Two broad lines of improvement.
The **Hedge** algorithm with learning rate $\eta$

\[
 w_{t+1}^k := \frac{e^{-\eta L_t^k}}{\sum_k e^{-\eta L_t^k}} \quad \text{where} \quad L_t^k = \sum_{s=1}^{t} \ell_s^k,
\]

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

\[
 R_T^k \prec \sqrt{T \ln K} \quad \text{for each expert } k
\]

which is tight for adversarial (worst-case) losses

but **underwhelming** in practice

Two broad lines of improvement.

- Second-order bounds
- Quantile bounds
Cesa-Bianchi et al. [2007], Hazan and Kale [2010], Chiang et al. [2012], De Rooij et al. [2014], Gaillard et al. [2014], Steinhardt and Liang [2014]

\[
R_T^k \prec \sqrt{V_T^k \ln K} \quad \text{for each expert } k.
\]

for some second-order quantity \( V_T^k \leq L_T^k \leq T \).
Second-order bounds

Cesa-Bianchi et al. [2007], Hazan and Kale [2010], Chiang et al. [2012], De Rooij et al. [2014], Gaillard et al. [2014], Steinhardt and Liang [2014]

\[ R_T^k < \sqrt{V_T^k \ln K} \quad \text{for each expert } k. \]

for some second-order quantity \( V_T^k \leq L_T^k \leq T \).

- Pro: stochastic case, learning sub-algorithms
- Con: specialized algorithms. hard-coded \( K \).
Quantile bounds

Hutter and Poland [2005], Chaudhuri et al. [2009], Chernov and Vovk [2010], Luo and Schapire [2014]

Prior $\pi$ on experts:

$$\min_{k \in \mathcal{K}} R_T^k \prec \sqrt{T \left( -\ln \pi(\mathcal{K}) \right)}$$

for each subset $\mathcal{K}$ of experts
Quantile bounds

Hutter and Poland [2005], Chaudhuri et al. [2009], Chernov and Vovk [2010], Luo and Schapire [2014]

Prior $\pi$ on experts:

$$\min_{k \in \mathcal{K}} R_T^k < \sqrt{T \left( -\ln \pi(\mathcal{K}) \right)} \quad \text{for each subset } \mathcal{K} \text{ of experts}$$

- **Pro**: over-discretized models, company baseline
- **Con**: specialized algorithms. Efficiency. Inescapable $T$. 
Our contribution

Squint [Koolen and Van Erven, 2015] guarantees

\[ R^K_T \prec \sqrt{V^K_T \left( -\ln \pi(K) + C_T \right)} \]

for each subset \( K \) of experts where

\[ R^K_T = \mathbb{E}_{\pi(k|K)} R^k_T \]

and

\[ V^K_T = \mathbb{E}_{\pi(k|K)} V^k_T \]

denote the average (under the prior \( \pi \)) among the reference experts \( k \in K \) of the regret \( R^k_T = \sum_{t=1}^{T} r^k_t \) and the (uncentered) variance of the excess losses \( V^k_T = \sum_{t=1}^{T} (r^k_t)^2 \) (where \( r^k_t = (w_t - e_k)^T \ell_t \)).
The cool . . .

- Squint aggregates over **all** learning rates
- While staying as efficient as Hedge
Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$. 
Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$. 

Potential function

$$\Phi_T := \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ e^{\eta R_k^T - \eta^2 V_k^T} \right],$$
Fix prior \( \pi(k) \) on experts and \( \gamma(\eta) \) on learning rates \( \eta \in [0, 1/2] \).

Potential function

\[
\Phi_T := \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ e^{\eta R^k_T - \eta^2 V^k_T} \right],
\]

Weights

\[
w^k_{T+1} := \frac{\pi(k) \mathbb{E}_{\gamma(\eta)} \left[ e^{\eta R^k_T - \eta^2 V^k_T} \right]}{\text{normalisation}}.
\]
Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$.

