Learning Faster from Easy Data II

Wouter Koolen

Tim van Erven

CWI Centrum Wiskunde & Informatica

Universiteit Leiden

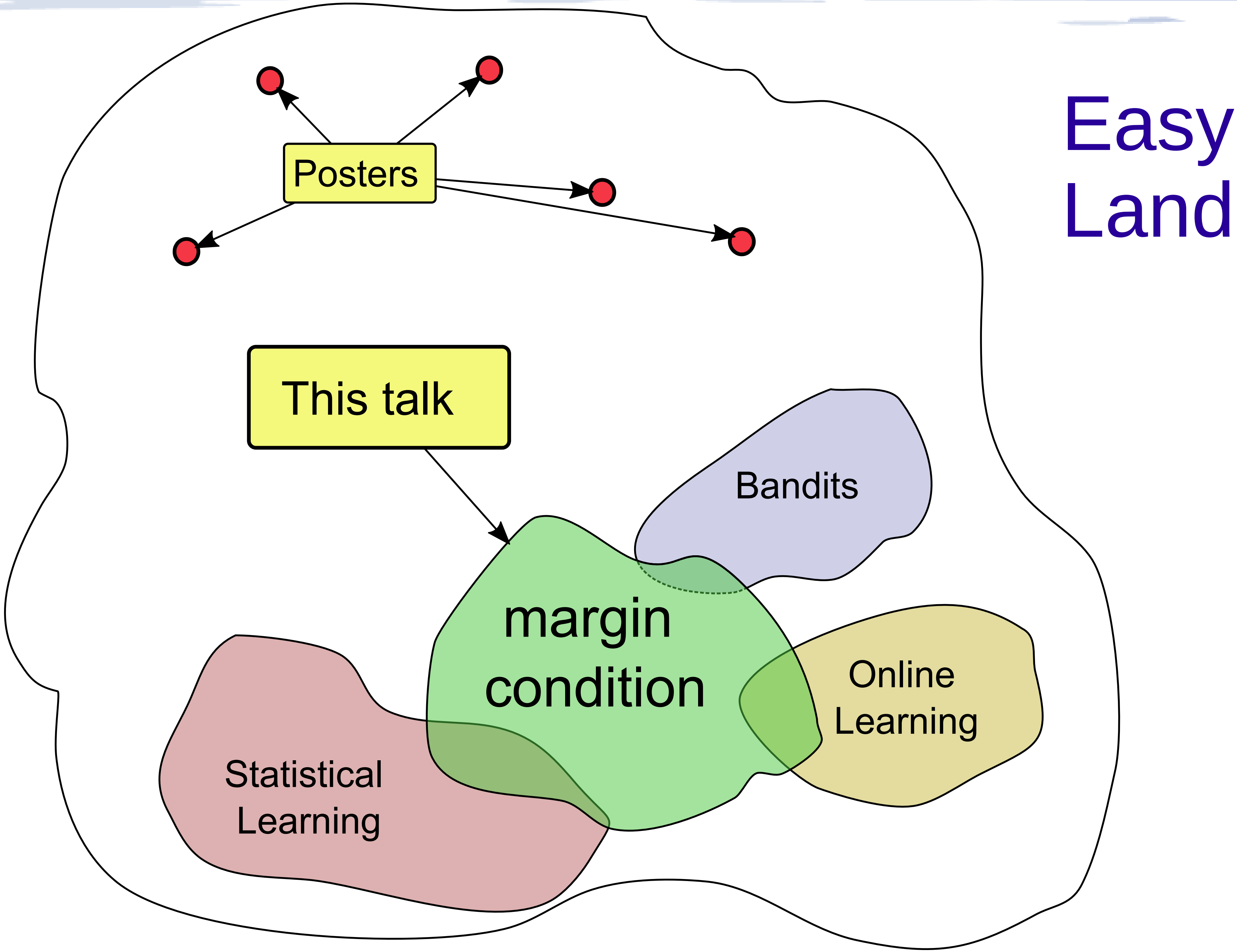
Aim of the Workshop

- Minimax analysis gives **robust algorithms**
- But in common **easy cases** these are overly **conservative**
  - Large gap between performance predicted by theory and observed in practice
- This workshop:
  - Bring together easy cases in different learning settings
  - New algorithms: robust to worst case, but ***automatically*** adapt to easy cases to learn faster

