

Advanced Machine Learning

Online Learning Basics

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Outline

- Prediction with expert advice
- Weighted Majority algorithm (WM)
- Randomized Majority algorithm (RWM)
- Online-to-batch conversion

Motivation

- PAC learning:
 - distribution fixed over time (training and test).
 - IID assumption.

- On-line learning:
 - no distributional assumption.
 - worst-case analysis (adversarial).
 - mixed training and test.
 - Performance measure: mistake model, regret.

General Online Setting

- For $t = 1$ to T do
 - receive instance $x_t \in X$.
 - predict $\hat{y}_t \in Y$.
 - receive label $y_t \in Y$.
 - incur loss $L(\hat{y}_t, y_t)$.
- Classification: $Y = \{0, 1\}$, $L(y, y') = |y' - y|$.
- Regression: $Y \subseteq \mathbb{R}$, $L(y, y') = (y' - y)^2$.
- **Objective:** minimize total loss $\sum_{t=1}^T L(\hat{y}_t, y_t)$.

Prediction with Expert Advice

- For $t = 1$ to T do
 - receive instance $x_t \in X$ and advice $\hat{y}_{t,i} \in Y, i \in [1, N]$.
 - predict $\hat{y}_t \in Y$.
 - receive label $y_t \in Y$.
 - incur loss $L(\hat{y}_t, y_t)$.
- **Objective:** minimize regret, i.e., difference of total loss incurred and that of the best expert,

$$\text{Regret}(T) = \sum_{t=1}^T L(\hat{y}_t, y_t) - \min_{i=1}^N \sum_{t=1}^T L(\hat{y}_{t,i}, y_t).$$

Halving Algorithm

[see (Mitchell, 1997)]

HALVING(H)

```
1    $H_1 \leftarrow H$ 
2   for  $t \leftarrow 1$  to  $T$  do
3       RECEIVE( $x_t$ )
4        $\hat{y}_t \leftarrow \text{MAJORITYVOTE}(H_t, x_t)$ 
5       RECEIVE( $y_t$ )
6       if  $\hat{y}_t \neq y_t$  then
7            $H_{t+1} \leftarrow \{c \in H_t : c(x_t) = y_t\}$ 
8   return  $H_{T+1}$ 
```

Halving Algorithm - Bound

(Littlestone, 1988)

- **Theorem:** Let H be a finite hypothesis set, then the number of mistakes made by the Halving algorithm is bounded as follows:

$$M_{\text{Halving}(H)} \leq \log_2 |H|.$$

- **Proof:** At each mistake, the hypothesis set is reduced at least by half.

Weighted Majority Algorithm

(Littlestone and Warmuth, 1988)

WEIGHTED-MAJORITY(N experts) $\triangleright y_t, y_{t,i} \in \{0, 1\}$.
 $\beta \in [0, 1]$.

```
1  for  $i \leftarrow 1$  to  $N$  do
2       $w_{1,i} \leftarrow 1$ 
3  for  $t \leftarrow 1$  to  $T$  do
4      RECEIVE( $x_t$ )
5       $\hat{y}_t \leftarrow 1_{\sum_{y_{t,i}=1}^N w_t \geq \sum_{y_{t,i}=0}^N w_t}$   $\triangleright$  weighted majority vote
6      RECEIVE( $y_t$ )
7      if  $\hat{y}_t \neq y_t$  then
8          for  $i \leftarrow 1$  to  $N$  do
9              if ( $y_{t,i} \neq y_t$ ) then
10                  $w_{t+1,i} \leftarrow \beta w_{t,i}$ 
11             else  $w_{t+1,i} \leftarrow w_{t,i}$ 
12  return  $\mathbf{w}_{T+1}$ 
```

Weighted Majority - Bound

- **Theorem:** Let m_t be the number of mistakes made by the WM algorithm till time t and m_t^* that of the best expert. Then, for all t ,

$$m_t \leq \frac{\log N + m_t^* \log \frac{1}{\beta}}{\log \frac{2}{1+\beta}}.$$

- Thus, $m_t \leq O(\log N) + \text{constant} \times \text{best expert.}$
- Realizable case: $m_t \leq O(\log N).$
- Halving algorithm: $\beta = 0.$

Weighted Majority - Proof

- Potential: $\Phi_t = \sum_{i=1}^N w_{t,i}$.
- Upper bound: after each error,

$$\Phi_{t+1} \leq [1/2 + 1/2 \beta] \Phi_t = \left[\frac{1 + \beta}{2} \right] \Phi_t.$$

