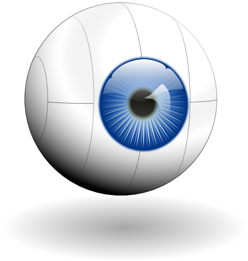
Second-order Quantile Methods



Wouter M. Koolen Tim van Erven

CWI Centrum Wiskunde & Informatica

Universiteit Leiden

Kyushu University, Monday 3rd 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

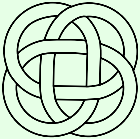
Two reasons data is often **easier** in practice:

Beyond the Worst Case

Two reasons data is often **easier** in practice:

Data complexity

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

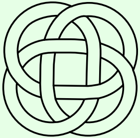


Beyond the Worst Case

Two reasons data is often **easier** in practice:

Data complexity

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



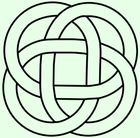
second-order

Beyond the Worst Case

Two reasons data is often **easier** in practice:

Data complexity

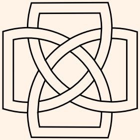
- Stochastic data (gap)
- Low noise
- Low variance



second-order

Model complexity

- Simple model is good
- Multiple good models

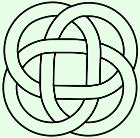


Beyond the Worst Case

Two reasons data is often **easier** in practice:

Data complexity

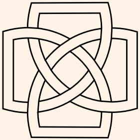
- Stochastic data (gap)
- Low noise
- Low variance



second-order

Model complexity

- Simple model is good
- Multiple good models



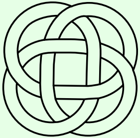
quantiles

Beyond the Worst Case

Two reasons data is often **easier** in practice:

Data complexity

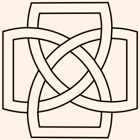
- Stochastic data (gap)
- Low noise
- Low variance



second-order

Model complexity

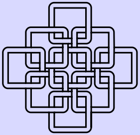
- Simple model is good
- Multiple good models



quantiles

Second-order & Quantiles

- Any combination

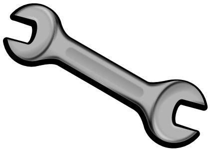


All we need is the right learning rate

Existing algorithms

(Hedge, Prod, . . .)

with

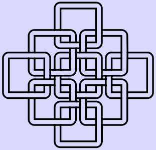


oracle

learning rate η

exploit

Sec-ord. & Quant.

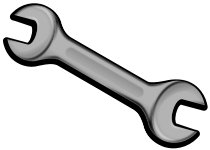


All we need is the right learning rate

Existing algorithms

(Hedge, Prod, . . .)

with

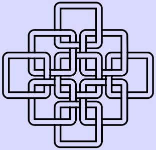


oracle

learning rate η

exploit

Sec-ord. & Quant.



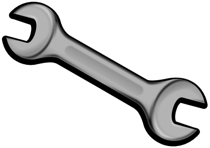
Can we exploit Second-order & Quantiles **on-line**?

All we need is the right learning rate

Existing algorithms

(Hedge, Prod, ...)

with

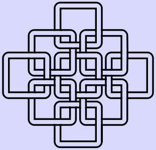


oracle

learning rate η

exploit

Sec-ord. & Quant.



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

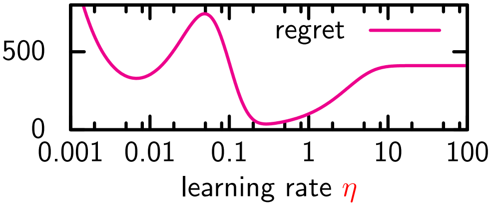
Can we **learn the learning rate**?

But everyone struggles with the learning rate

Oracle η

- **not** monotonic,
- **not** smooth

over time.

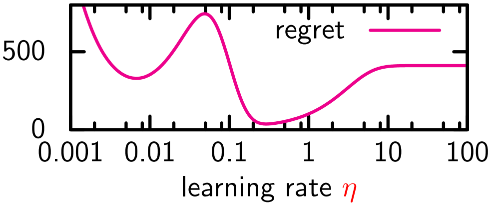


But everyone struggles with the learning rate

Oracle η

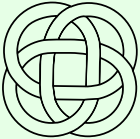
- **not** monotonic,
- **not** smooth

over time.