Learning Settings Easy Cases

(non-exhaustive list)

Standard statistical learning **Active learning**	• Margin condition (classification), Bernstein condition • Data fit low-complexity model • Sparsity
Online learning	• Curvature of the loss: strong convexity, exp-concavity, mixability • Small variance: 2nd-order bounds, IID losses + gap, small losses, ... • Many “good” experts
Bandits	• Stochastic = IID losses + gap
Clustering	• K-Means “works”


Easy
Land
Posters
This talk
Bandits
margin
condition
Online
Learning
Statistical
Learning

Outline

- **Easy data**
    - statistical learning
    - online learning
    - bandits
- How to exploit easy data
    - statistical learning
    - online learning
- The price of adaptivity

Statistical Learning

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ X_n \end{pmatrix} \sim P$$

$$\downarrow$$

$$\hat{f}$$

$$\downarrow$$

small risk $\quad R(\hat{f}) = \mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\mathrm{loss}(X, Y, \hat{f})\right]$

compared to minimizer f^* of risk in model $\mathcal{F}$

Easy Data in Classification

For **worst-case** P learning is slow:

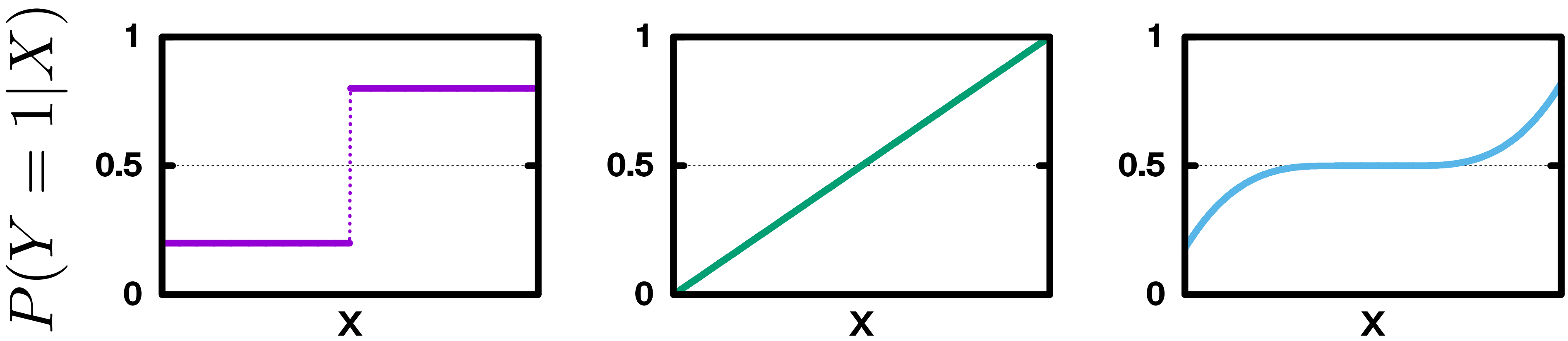
$$R(\hat{f}) - R(f^*) = O\left(\sqrt{\frac{\text{complexity}(\mathcal{F})}{n}}\right)$$

Margin condition: [Tsybakov, 2004]

- common case: $P(Y|X)$ not too close to $\frac{1}{2}$
- then learning is much faster, up to

$$R(\hat{f}) - R(f^*) = O\left(\frac{\text{complexity}(\mathcal{F})}{n}\right)$$

The Margin Condition



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

$$P\left(|P(Y|X) - \tfrac{1}{2}| \leq t\right) \leq ct^{\alpha}$$

Large Margin Reduces Variance

- Important source of excess risk $R(\hat{f}) - R(f^*)$ is **variance in excess loss**:

$$V(\hat{f}, f^*) = \mathbb{E}\left(\mathrm{loss}(X, Y, \hat{f}) - \mathrm{loss}(X, Y, f^*)\right)^2$$

- **Margin condition** $\Longleftrightarrow$ **Bernstein condition**:

$$V(\hat{f}, f^*) \leq c\big(R(\hat{f}) - R(f^*)\big)^{\kappa}$$

- Smaller excess risk $\Longrightarrow$ smaller variance

Large Margin Reduces Variance

- Important source of excess risk $R(\hat{f}) - R(f^*)$ is **variance in excess loss**:

$$V(\hat{f}, f^*) = \mathbb{E}\left(\text{loss}(X, Y, \hat{f}) - \text{loss}(X, Y, f^*)\right)^2$$