Thus, $\Phi_t \leq \left[\frac{1 + \beta}{2} \right]^{m_t} N.$

- Lower bound: for any expert i , $\Phi_t \geq w_{t,i} = \beta^{m_{t,i}}$.
- Comparison:
$$\beta^{m_t^*} \leq \left[\frac{1 + \beta}{2} \right]^{m_t} N$$
$$\Rightarrow m_t^* \log \beta \leq \log N + m_t \log \left[\frac{1 + \beta}{2} \right]$$
$$\Rightarrow m_t \log \left[\frac{2}{1 + \beta} \right] \leq \log N + m_t^* \log \frac{1}{\beta}.$$

Weighted Majority - Notes

- **Advantage:** remarkable bound requiring no assumption.
- **Disadvantage:** no deterministic algorithm can achieve a regret $R_T = o(T)$ with the binary loss.
 - better guarantee with randomized WM.
 - better guarantee for WM with convex losses.

Exponential Weighted Average

Algorithm:

- weight update: $w_{t+1,i} \leftarrow w_{t,i} e^{-\eta L(\hat{y}_{t,i}, y_t)} = e^{-\eta L_{t,i}}$.
- prediction: $\hat{y}_t = \frac{\sum_{i=1}^N w_{t,i} \hat{y}_{t,i}}{\sum_{i=1}^N w_{t,i}}$.

total loss incurred by
expert i up to time t

Theorem:

assume that L is convex in its first argument and takes values in $[0, 1]$. Then, for any $\eta > 0$ and any sequence $y_1, \dots, y_T \in Y$, the regret at time T satisfies

$$\text{Regret}(T) \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

For $\eta = \sqrt{8 \log N / T}$,

$$\boxed{\text{Regret}(T) \leq \sqrt{(T/2) \log N}}.$$

EW - Proof

- Potential: $\Phi_t = \log \sum_{i=1}^N w_{t,i}$.
- Upper bound:

$$\begin{aligned}\Phi_t - \Phi_{t-1} &= \log \frac{\sum_{i=1}^N w_{t-1,i} e^{-\eta L(\hat{y}_{t,i}, y_t)}}{\sum_{i=1}^N w_{t-1,i}} \\ &= \log \left(\underset{w_{t-1}}{\text{E}} [e^{-\eta L(\hat{y}_{t,i}, y_t)}] \right) \\ &= \log \left(\underset{w_{t-1}}{\text{E}} \left[\exp \left(-\eta \left(L(\hat{y}_{t,i}, y_t) - \underset{w_{t-1}}{\text{E}} [L(\hat{y}_{t,i}, y_t)] \right) - \eta \underset{w_{t-1}}{\text{E}} [L(\hat{y}_{t,i}, y_t)] \right) \right] \right) \\ &\leq -\eta \underset{w_{t-1}}{\text{E}} [L(\hat{y}_{t,i}, y_t)] + \frac{\eta^2}{8} \quad (\text{Hoeffding's ineq.}) \\ &\leq -\eta L(\underset{w_{t-1}}{\text{E}} [\hat{y}_{t,i}], y_t) + \frac{\eta^2}{8} \quad (\text{convexity of first arg. of } L) \\ &= -\eta L(\hat{y}_t, y_t) + \frac{\eta^2}{8}.\end{aligned}$$

EW - Proof

- Upper bound: summing up the inequalities yields

$$\Phi_T - \Phi_0 \leq -\eta \sum_{t=1}^T L(\hat{y}_t, y_t) + \frac{\eta^2 T}{8}.$$

- Lower bound:

$$\begin{aligned}\Phi_T - \Phi_0 &= \log \sum_{i=1}^N e^{-\eta L_{T,i}} - \log N \geq \log \max_{i=1}^N e^{-\eta L_{T,i}} - \log N \\ &= -\eta \min_{i=1}^N L_{T,i} - \log N.\end{aligned}$$

- Comparison:

$$\begin{aligned}-\eta \min_{i=1}^N L_{T,i} - \log N &\leq -\eta \sum_{t=1}^T L(\hat{y}_t, y_t) + \frac{\eta^2 T}{8} \\ \Rightarrow \sum_{t=1}^T L(\hat{y}_t, y_t) - \min_{i=1}^N L_{T,i} &\leq \frac{\log N}{\eta} + \frac{\eta T}{8}.\end{aligned}$$

EW - Proof

- **Advantage:** bound on regret per bound is of the form

$$\frac{R_T}{T} = O\left(\sqrt{\frac{\log(N)}{T}}\right).$$

- **Disadvantage:** choice of η requires knowledge of horizon T .

