Learning Faster from Easy Data

Peter Grünwald Wouter M. Koolen
Sasha Rakhlin Karthik Sridharan

How Natural is the Worst Case?

Predict T coin flips

Regret $=$ My total loss$-\min\{$All-heads total loss, All-tails total loss$\}$

Minimax regret is $\sqrt{T}$ (IID fair coin)

How Natural is the Worst Case?

Predict T coin flips

Regret $=$ My total loss$-\min\{$All-heads total loss, All-tails total loss$\}$

Minimax regret is $\sqrt{T}$ (IID fair coin)

Any other IID coin:

- FTL gives constant regret ...

How Natural is the Worst Case?

Predict T coin flips

Regret $=$ My total loss$-\min\{$All-heads total loss, All-tails total loss$\}$

Minimax regret is $\sqrt{T}$ (IID fair coin)

Any other IID coin:

- FTL gives constant regret . . .
- . . . but is no solution: terrible worst-case regret (010101. . .)

How Natural is the Worst Case?

Predict T coin flips

Regret $=$ My total loss$-\min\{$All-heads total loss, All-tails total loss$\}$

Minimax regret is $\sqrt{T}$ (IID fair coin)

Any other IID coin:

- FTL gives constant regret . . .
- . . . but is no solution: terrible worst-case regret (010101. . .)
- . . . yet standard low regret algorithms retain $\sqrt{T}$ regret.

How Natural is the Worst Case?

Predict T coin flips

Regret $=$ My total loss$-\min\{$All-heads total loss, All-tails total loss$\}$

Minimax regret is $\sqrt{T}$ (IID fair coin)

Any other IID coin:

- FTL gives constant regret ...
- ... but is no solution: terrible worst-case regret (010101...)
- ... yet standard low regret algorithms retain $\sqrt{T}$ regret.

Not useful in practice ☹

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R = \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R \;=\; \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

Easy case: Stochastic w. gap

$$R \;=\; c \cdot \frac{\ln K}{T}$$

Achieved by FTL/EW with const η

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R \;=\; \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

const η is bad

Easy case: Stochastic w. gap

$$R \;=\; c \cdot \frac{\ln K}{T}$$

Achieved by FTL/EW with const η

$\eta = \frac{1}{\sqrt{T}}$ is bad

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R = \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

const η is bad

Stochastic IID: $R =$ Excess Risk

$$\min_{\text{alg}} \max_{\text{dist}} R = \sqrt{\frac{\ln K_T}{T}}$$

Achieved by ERM

Easy case: Stochastic w. gap

$$R = c \cdot \frac{\ln K}{T}$$

Achieved by FTL/EW with const η

$\eta = \frac{1}{\sqrt{T}}$ is bad

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R \;=\; \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

const η is bad

Easy case: Stochastic w. gap

$$R \;=\; c \cdot \frac{\ln K}{T}$$

Achieved by FTL/EW with const η

$\eta = \frac{1}{\sqrt{T}}$ is bad

Stochastic IID: $R = \text{Excess Risk}$

$$\min_{\text{alg}} \max_{\text{dist}} R \;=\; \sqrt{\frac{\ln K_T}{T}}$$

Achieved by ERM

Easy case: Tsybakov(κ) condition

$$R \;=\; \left(\frac{\ln K_T}{T}\right)^{\frac{\kappa}{2\kappa-1}}$$

Exploited by ERM

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R = \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

const η is bad

Easy case: Stochastic w. gap

$$R = c \cdot \frac{\ln K}{T}$$

Achieved by FTL/EW with const η

$\eta = \frac{1}{\sqrt{T}}$ is bad

Stochastic IID: $R = \text{Excess Risk}$

$$\min_{\text{alg}} \max_{\text{dist}} R = \sqrt{\frac{-\ln \pi(\text{best})}{T}}$$

Achieved by "Bayes" with $\eta = \frac{1}{\sqrt{T}}$

Easy case: Tsybakov(κ) condition

$$R = \left(\frac{-\ln \pi(\text{best})}{T}\right)^{\frac{\kappa}{2\kappa-1}}$$

Achieved by Bayes w. $\eta = T^{1-\frac{\kappa}{2\kappa-1}}$

This Problem is Everywhere

Individual Sequence: $R = \frac{\text{Regret}}{T}$

$$\min_{\text{alg}} \max_{\text{data}} R = \sqrt{\frac{\ln K}{T}}$$

Achieved by Hedge/EW with $\eta = \frac{1}{\sqrt{T}}$

const η is bad

Easy case: Stochastic w. gap

$$R = c \cdot \frac{\ln K}{T}$$

Achieved by FTL/EW with const η

$\eta = \frac{1}{\sqrt{T}}$ is bad

Stochastic IID: $R =$ Excess Risk

$$\min_{\text{alg}} \max_{\text{dist}} R = \sqrt{\frac{-\ln \pi(\text{best})}{T}}$$