- **Margin condition** $\Longleftrightarrow$ **Bernstein condition**:

$$V(\hat{f}, f^*) \leq c\big(R(\hat{f}) - R(f^*)\big)^{\kappa}$$

- Smaller excess risk $\Longrightarrow$ smaller variance

Online Learning

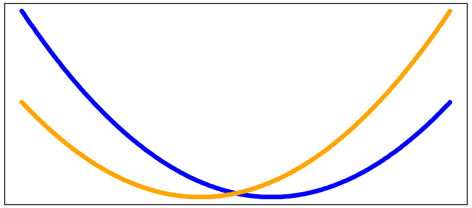
$$\begin{tikzcd}
 & \begin{pmatrix} Y_1 \\ X_1 \end{pmatrix} \arrow[dr] & & \begin{pmatrix} Y_2 \\ X_2 \end{pmatrix} \arrow[dr] & & \begin{pmatrix} Y_3 \\ X_3 \end{pmatrix} & \cdots \arrow[dr] & & \begin{pmatrix} Y_T \\ X_T \end{pmatrix} \\
\hat{f}_1 \arrow[ur] & & \hat{f}_2 \arrow[ur] & & \hat{f}_3 \arrow[ur] & & & \hat{f}_T \arrow[ur] &
\end{tikzcd}$$

small cumulative loss $L(\hat{f}) = \sum_{t=1}^{T} \mathrm{loss}(X_t, Y_t, \hat{f}_t)$

compared to minimizer f^* of cumulative loss in model $\mathcal{F}$

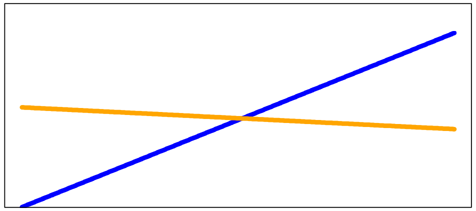
Easy Data in Online Learning

- Curved losses:



strongly convex,
exp-concave, mixable

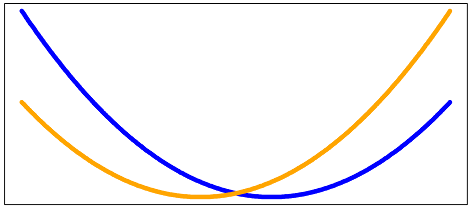
easier
than



linear loss

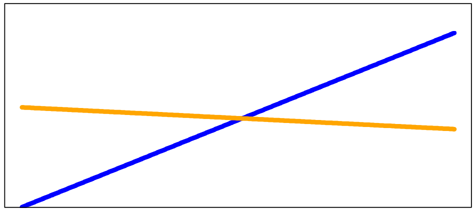
Easy Data in Online Learning

- Curved losses:



strongly convex,
exp-concave, mixable

easier than



linear loss

- Small empirical variance in excess losses:

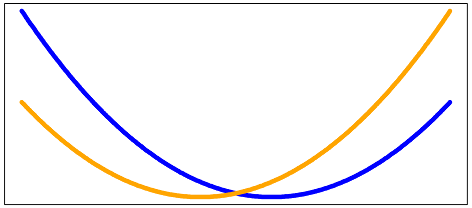
$$V = \frac{1}{T}\sum_{t=1}^{T}\Big(\mathrm{loss}(X_t, Y_t, \hat{f}_t) - \mathrm{loss}(X_t, Y_t, f^*)\Big)^2$$

Implied by:

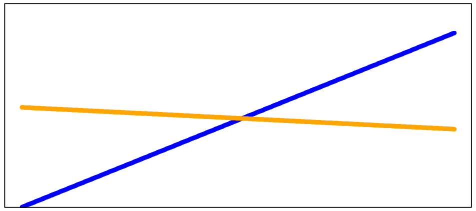
  - small losses (L^*-bounds)
  - i.i.d. losses + gap

Easy Data in Online Learning

- Curved losses:



easier than



strongly convex, exp-concave, mixable

linear loss

- Small empirical variance in excess losses:

$$V = \frac{1}{T}\sum_{t=1}^{T}\Big(\mathrm{loss}(X_t, Y_t, \hat{f}_t) - \mathrm{loss}(X_t, Y_t, f^*)\Big)^2$$

Implied by:

  - small losses (L^*-bounds)
  - i.i.d. losses + gap
  - Bernstein condition! ← **Grünwald**

i_2

Bandit Online Learning

$$\begin{tikzcd}[column sep=small]
 & \ell_{1,i_1} \arrow[dr] & & \ell_{2,i_2} \arrow[dr] & & \ell_{3,i_3} & & \cdots \arrow[dr] & & \ell_{T,i_T} \\
\hat{p}_1 \arrow[ur] & & \hat{p}_2 \arrow[ur] & & \hat{p}_3 \arrow[ur] & & & & \hat{p}_T \arrow[ur] &
\end{tikzcd}$$

- **K arms**/treatments with losses $\ell_{t,1}, \ldots, \ell_{t,K}$
- Only **observe** own (randomized) choice $i_t \sim \hat{p}_t$

small cumulative loss $\displaystyle\sum_{t=1}^{T} \ell_{t,i_t}$

compared to best fixed arm i^*

Easy Data for Bandits

- Stochastic bandits (easier):
  - Losses for arms are **independent, identically distributed** (i.i.d.)
  - Positive **gap** between expected performance of best arm and all others
- Adversarial bandits (harder):
  - Losses can be anything, even chosen to make learning as **difficult** as possible

Easy Data for Bandits

- Stochastic bandits (easier):
  - Losses for arms are **independent, identically distributed** (i.i.d.)
  - Positive **gap** between expected performance of best arm and all others
- Adversarial bandits (harder):
  - Losses can be anything, even chosen to make learning as **difficult** as possible
- Can a single algorithm adapt to:
  - **iid+gap + adversarial**? Auer
  - small losses + adversarial? **Neu**
  - small variance in general + adversarial?

Outline

- Easy data
    - statistical learning
    - online learning
    - bandits
- **How to exploit easy data**
    - statistical learning
    - online learning
- The price of adaptivity

Adaptive Statistical Learning

We consider exploiting κ-**Bernstein** cases:

$$V(\hat{f}, f^*) \leq c\big(R(\hat{f}) - R(f^*)\big)^{\kappa} \quad \kappa \in [0, 1]$$

Method: **penalized ERM $\hat{f}$ minimizes**

$$\sum_{i=1}^{n} \mathrm{loss}(X_i, Y_i, f) + \lambda \ln \frac{1}{\pi(f)}$$

(for simplicity: prior π on countable model $\mathcal{F}$)

How to tune λ?

Adaptive Statistical Learning

- Knowing κ, penalized ERM with $\lambda = \left(\frac{n}{\log \frac{1}{\pi(f^*)}}\right)^{\frac{1-\kappa}{2-\kappa}}$

$$R(\hat{f}) - R(f^*) = O\left(\left(\frac{\ln \frac{1}{\pi(f^*)}}{n}\right)^{\frac{1}{2-\kappa}}\right)$$

- **Adaptive** method through **holdout estimate**
- More sophisticated adaptive methods:
  - Slope heuristic [Birgé, Massart]
  - Lepski's method
  - Safe Bayes [Grünwald]

Adaptive Online Learning: Probabilistic Estimators

- Penalized ERM:

$$\min_f \sum_{i=1}^n \operatorname{loss}(X_i, Y_i, f) + \lambda \ln \frac{1}{\pi(f)}$$

- Allow probability distributions $p(f)$:

$$\min_p \mathop{\mathbb{E}}_{p(f)} \left[\sum_{i=1}^n \operatorname{loss}(X_i, Y_i, f) \right] + \frac{1}{\eta} \mathrm{KL}(p \| \pi)$$

Adaptive Online Learning: Probabilistic Estimators

- Penalized ERM:

$$\min_f \sum_{i=1}^n \mathrm{loss}(X_i, Y_i, f) + \lambda \ln \frac{1}{\pi(f)}$$

- Allow probability distributions $p(f)$:

$$\min_p \mathop{\mathbb{E}}_{p(f)} \left[\sum_{i=1}^n \mathrm{loss}(X_i, Y_i, f) \right] + \frac{1}{\eta} \mathrm{KL}(p \| \pi)$$

- Solution: **exponential weights**

$$p(f) = \frac{\pi(f) e^{-\eta \sum_{i=1}^n \mathrm{loss}(X_i, Y_i, f)}}{\mathrm{normalization}}$$

Adaptive Online Learning: Probabilistic Estimators

- Penalized ERM:

Remark: Obtain other methods like **gradient descent** by:
- changing KL to other regularizers +
- more general sets for p

- Allow probability distributions $p(f)$

$$\min_{p} \mathop{\mathbb{E}}_{p(f)} \left[\sum_{i=1}^{n} \mathrm{loss}(X_i, Y_i, f) \right] + \frac{1}{\eta} \mathrm{KL}(p \| \pi)$$

- Solution: **exponential weights**

$$p(f) = \frac{\pi(f) e^{-\eta \sum_{i=1}^{n} \mathrm{loss}(X_i, Y_i, f)}}{\mathrm{normalization}}$$

Adaptive Online Learning

$$p_{t+1}(f) = \frac{e^{-\eta \sum_{s=1}^{t} \mathrm{loss}_s(f)} \pi(f)}{\text{normalization}}$$

- For convex losses, play mean: $\hat{f}_t = \sum_f p_t(f) f$
- Standard tuning for the worst case

$$\eta \approx 1/\sqrt{T}$$

- Gives worst-case regret bound

$$O(\sqrt{T})$$

- Can we do **better** if we get κ-**Bernstein** data?

Adaptive Online Learning

- Turns out can indeed exploit κ-Bernstein data with correctly tuned η. In fact want $\eta = 1/\lambda$.
- But cannot do **holdout**
- Then how to tune eta?
  - One approach: tune η in terms of **upper bound** on regret that includes some **measure of variance**
  - Next slide: **learn** empirically **best learning rate** η for data at hand

Squint [Koolen and Van Erven 2015]

- **Exponential weights**: η needs external tuning

$$p_{t+1}(f) \propto \pi(f) e^{-\eta \sum_{s=1}^{t} \mathrm{loss}_s(f)} \propto \pi(f) e^{\eta R_t(f)}$$

exponential in regret $R_t(f) = \sum_{s=1}^{t} \mathrm{loss}_s(\hat{f}_s) - \mathrm{loss}_s(f)$.

Squint [Koolen and Van Erven 2015]

- **Exponential weights**: η needs external tuning

$$p_{t+1}(f) \propto \pi(f) e^{-\eta \sum_{s=1}^{t} \operatorname{loss}_s(f)} \propto \pi(f) e^{\eta R_t(f)}$$

exponential in regret $R_t(f) = \sum_{s=1}^{t} \operatorname{loss}_s(\hat{f}_s) - \operatorname{loss}_s(f)$.

- **Squint**: **learn** best η for the data

$$p_{t+1}(f) \propto \pi(f) \int_0^{1/2} e^{\eta R_t(f) - \eta^2 V_t(f)} \mathrm{d}\eta$$

with variance penalty $V_t(f) = \sum_{s=1}^{t} \Big(\operatorname{loss}_s(\hat{f}_s) - \operatorname{loss}_s(f)\Big)^2$.

Squint

- Philosophy: **learn** best η for the data

$$p_{t+1}(f) \propto \pi(f) \int_0^{1/2} e^{\eta R_t(f) - \eta^2 V_t(f)} \mathrm{d}\eta$$

- Important for current overview:
  - Optimal rate in **Bernstein** cases
- Further advantages beyond **stochastic case**:
  - Fast rates on **sub-adversarial** data
  - **Second-order** and **quantile** adaptivity

Outline

- Easy data
  - statistical learning
  - online learning
  - bandits
- How to exploit easy data
  - statistical learning
  - online learning
- **The price of adaptivity**

Price of adaptivity

- Settings where adaptivity is **cheap**
  - **Statistical learning**: holdout, etc.
  - **Online learning** (full inf.): Squint **(Grünwald, Foster)**
- Settings where adaptivity **subtle/unknown**
  - **Bandits** (IID stochastic / adversarial)
    - Adaptivity to both settings affordable **(Auer)**.
    - Can adapt to small losses (L^*) but general intermediate case very very tricky **(Neu)**.
  - **Active learning (Singh)**
  - **Online boosting**: **(Kale)**
    - Newly introduced setting (ICML best paper)
    - Seems some cost for adaptivity
  - **Clustering**: **(Ben-David)**
  - ...

Schedule

- Invited speakers
- **Spotlights + posters**:
  - Online learning, online convex optimization
  - Clustering
  - Statistical learning
  - Non-i.i.d. data
  - Bandits
- **Panel discussion**

Easy Land: great unknowns