Doubling Trick

- **Idea:** divide time into periods $[2^k, 2^{k+1} - 1]$ of length 2^k with $k = 0, \dots, n, T \geq 2^n - 1$, and choose $\eta_k = \sqrt{\frac{8 \log N}{2^k}}$ in each period.
- **Theorem:** with the same assumptions as before, for any T , the following holds:

$$\text{Regret}(T) \leq \frac{\sqrt{2}}{\sqrt{2} - 1} \sqrt{(T/2) \log N} + \sqrt{\log N/2}.$$

Doubling Trick - Proof

- By the previous theorem, for any $I_k = [2^k, 2^{k+1} - 1]$,

$$L_{I_k} - \min_{i=1}^N L_{I_k, i} \leq \sqrt{2^k / 2 \log N}.$$

$$\begin{aligned} \text{Thus, } L_T &= \sum_{k=0}^n L_{I_k} \leq \sum_{k=0}^n \min_{i=1}^N L_{I_k, i} + \sum_{k=0}^n \sqrt{2^k (\log N) / 2} \\ &\leq \min_{i=1}^N L_{T,i} + \sum_{k=0}^n 2^{\frac{k}{2}} \sqrt{(\log N) / 2}. \end{aligned}$$

with

$$\sum_{k=0}^n 2^{\frac{k}{2}} = \frac{\sqrt{2}^{n+1} - 1}{\sqrt{2} - 1} = \frac{2^{(n+1)/2} - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}\sqrt{T+1} - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}(\sqrt{T} + 1) - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}\sqrt{T}}{\sqrt{2} - 1} + 1.$$

Notes

- Doubling trick used in a variety of other contexts and proofs.
- More general method, learning parameter function of time: $\eta_t = \sqrt{(8 \log N)/t}$. Constant factor improvement:

$$\text{Regret}(T) \leq 2\sqrt{(T/2) \log N} + \sqrt{(1/8) \log N}.$$

General Setting

- Adversarial model with action set $X = \{1, \dots, N\}$.
- For $t=1$ to T do
 - player selects distribution p_t over X .
 - adversary selects loss $\mathbf{l}_t = (l_{t,1}, \dots, l_{t,N})$.
 - player receives \mathbf{l}_t .
 - player incurs loss $\sum_{i=1}^N p_{t,i} l_{t,i}$.
- **Objective:** minimize (external) regret

$$R_T = \sum_{t=1}^T \mathbb{E}_{i \sim p_t} [l_{t,i}] - \min_{i=1}^N \sum_{t=1}^T l_{t,i}.$$

Different Setups

- Deterministic vs. randomized.
- Full information vs. partial information (e.g. bandit setting).
- More general competitor class (e.g., swap regret).
- Oblivious vs. non-oblivious (or adaptive).
- Bounded memory.

Randomized Weighted Majority

(Littlestone and Warmuth, 1988)

RANDOMIZED-WEIGHTED-MAJORITY (N)

```
1  for  $i \leftarrow 1$  to  $N$  do
2       $w_{1,i} \leftarrow 1$ 
3       $p_{1,i} \leftarrow 1/N$ 
4  for  $t \leftarrow 1$  to  $T$  do
5      for  $i \leftarrow 1$  to  $N$  do
6          if ( $l_{t,i} = 1$ ) then
7               $w_{t+1,i} \leftarrow \beta w_{t,i}$ 
8          else  $w_{t+1,i} \leftarrow w_{t,i}$ 
9       $W_{t+1} \leftarrow \sum_{i=1}^N w_{t+1,i}$ 
10     for  $i \leftarrow 1$  to  $N$  do
11          $p_{t+1,i} \leftarrow w_{t+1,i}/W_{t+1}$ 
12 return  $\mathbf{w}_{T+1}$ 
```

RWM - Bound

- **Theorem:** fix $\beta \in [\frac{1}{2}, 1)$. Then, for any $T \geq 1$, the expected cumulative loss of RWM can be bounded as follows:

$$\mathcal{L}_T \leq \frac{\log N}{1 - \beta} + (2 - \beta)\mathcal{L}_T^{\min}.$$

For $\beta = \max \left\{ \frac{1}{2}, 1 - \sqrt{\frac{\log N}{T}} \right\}$,

$$\mathcal{L}_T \leq \mathcal{L}_T^{\min} + 2\sqrt{T \log N}.$$

RWM - Proof

■ Potential: $W_t = \sum_{i=1}^N w_{t,i}$.

■ Upper bound:

$$\begin{aligned} W_{t+1} &= \sum_{i: l_{t,i}=0} w_{t,i} + \beta \sum_{i: l_{t,i}=1} w_{t,i} = W_t + (\beta - 1) \sum_{i: l_{t,i}=1} w_{t,i} \\ &= W_t + (\beta - 1) W_t \sum_{i: l_{t,i}=1} p_{t,i} \\ &= W_t + (\beta - 1) W_t L_t \\ &= W_t (1 - (1 - \beta) L_t). \end{aligned}$$

Thus, $W_{T+1} = N \prod_{t=1}^T (1 - (1 - \beta) L_t)$.