Achieved by "Bayes" with $\eta = \frac{1}{\sqrt{T}}$

higher η are bad

Easy case: Tsybakov(κ) condition

$$R = \left(\frac{-\ln \pi(\text{best})}{T}\right)^{\frac{\kappa}{2\kappa-1}}$$

Achieved by Bayes w. $\eta = T^{1-\frac{\kappa}{2\kappa-1}}$

other η are bad

Punchline

No single algorithm seems to work in general

Different degrees of easiness seem to require different algorithms

Punchline

No single algorithm seems to work in general

Different degrees of easiness seem to require different algorithms

or do they . . . ?

Punchline

No single algorithm seems to work in general

Different degrees of easiness seem to require different algorithms

or do they . . . ?

Adaptive algorithms exist adapting to some types of luckiness in some settings, while preserving minimax guarantees:

- Srebro low target error in non-parametric setting
- Agarwal high margin in active learning setting
- Sridharan past proves future cannot be worst-case
- Van Erven data for which FTL works well (e.g. stochastic)
- Bubeck stochastic bandit feedback

Goals of this workshop

- Develop general methods for constructing algorithms that adapt to general types of easiness
- Determine classes of easiness worth exploiting in practice

Recent developments suggest answers may be within our reach

Partial Unification of Easiness Notions

[vEGRW12] subsume three important easiness criteria

- Statistical learning
- Density estimation when model wrong
- Ind. seq. prediction with easy loss fn.

- (Generalised) Tsybakov condition
- Barron-Li-Van der Vaart martingale condition
- Vovk mixability
  $\supset$ exp-concavity
  $\supset$ strong convexity

} Stochastic mixability

for every action a: $\mathop{\mathbb{E}}_{Y \sim P} \left[\frac{e^{-\eta \ell(Y,a)}}{e^{-\eta \ell(Y,a^*)}} \right] \leq 1$ (SM-η)

Partial Unification of Easiness Notions

[vEGRW12] subsume three important easiness criteria

- Statistical learning
- Density estimation when model wrong
- Ind. seq. prediction with easy loss fn.
- (Generalised) Tsybakov condition
- Barron-Li-Van der Vaart martingale condition
- Vovk mixability
  $\supset$ exp-concavity
  $\supset$ strong convexity

} Stochastic mixability

for every action a: $\mathop{\mathbb{E}}_{Y \sim P} \left[\frac{e^{-\eta \ell(Y,a)}}{e^{-\eta \ell(Y,a^*)}} \right] \leq 1$ (SM-η)

Loss Vovk mixable iff stochastically mixable for all distributions

Easiness sans Stochastics

Small regret when

- Prior luckiness
  - simple (high prior) best expert [Hutter & Poland, 2005]
  - many good experts [Chaudhuri, Freund & Hsu 2009]
  - few leaders [Gofer, Cesa-Bianchi, Gentile & Mansour 2013]

Easiness sans Stochastics

Small regret when

- Prior luckiness
  - simple (high prior) best expert [Hutter & Poland, 2005]
  - many good experts [Chaudhuri, Freund & Hsu 2009]
  - few leaders [Gofer, Cesa-Bianchi, Gentile & Mansour 2013]
- IID type luckiness
  - best expert has low loss [Auer, Cesa-Bianchi & Gentile 2002]
  - algorithm issues low variance predictions
    [Cesa-Bianchi, Mansour & Stoltz 2007]
  - best expert loss has low variance [Hazan & Kale 2008]

Easiness sans Stochastics

Small regret when

- Prior luckiness
  - simple (high prior) best expert [Hutter & Poland, 2005]
  - many good experts [Chaudhuri, Freund & Hsu 2009]
  - few leaders [Gofer, Cesa-Bianchi, Gentile & Mansour 2013]
- IID type luckiness
  - best expert has low loss [Auer, Cesa-Bianchi & Gentile 2002]
  - algorithm issues low variance predictions
    [Cesa-Bianchi, Mansour & Stoltz 2007]
  - best expert loss has low variance [Hazan & Kale 2008]
- Non-stationary luckiness
  - expert losses evolve slowly over time
    [Chiang, Yang, Lee, Mahdavi, Lu, Jin & Zhu 2012]
  - expert losses are predictable [Rakhlin & Karthik 2013]
- ...

We insist: your next algorithm is both

robust in the worst case

and

optimal in the lucky case

Enjoy!