Potential function

$$\Phi_T := \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ e^{\eta R^k_T - \eta^2 V^k_T} \right],$$

Weights

$$w^k_{T+1} := \frac{\pi(k) \mathbb{E}_{\gamma(\eta)} \left[ e^{\eta R^k_T - \eta^2 V^k_T} \right]}{\text{normalisation}}.$$

Next:

- Argue weights ensure $1 = \Phi_0 \geq \Phi_1 \geq \Phi_2 \geq \cdots$.
- Derive second-order quantile bound from $\Phi_T \leq 1$. 
Squint Analysis: Potential Decreases

Theorem

Squint ensures: \( 1 = \Phi_0 \geq \Phi_1 \geq \Phi_2 \geq \cdots \)

Proof.

Let \( f_{T}^{k,\eta} := e^{\eta R_{T}^{k}} - \eta^2 V_{T}^{k} \) so that \( \Phi_{T} = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ f_{T}^{k,\eta} \right] \).
Squint Analysis: Potential Decreases

**Theorem**

*Squint ensures: \( 1 = \Phi_0 \geq \Phi_1 \geq \Phi_2 \geq \cdots \)*

**Proof.**

Let \( f_{T}^{k,\eta} := e^{\eta R_{T}^{k} - \eta^2 V_{T}^{k}} \) so that \( \Phi_{T} = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ f_{T}^{k,\eta} \right] \). Then

\[
\Phi_{T+1} = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ f_{T+1}^{k,\eta} \right] = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ f_{T}^{k,\eta} e^{\eta r_{T+1}^{k} - (\eta r_{T+1}^{k})^2} \right] \\
\leq \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ f_{T}^{k,\eta} (1 + \eta r_{T+1}^{k}) \right] \\
= \Phi_{T} + \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ f_{T}^{k,\eta} (w_{T+1} - e_k) \right]^T \ell_{T+1}
\]
Squint Analysis: Potential Decreases

Theorem

Squint ensures: $1 = \Phi_0 \geq \Phi_1 \geq \Phi_2 \geq \cdots$

Proof.

Let $f_{T}^{k,\eta} := e^{\eta R_{T}^{k} - \eta^2 V_{T}^{k}}$ so that $\Phi_T = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta}\right]$. Then

$$
\Phi_{T+1} = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T+1}^{k,\eta}\right] = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta} e^{\eta r_{T+1}^k - (\eta r_{T+1}^k)^2}\right] 
$$

$$
\leq \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta}(1 + \eta r_{T+1}^k)\right] 
$$

$$
= \Phi_T + \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta}(w_{T+1} - e_k)\right]^{T} \ell_{T+1}
$$

and the weights $w_{T+1} \propto \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta} e_k\right]$ ensure

$$
\mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta}(w_{T+1} - e_k)\right] = \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta}\right] w_{T+1} - \mathbb{E}_{\pi(k)\gamma(\eta)} \left[f_{T}^{k,\eta} e_k\right] = 0.
$$
Squint Analysis: Regret Bound

We have $1 \geq \Phi_T$. So for any $k$ and $\eta$

\[
0 \geq \ln \Phi_T = \ln \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ e^{\eta R^k_T - \eta^2 V^k_T} \right] \\
\geq \ln \left( \pi(k)\gamma(\eta)e^{\eta R^k_T - \eta^2 V^k_T} \right) \\
= \ln \pi(k) + \ln \gamma(\eta) + \eta R^k_T - \eta^2 V^k_T
\]
Squint Analysis: Regret Bound

We have $1 \geq \Phi_T$. So for any $k$ and $\eta$

$$0 \geq \ln \Phi_T = \ln \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ e^{\eta R_T^k - \eta^2 V_T^k} \right]$$

$$\geq \ln \left( \pi(k)\gamma(\eta)e^{\eta R_T^k - \eta^2 V_T^k} \right)$$

$$= \ln \pi(k) + \ln \gamma(\eta) + \eta R_T^k - \eta^2 V_T^k$$

Now $\max_{\eta} \left\{ \eta R_T^k - \eta^2 V_T^k \right\} = \frac{(R_T^k)^2}{4V_T^k}$ at $\hat{\eta} = \frac{R_T^k}{2V_T^k}$ and hence

$$\frac{(R_T^k)^2}{4V_T^k} \leq - \ln \pi(k) - \ln \gamma(\hat{\eta}),$$
Squint Analysis: Regret Bound