■ Lower bound: $W_{T+1} \geq \max_{i \in [1, N]} w_{T+1,i} = \beta^{\mathcal{L}_T^{\min}}$.

RWM - Proof

■ Comparison:

$$\beta^{\mathcal{L}_T^{\min}} \leq N \prod_{t=1}^T (1 - (1 - \beta)L_t) \implies \mathcal{L}_T^{\min} \log \beta \leq \log N + \sum_{t=1}^T \log(1 - (1 - \beta)L_t)$$

$$(\forall x < 1, \log(1 - x) \leq -x) \implies \mathcal{L}_T^{\min} \log \beta \leq \log N - (1 - \beta) \sum_{t=1}^T L_t$$

$$\implies \mathcal{L}_T^{\min} \log \beta \leq \log N - (1 - \beta)\mathcal{L}_T$$

$$\implies \mathcal{L}_T \leq \frac{\log N}{1 - \beta} - \frac{\log \beta}{1 - \beta} \mathcal{L}_T^{\min}$$

$$\implies \mathcal{L}_T \leq \frac{\log N}{1 - \beta} - \frac{\log(1 - (1 - \beta))}{1 - \beta} \mathcal{L}_T^{\min}$$

$$(\forall x \in [0, 1/2], -\log(1 - x) \leq x + x^2) \implies \mathcal{L}_T \leq \frac{\log N}{1 - \beta} + (2 - \beta)\mathcal{L}_T^{\min}.$$

■ For the second statement, use

$$\mathcal{L}_T \leq \frac{\log N}{1 - \beta} + (2 - \beta)\mathcal{L}_T^{\min} \leq \frac{\log N}{1 - \beta} + (1 - \beta)T + \mathcal{L}_T^{\min}.$$

Lower Bound

- **Theorem:** let $N = 2$. There is a stochastic sequence of losses for which the expected regret of any algorithm verifies $E[R_T] \geq \sqrt{T/8}$.

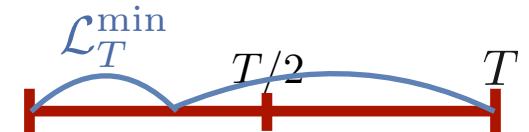
Lower Bound - Proof

- let \mathbf{l}_t take values $\mathbf{l}_{01} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ or $\mathbf{l}_{10} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ with equal probability. Then,

$$\mathbb{E}[\mathcal{L}_T] = \mathbb{E}\left[\sum_{t=1}^T \mathbf{p}_t \cdot \mathbf{l}_t\right] = \sum_{t=1}^T \mathbf{p}_t \cdot \mathbb{E}[\mathbf{l}_t] = \sum_{t=1}^T \frac{1}{2}p_{t,1} + \frac{1}{2}p_{t,2} = T/2.$$

- Since $\mathcal{L}_{T,1} + \mathcal{L}_{T,2} = T$,

$$\mathcal{L}_T^{\min} = T/2 - |\mathcal{L}_{T,1} - T/2|.$$



- Thus, by the Khintchine-Kahane ineq.,

$$\begin{aligned}\mathbb{E}[R_T] &= \mathbb{E}[\mathcal{L}_T] - \mathbb{E}[\mathcal{L}_T^{\min}] = \mathbb{E}[|\mathcal{L}_{T,1} - T/2|] \\ &= \mathbb{E}\left[\left|\sum_{t=1}^T \frac{1 + \sigma_t}{2} - T/2\right|\right] = \mathbb{E}\left[\left|\sum_{t=1}^T \frac{\sigma_t}{2}\right|\right] \geq \sqrt{T/8}.\end{aligned}$$

Online-to-Batch Conversion

■ Problem:

- sample $((x_1, y_1), \dots, (x_T, y_T)) \in (X \times Y)^T$ drawn i.i.d. according to D .
- loss function L bounded by $M > 0$.
- how do we combine the sequence h_1, \dots, h_{T+1} of hypotheses generated by a regret minimization algorithm to achieve a small generalization error?