State of the art:

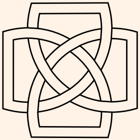
Second-order



Cesa-Bianchi, Mansour, and Stoltz 2007, Hazan and Kale 2010, Chiang, Yang, Lee, Mahdavi, Lu, Jin, and Zhu 2012, De Rooij, Van Erven, Grünwald, and Koolen 2014, Gaillard, Stoltz, and Van Erven 2014, Steinhardt and Liang 2014

or

Quantiles



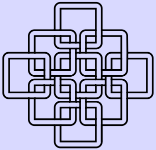
Hutter and Poland 2005, Chaudhuri, Freund, and Hsu 2009, Chernov and Vovk 2010, Luo and Schapire 2014

Main Result

Our new algorithm **Squint**

learns the learning rate. It offers

Second-order & Quantiles



- 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

   ...

Fundamental model for learning: Hedge setting

- K experts

   $\ldots$

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

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

Fundamental model for learning: 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.

Classic Hedge Result

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

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

Classic Hedge 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

Classic Hedge 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

but **underwhelming** in practice

Classic Hedge 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

but **underwhelming** in practice

Two broad lines of improvement.

Classic Hedge 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

but **underwhelming** in practice

Two broad lines of improvement.

Second-order bounds

Quantile bounds

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]

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

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

Second-order bounds

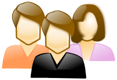
Cesa-Bianchi et al. [2007], Hazan and Kale [2010], Chiang et al. [2012], De Rooij et al. [2014], Gaillard et al. [2014], Steinhardt and Liang [2014]

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

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

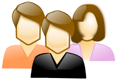


Hutter and Poland [2005], Chaudhuri et al. [2009], Chernov and Vovk [2010], Luo and Schapire [2014]

Prior π on experts:

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

Quantile bounds



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

Prior π on experts:

$$\min_{k \in \mathcal{K}} R_T^k \prec \sqrt{T\big(-\ln \pi(\mathcal{K})\big)} \qquad \text{for each subset } \mathcal{K} \text{ of experts}$$

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

Our contribution

Squint [Koolen and Van Erven, 2015] guarantees

$$R_T^{\mathcal{K}} \prec \sqrt{V_T^{\mathcal{K}}\left(-\ln \pi(\mathcal{K}) + C_T\right)} \qquad \text{for each subset } \mathcal{K} \text{ of experts}$$

where $R_T^{\mathcal{K}} = \mathbb{E}_{\pi(k|\mathcal{K})} R_T^k$ and $V_T^{\mathcal{K}} = \mathbb{E}_{\pi(k|\mathcal{K})} V_T^k$ denote the average (under the prior π) among the reference experts $k \in \mathcal{K}$ of the regret $R_T^k = \sum_{t=1}^T r_t^k$ and the (uncentered) variance of the excess losses $V_T^k = \sum_{t=1}^T (r_t^k)^2$ (where $r_t^k = (\boldsymbol{w}_t - \boldsymbol{e}_k)^\intercal \boldsymbol{\ell}_t$).

The cool . . .

- Squint aggregates over **all** learning rates
- While staying as efficient as Hedge

Squint

Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$.

Squint

Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$.

Potential function

$$\Phi_T \;\; := \;\; \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)} \left[e^{\eta R_T^k - \eta^2 V_T^k} \right],$$

Squint

Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$.

Potential function

$$\Phi_T \;\; := \;\; \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)} \left[e^{\eta R_T^k - \eta^2 V_T^k} \right],$$

Weights

$$w_{T+1}^k \;\; := \;\; \frac{\pi(k)\, \mathbb{E}_{\gamma(\eta)} \left[e^{\eta R_T^k - \eta^2 V_T^k} \eta \right]}{\text{normalisation}}.$$

Squint

Fix prior $\pi(k)$ on experts and $\gamma(\eta)$ on learning rates $\eta \in [0, 1/2]$.