We have $1 \geq \Phi_T$. So for any $k$ and $\eta$

$$0 \geq \ln \Phi_T = \ln \mathbb{E}_{\pi(k)\gamma(\eta)} \left[ e^{\eta R^k_T - \eta^2 V^k_T} \right]$$

$$\geq \ln \left( \pi(k) \gamma(\eta) e^{\eta R^k_T - \eta^2 V^k_T} \right)$$

$$= \ln \pi(k) + \ln \gamma(\eta) + \eta R^k_T - \eta^2 V^k_T$$

Now $\max_\eta \left\{ \eta R^k_T - \eta^2 V^k_T \right\} = \frac{(R^k_T)^2}{4 V^k_T}$ at $\hat{\eta} = \frac{R^k_T}{2 V^k_T}$ and hence

$$\frac{(R^k_T)^2}{4 V^k_T} \leq - \ln \pi(k) - \ln \gamma(\hat{\eta})$$

so

$$R^k_T \leq 2 \sqrt{V^k_T \left( - \ln \pi(k) - \ln \gamma(\hat{\eta}) \right)} \quad \text{for all } k.$$
Three priors

Idea: have prior $\gamma(\eta)$ put sufficient mass around optimal $\hat{\eta}$
Three priors

Idea: have prior $\gamma(\eta)$ put sufficient mass around optimal $\hat{\eta}$

1. Uniform prior (generalizes to conjugate)

   $\gamma(\eta) = 2$

   Efficient algorithm, $C_T = \ln V_T^K$.
Three priors

Idea: have prior $\gamma(\eta)$ put sufficient mass around optimal $\hat{\eta}$

1. Uniform prior (generalizes to conjugate)

$$\gamma(\eta) = 2$$

Efficient algorithm, $C_T = \ln V_T^K$.

2. Chernov and Vovk [2010] prior

$$\gamma(\eta) = \frac{\ln 2}{\eta \ln^2(\eta)}$$

Not efficient, $C_T = \ln \ln V_T^K$. 

Three priors

Idea: have prior $\gamma(\eta)$ put sufficient mass around optimal $\hat{\eta}$

1. Uniform prior (generalizes to conjugate)

$$\gamma(\eta) = 2$$

Efficient algorithm, $C_T = \ln V_T^K$.

2. Chernov and Vovk [2010] prior

$$\gamma(\eta) = \frac{\ln 2}{\eta \ln^2(\eta)}$$

Not efficient, $C_T = \ln \ln V_T^K$.

3. Improper(!) log-uniform prior

$$\gamma(\eta) = \frac{1}{\eta}$$

Efficient algorithm, $C_T = \ln \ln T$
Implementation of Squint w. log-uniform prior

Closed-form expression for weights:

$$w^k_{T+1} \propto \pi(k) \int_0^{1/2} e^{\eta R^k_T - \eta^2 V^k_T} \frac{1}{\eta} d\eta$$

$$\propto \pi(k) e^{\frac{(R^k_T)^2}{4V^k_T}} \frac{\text{erf} \left( \frac{R^k_T}{2\sqrt{V^k_T}} \right) - \text{erf} \left( \frac{R^k_T - V^k_T}{2\sqrt{V^k_T}} \right)}{\sqrt{V^k_T}}.$$  

Note: erf part of e.g. C99 standard.
Constant time per expert per round
Extensions I

Combinatorial concept class $C \subseteq \{0, 1\}^K$:

- Shortest path
- Spanning trees
- Permutations
- ...