Average Generalization

- **Lemma:** for any $\delta > 0$, with probability at least $1 - \delta$, the following holds

$$\frac{1}{T} \sum_{t=1}^T R(h_t) \leq \frac{1}{T} \sum_{t=1}^T L(h_t(x_t), y_t) + M \sqrt{\frac{2 \log \frac{1}{\delta}}{T}}.$$

- **Proof:** for any $t \in [1, T]$, let $V_t = R(h_t) - L(h_t(x_t), y_t)$.
 - Then,
$$\mathbb{E}[V_t | x_{1:t-1}] = R(h_t) - \mathbb{E}[L(h_t(x_t), y_t) | x_{1:t-1}] = R(h_t) - R(h_t) = 0.$$
 - By Azuma's inequality,

$$\Pr\left[\frac{1}{T} \sum_{i=1}^T V_i \geq \epsilon\right] \leq \exp(-2T\epsilon^2/(2M)^2)).$$

Online-to-Batch Guarantee

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for a convex loss:

$$R\left(\frac{1}{T} \sum_{i=1}^T h_i\right) \leq \frac{1}{T} \sum_{i=1}^T L(h_i(x_i), y_i) + M \sqrt{\frac{2 \log \frac{1}{\delta}}{T}}$$

$$R\left(\frac{1}{T} \sum_{i=1}^T h_i\right) \leq \inf_{h \in H} R(h) + \frac{R_T}{T} + 2M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}}.$$

Proof

- Convexity: $L\left(\frac{1}{T} \sum_{i=1}^T h_i(x), y\right) \leq \frac{1}{T} \sum_{i=1}^T L(h_i(x), y).$

Thus,

$$R\left(\frac{1}{T} \sum_{i=1}^T h_i\right) \leq \frac{1}{T} \sum_{i=1}^T R(h_i).$$

- By definition of the regret, with probability at least $1 - \delta/2$,

$$\begin{aligned} R\left(\frac{1}{T} \sum_{i=1}^T h_i\right) &\leq \frac{1}{T} \sum_{i=1}^T L(h_i(x_i), y_i) + M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}} \\ &\leq \inf_{h \in H} \frac{1}{T} \sum_{i=1}^T L(h(x_i), y_i) + \frac{R_T}{T} + M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}}. \end{aligned}$$

Proof

- Assume that the infimum is reached at h^* . By Hoeffding's inequality, with probability at least $1 - \delta/2$,

$$\frac{1}{T} \sum_{i=1}^T L(h^*(x_i), y_i) \leq R(h^*) + M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}}.$$

- Thus, with probability at least $1 - \delta$,

$$\begin{aligned} R\left(\frac{1}{T} \sum_{i=1}^T h_i\right) &\leq \frac{1}{T} \sum_{i=1}^T L(h^*(x_i), y_i) + \frac{R_T}{T} + M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}} \\ &\leq R(h^*) + M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}} + \frac{R_T}{T} + M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}} \\ &= R(h^*) + \frac{R_T}{T} + 2M \sqrt{\frac{2 \log \frac{2}{\delta}}{T}}. \end{aligned}$$

References

- Nicolò Cesa-Bianchi, Alex Conconi, Claudio Gentile: On the Generalization Ability of On-Line Learning Algorithms. *IEEE Transactions on Information Theory* 50(9): 2050-2057. 2004.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Yoav Freund and Robert Schapire. Large margin classification using the perceptron algorithm. In *Proceedings of COLT 1998*. ACM Press, 1998.
- Adam T. Kalai, Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71(3): 291-307. 2005.
- Nick Littlestone. From On-Line to Batch Learning. *COLT 1989*: 269-284.
- Nick Littlestone. "Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm" *Machine Learning* 285-318(2). 1988.

References

- Nick Littlestone, Manfred K. Warmuth: The Weighted Majority Algorithm. *FOCS* 1989: 256-261.
- Tom Mitchell. *Machine Learning*, McGraw Hill, 1997.
- Novikoff, A. B. (1962). On convergence proofs on perceptrons. *Symposium on the Mathematical Theory of Automata*, 12, 615-622. Polytechnic Institute of Brooklyn.

Advanced Machine Learning

Follow-The-Perturbed Leader

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

General Ideas

- Linear loss: decomposition as a sum along substructures.
 - sum of edge losses in a tree.
 - sum of edge losses along a path.
 - sum of other substructures losses in a discrete problem.
 - includes expert setting.

FPL

(Kalai and Vempala, 2004)

■ General linear decision problem:

- player selects $\mathbf{w}_t \in \mathcal{W} \subseteq \mathbb{R}^N$, $l_1\text{-diam}(\mathcal{W}) \leq W_1$.
- player receives $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^N$, $\mathcal{X} \subseteq \{\mathbf{x}: \|\mathbf{x}\|_1 \leq X_1\}$.
- player incurs loss $\mathbf{w}_t \cdot \mathbf{x}_t$, $\sup_{\mathbf{w} \in \mathcal{W}, \mathbf{x} \in \mathcal{X}} |\mathbf{w} \cdot \mathbf{x}| \leq R$.

■ Objective: minimize cumulative loss or regret.