Potential function

$$\Phi_T \; := \; \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)} \left[e^{\eta R_T^k - \eta^2 V_T^k} \right],$$

Weights

$$w_{T+1}^k \; := \; \frac{\pi(k) \, \mathbb{E}_{\gamma(\eta)} \left[e^{\eta R_T^k - \eta^2 V_T^k} \eta \right]}{\text{normalisation}}.$$

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

$$
\begin{aligned}
\Phi_{T+1} &= \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_{T+1}^{k,\eta}\right] = \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\; e^{\eta r_{T+1}^k - (\eta r_{T+1}^k)^2}\right] \\
&\leq \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}(1 + \eta r_{T+1}^k)\right] \\
&= \Phi_T + \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\eta(\boldsymbol{w}_{T+1} - \boldsymbol{e}_k)\right]^{\intercal} \boldsymbol{\ell}_{T+1}
\end{aligned}
$$

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

$$\begin{aligned}
\Phi_{T+1} &= \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_{T+1}^{k,\eta}\right] = \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\, e^{\eta r_{T+1}^k - (\eta r_{T+1}^k)^2}\right] \\
&\leq \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}(1 + \eta r_{T+1}^k)\right] \\
&= \Phi_T + \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\eta(\boldsymbol{w}_{T+1} - \boldsymbol{e}_k)\right]^\intercal \boldsymbol{\ell}_{T+1}
\end{aligned}$$

and the weights $\boldsymbol{w}_{T+1} \propto \mathbb{E}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\eta\boldsymbol{e}_k\right]$ ensure

$$\mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\eta(\boldsymbol{w}_{T+1} - \boldsymbol{e}_k)\right] = \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\eta\right]\boldsymbol{w}_{T+1} - \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)}\left[f_T^{k,\eta}\eta\boldsymbol{e}_k\right] = \boldsymbol{0}.$$

$\square$

Squint Analysis: Regret Bound

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

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

Squint Analysis: Regret Bound

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

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

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

$$\frac{(R_T^k)^2}{4 V_T^k} \;\leq\; -\ln \pi(k) - \ln \gamma(\hat{\eta}),$$

Squint Analysis: Regret Bound

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

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

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

$$
\frac{(R_T^k)^2}{4 V_T^k} \;\leq\; -\ln \pi(k) - \ln \gamma(\hat{\eta}),
$$

so

$$
R_T^k \;\leq\; 2\sqrt{V_T^k \left(-\ln \pi(k) \underbrace{- \ln \gamma(\hat{\eta})}_{C_T} \right)} \qquad \text{for all } k.
$$

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, $C_T = \ln V_T^{\mathcal{K}}$.

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, $C_T = \ln V_T^{\mathcal{K}}$.

2. Chernov and Vovk [2010] prior

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

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

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, $C_T = \ln V_T^{\mathcal{K}}$.

2. Chernov and Vovk [2010] prior

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

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

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:

$$
\begin{aligned}
w_{T+1}^{k} &\propto \pi(k) \int_0^{1/2} e^{\eta R_T^k - \eta^2 V_T^k} \eta \frac{1}{\eta} \, \mathrm{d}\eta \\
&\propto \pi(k) e^{\frac{(R_T^k)^2}{4 V_T^k}} \frac{\operatorname{erf}\left(\frac{R_T^k}{2\sqrt{V_T^k}}\right) - \operatorname{erf}\left(\frac{R_T^k - V_T^k}{2\sqrt{V_T^k}}\right)}{\sqrt{V_T^k}}.
\end{aligned}
$$

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

Extensions I

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

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

Extensions I

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

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

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

$$R_T^{\boldsymbol{u}} \prec \sqrt{V_T^{\boldsymbol{u}}\big(\mathrm{comp}(\boldsymbol{u}) + K\,C_T\big)} \qquad \text{for each } \boldsymbol{u} \in \mathrm{conv}(\mathcal{C}).$$

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

Extensions I

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

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

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

$$R_T^{\boldsymbol{u}} \prec \sqrt{V_T^{\boldsymbol{u}}\left(\operatorname{comp}(\boldsymbol{u}) + K\,C_T\right)} \qquad \text{for each } \boldsymbol{u} \in \operatorname{conv}(\mathcal{C}).$$

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

No range factor. Drop-in replacement for Component Hedge

[Koolen, Warmuth, and Kivinen, 2010]

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

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:

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

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:

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

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

Conclusion

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!