Component $i$Prod [Koolen and Van Erven, 2015] guarantees:

$$R_u T \preceq \sqrt{V_u T (\text{comp}(u) + K C_T)}$$

for each $u \in \text{conv}(C)$. The reference set of experts $K$ is subsumed by an "average concept" vector $u \in \text{conv}(C)$, for which our bound relates the coordinate-wise average regret $R_u T = \sum_t, k u_k r_k t$ to the averaged variance $V_u T = \sum_t, k u_k (r_k t)^2$ and the prior entropy $\text{comp}(u)$. No range factor. Drop-in replacement for Component Hedge [Koolen, Warmuth, and Kivinen, 2010].
Extensions I

Combinatorial concept class $\mathcal{C} \subseteq \{0, 1\}^K$:

- Shortest path
- Spanning trees
- Permutations
- . . .

Component iProd [Koolen and Van Erven, 2015] guarantees:

$$R_T^u \prec \sqrt{V_T^u (\text{comp}(u) + KC_T)}$$

for each $u \in \text{conv}(\mathcal{C})$.

The reference set of experts $\mathcal{K}$ is subsumed by an “average concept” vector $u \in \text{conv}(\mathcal{C})$, for which our bound relates the coordinate-wise average regret $R_T^u = \sum_{t,k} u_k r_t^k$ to the averaged variance $V_T^u = \sum_{t,k} u_k (r_t^k)^2$ and the prior entropy $\text{comp}(u)$. 
Extensions I

Combinatorial concept class $C \subseteq \{0, 1\}^K$:
- Shortest path
- Spanning trees
- Permutations
- ... 

Component iProd [Koolen and Van Erven, 2015] guarantees:

$$R_T^u \prec \sqrt{V_T^u (\text{comp}(u) + KC_T)}$$
for each $u \in \text{conv}(C)$.

The reference set of experts $\mathcal{K}$ is subsumed by an “average concept” vector $u \in \text{conv}(C)$, for which our bound relates the coordinate-wise average regret $R_T^u = \sum_{t,k} u_k r_t^k$ to the averaged variance $V_T^u = \sum_{t,k} u_k (r_t^k)^2$ and the prior entropy $\text{comp}(u)$.

No range factor. Drop-in replacement for Component Hedge

[Koolen, Warmuth, and Kivinen, 2010]
Extensions II

Setup generalized to

- Continuous (bounded) domain $U \subseteq \mathbb{R}^d$
- Convex loss functions $f_t : U \rightarrow \mathbb{R}$

Includes:

- Previous settings (linear)
- Online convex optimization

MetaGrad \cite{Van Erven and Koolen, 2016} guarantees:

$$R_u T \preceq \sqrt{V_u T \ln T}$$

for each $u \in U$.

- Weights become Gaussians.
- Run-time $O(d^2)$ per round (like Online Newton Step).
Extensions II

Setup generalized to

- Continuous (bounded) domain $\mathcal{U} \subseteq \mathbb{R}^d$
- Convex loss functions $f_t : \mathcal{U} \to \mathbb{R}$

Includes:

- Previous settings (linear)
- Online convex optimization

MetaGrad [Van Erven and Koolen, 2016] guarantees:

$$R^u_T \prec \sqrt{V^u_T \ln T}$$

for each $u \in \mathcal{U}$. 

Weights become Gaussians.

Run-time $O(d^2)$ per round (like Online Newton Step).
Extensions II

Setup generalized to

- Continuous (bounded) domain $\mathcal{U} \subseteq \mathbb{R}^d$
- Convex loss functions $f_t : \mathcal{U} \rightarrow \mathbb{R}$

Includes:

- Previous settings (linear)
- Online convex optimization

**MetaGrad** [Van Erven and Koolen, 2016] guarantees:

$$R_T^u \prec \sqrt{V_T^u d \ln T} \quad \text{for each } u \in \mathcal{U}.$$

- Weights become Gaussians.
- Run-time $O(d^2)$ per round (like Online Newton Step).
Central idea: **learning the learning rate**

A new set of tools
- fresh
- different
- efficient

for the well-studied experts problem.

Powerful generalizations to more complex problems.
Thank you!