■ Notation: $M(\mathbf{x}) = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \mathbf{w} \cdot \mathbf{x}$.

FL

- Follow the Leader (FL): use M at every round (aka **fictitious play**).
- FL problem: Suppose $N = 2$ and consider a sequence starting with $\begin{pmatrix} 0 \\ 1/2 \end{pmatrix}$ and then alternating $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Then,
 - FL incurs loss 1 at every round, T overall.
 - any single expert incurs loss $T/2$ overall.

FPL Algorithms

(Hannan 1957; Kalai and Vempala, 2004)

■ Additive bound Follow the Perturbed Leader (FPL):

- $\mathbf{p}_t \sim U([0, 1/\epsilon]^N)$.
- $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \sum_{s=1}^{t-1} \mathbf{w} \cdot \mathbf{x}_s + \mathbf{w} \cdot \mathbf{p}_t$
 $= M(\mathbf{x}_{1:t-1} + \mathbf{p}_t)$.

■ Multiplicative bound Follow the Perturbed Leader (FPL *):

- $\mathbf{p}_t \sim \text{Laplacian with density } f(\mathbf{x}) = \frac{\epsilon}{2} e^{-\epsilon \|\mathbf{x}\|_1}$.
- $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \sum_{s=1}^{t-1} \mathbf{w} \cdot \mathbf{x}_s + \mathbf{w} \cdot \mathbf{p}_t$
 $= M(\mathbf{x}_{1:t-1} + \mathbf{p}_t)$.

$\mathbf{x}_{1:t}$: sum of \mathbf{x}_{-1} to \mathbf{x}_t

FPL - Bound

- **Theorem:** fix $\epsilon > 0$. Then, the expected cumulative loss of additive FPL(ϵ) is bounded as follows

$$E[\mathcal{L}_T] \leq \mathcal{L}_T^{\min} + \epsilon R X_1 T + \frac{W_1}{\epsilon}.$$

For $\epsilon = \sqrt{\frac{W_1}{R X_1 T}}$

$$E[\mathcal{L}_T] \leq \mathcal{L}_T^{\min} + 2\sqrt{X_1 W_1 R T}.$$

FPL* - Bound

- **Theorem:** fix $\epsilon > 0$ and assume that $\mathcal{W}, \mathcal{X} \subseteq \mathbb{R}_+^N$. Then, the expected cumulative loss of (multiplicative) FPL*($\epsilon/2X_1$) is bounded as follows

$$\mathbb{E}[\mathcal{L}_T] \leq (1 + \epsilon)\mathcal{L}_T^{\min} + \frac{2X_1 W_1(1 + \log N)}{\epsilon}.$$

For $\epsilon = \min\left(1/2X_1, \sqrt{W_1(1 + \log N)/X_1 \mathcal{L}_T^{\min}}\right)$

$$\mathbb{E}[\mathcal{L}_T] \leq \mathcal{L}_T^{\min} + 4\sqrt{\mathcal{L}_T^{\min} X_1 W_1(1 + \log N)} + 4X_1 W_1(1 + \log N).$$

Proof Outline

■ Be the perturbed leader (BPL): $\mathbf{w}_t = M(\mathbf{x}_{1:t} + \mathbf{p}_t)$.

1. Bound on regret of BPL: $E[R_T(\text{BPL})] \leq \frac{W_1}{\epsilon}$.

2. Bound on difference of regrets of FPL and BPL:

$$E[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] - E[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t].$$

3. Difference of expectations small because similar distributions.

Proof: BL Regret

- Lemma 1: $\sum_{t=1}^T M(\mathbf{x}_{1:t}) \cdot \mathbf{x}_t \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T}$.
- Proof: case $T = 1$ is clear. By induction,

$$\begin{aligned} & \sum_{t=1}^{T+1} M(\mathbf{x}_{1:t}) \cdot \mathbf{x}_t \\ & \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{T+1} \quad (\text{induction}) \\ & \leq M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{1:T} + M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{T+1} \quad (\text{def. of } M(\mathbf{x}_{1:T}) \text{ as minimizer}) \\ & = M(\mathbf{x}_{1:T+1}) \cdot \mathbf{x}_{1:T+1}. \end{aligned}$$

Proof: BPL Regret

■ **Lemma 2:** let $p_0 = 0$. Then, the following holds:

$$\sum_{t=1}^T M(\mathbf{x}_{1:t} + \mathbf{p}_t) \cdot \mathbf{x}_t \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + W_1 \sum_{t=1}^T \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_\infty.$$

■ **Proof:** use Lemma 1 with $\mathbf{x}'_t = \mathbf{x}_t + \mathbf{p}_t - \mathbf{p}_{t-1}$, then

$$\begin{aligned} \sum_{t=1}^T M(\mathbf{x}_{1:t} + \mathbf{p}_t) \cdot (\mathbf{x}_t + \mathbf{p}_t - \mathbf{p}_{t-1}) &\leq M(\mathbf{x}_{1:T} + \mathbf{p}_T) \cdot (\mathbf{x}_{1:T} + \mathbf{p}_T) \\ &\leq M(\mathbf{x}_{1:T}) \cdot (\mathbf{x}_{1:T} + \mathbf{p}_T) \\ &= M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + M(\mathbf{x}_{1:T}) \cdot \sum_{t=1}^T \mathbf{p}_t - \mathbf{p}_{t-1}. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{t=1}^T M(\mathbf{x}_{1:t} + \mathbf{p}_t) \cdot \mathbf{x}_t &\leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + \sum_{t=1}^T [M(\mathbf{x}_{1:T}) - M(\mathbf{x}_{1:t} + \mathbf{p}_t)] \cdot [\mathbf{p}_t - \mathbf{p}_{t-1}] \\ &\leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + W_1 \sum_{t=1}^T \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_\infty. \end{aligned}$$

Proof: FPL vs. BPL Regrets

- **Proof:** for the expected loss, we can just choose $\mathbf{p}_t = \mathbf{p}_1$ for all $t > 0$, which yields:

$$\sum_{t=1}^T M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t \leq M(\mathbf{x}_{1:T}) \cdot \mathbf{x}_{1:T} + W_1 \|\mathbf{p}_1\|_\infty.$$

- Thus,

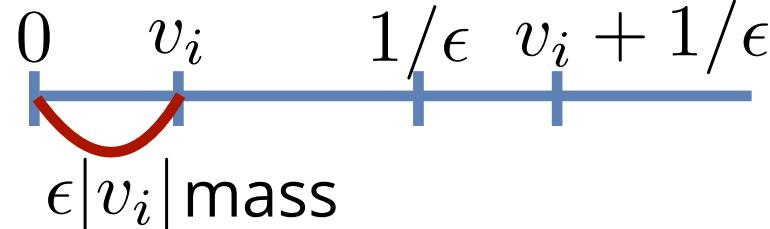
$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] \\ &= \sum_{t=1}^T \mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] - \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t] + \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t] \\ &\leq \sum_{t=1}^T \left[\mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] - \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t] \right] + \mathcal{L}_T^{\min} + W_1 \|\mathbf{p}_1\|_\infty. \end{aligned}$$

Proof: FPL

- By definition of the perturbation, $\|\mathbf{p}_1\|_\infty \leq \frac{1}{\epsilon}$.
- Now, $\mathbf{x}_{1:t} + \mathbf{p}_1$ and $\mathbf{x}_{1:t-1} + \mathbf{p}_1$ both follow a uniform distribution over a cube. Thus,

$$\mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] - \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t] \leq R(1 - \text{fraction of overlap}).$$

- Two cubes $[0, 1/\epsilon]^N$ and $\mathbf{v} + [0, 1/\epsilon]^N$ overlap over at least the fraction $(1 - \epsilon \|\mathbf{v}\|_1)$:
 - if $\mathbf{x} \in [0, 1/\epsilon]^N$ but $\mathbf{x} \notin \mathbf{v} + [0, 1/\epsilon]^N$ then for at least one i , $x_i \notin v_i + [0, 1/\epsilon]$, which has probability at most $\epsilon |v_i|$.



Proof: FPL

■ Thus,

$$\mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] - \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t] \leq R\epsilon \|\mathbf{x}_t\|_1 \leq R\epsilon X_1.$$

■ And,

$$\mathbb{E}[R_T] \leq R\epsilon X_1 T + \frac{W_1}{\epsilon}.$$

Proof: FPL*

■ Lemma 3:

$$\mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] \leq e^{\epsilon X_1} \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t].$$

■ Proof:

$$\begin{aligned} & \mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] \\ &= \int_{\mathbb{R}^N} M(\mathbf{x}_{1:t-1} + \mathbf{u}) \cdot \mathbf{x}_t d\mu(\mathbf{u}) \\ &= \int_{\mathbb{R}^N} M(\mathbf{x}_{1:t} + \mathbf{v}) \cdot \mathbf{x}_t d\mu(\mathbf{x}_t + \mathbf{v}) \quad (\text{change of var. } \mathbf{u} = \mathbf{v} + \mathbf{x}_t) \\ &= \int_{\mathbb{R}^N} M(\mathbf{x}_{1:t} + \mathbf{v}) \cdot \mathbf{x}_t \underbrace{e^{-\epsilon(\|\mathbf{x}_t + \mathbf{v}\|_1 - \|\mathbf{v}\|_1)}}_{\leq e^{\epsilon X_1}} d(\mathbf{v}) \\ &\leq e^{\epsilon X_1} \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t]. \end{aligned}$$

Proof: FPL*

- For $\epsilon \leq 1/X_1$, $e^{\epsilon X_1} \leq (1 + 2\epsilon X_1)$, thus,

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}[M(\mathbf{x}_{1:t-1} + \mathbf{p}_1) \cdot \mathbf{x}_t] &\leq \sum_{t=1}^T (1 + 2\epsilon X_1) \mathbb{E}[M(\mathbf{x}_{1:t} + \mathbf{p}_1) \cdot \mathbf{x}_t] \\ &\leq \sum_{t=1}^T (1 + 2\epsilon X_1)(\mathcal{L}_T^{\min} + W_1 \mathbb{E}[\|\mathbf{p}_1\|_\infty]). \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}[\|\mathbf{p}_1\|_\infty] &= \mathbb{E}\left[\max_{i \in [1, N]} |p_{1,i}|\right] = \int_0^{+\infty} \Pr\left[\max_{i \in [1, N]} |p_{1,i}| > t\right] dt \\ &\leq 2 \int_0^{+\infty} \Pr\left[\max_{i \in [1, N]} p_{1,i} > t\right] dt \\ &= 2 \int_0^u \Pr\left[\max_{i \in [1, N]} p_{1,i} > t\right] dt + \int_u^{+\infty} \Pr\left[\max_{i \in [1, N]} p_{1,i} > t\right] dt \\ &\leq 2u + N \int_u^{+\infty} \Pr\left[p_{1,1} > t\right] dt \\ &= 2u + N \frac{e^{-\epsilon u}}{\epsilon} \leq \frac{2(1 + \log N)}{\epsilon} \quad (\text{best choice of } u). \end{aligned}$$

Expert Setting

- $W_1 = 1, X_1 = N$, and $R = 1$; for $\text{FLP}^*(\epsilon)$,

$$\mathbb{E}[\mathcal{L}_T] \leq (1 + 2N\epsilon)\mathcal{L}_T^{\min} + \frac{2(1+\log(N))}{\epsilon}.$$

- More favorable bound:

- $\mathbf{x}_t \rightarrow x_{t,1}\mathbf{e}_1 \dots x_{t,N}\mathbf{e}_N$.
- new $\mathcal{L}_{NT}^{\min} = \text{old } \mathcal{L}_T^{\min}$. name: small loss gaurantee
- $\mathbb{E}[\mathcal{L}_T^{\text{old}}] \leq \mathbb{E}[\mathcal{L}_{TN}^{\text{new}}]$.
- new guarantee: for $\text{FLP}^*(\epsilon)$,

$$\mathbb{E}[\mathcal{L}_T] \leq (1 + 2\epsilon)\mathcal{L}_T^{\min} + \frac{2(1+\log(NT))}{\epsilon}.$$

$$\rightarrow \mathbb{E}[R_T] \leq 2\sqrt{2\mathcal{L}_T^{\min}(1 + \log(NT))}.$$

RWM = FPL

- Let $\text{FPL}(\eta)$ be an instance of the general FPL algorithm with a perturbation defined by

$$\mathbf{p}_1 = \left[\frac{\log(-\log(u_1))}{\eta}, \dots, \frac{\log(-\log(u_N))}{\eta} \right]^\top,$$

- where u_j is drawn according to the uniform distribution over $[0, 1]$.
- Then, $\text{FPL}(\eta)$ and $\text{RWM}(\eta)$ coincide.

References

- Nicolò Cesa-Bianchi, Alex Conconi, Claudio Gentile: On the Generalization Ability of On-Line Learning Algorithms. *IEEE Transactions on Information Theory* 50(9): 2050-2057. 2004.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Yoav Freund and Robert Schapire. Large margin classification using the perceptron algorithm. In *Proceedings of COLT 1998*. ACM Press, 1998.
- Adam T. Kalai, Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71(3): 291-307. 2005.
- Nick Littlestone. From On-Line to Batch Learning. *COLT 1989*: 269-284.
- Nick Littlestone. "Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm" *Machine Learning* 285-318(2). 1988.

References

- Nick Littlestone, Manfred K. Warmuth: The Weighted Majority Algorithm. *FOCS* 1989: 256-261.
- Tom Mitchell. *Machine Learning*, McGraw Hill, 1997.
- Novikoff, A. B. (1962). On convergence proofs on perceptrons. *Symposium on the Mathematical Theory of Automata*, 12, 615-622. Polytechnic Institute of Brooklyn.

Advanced Machine Learning

Learning and Games

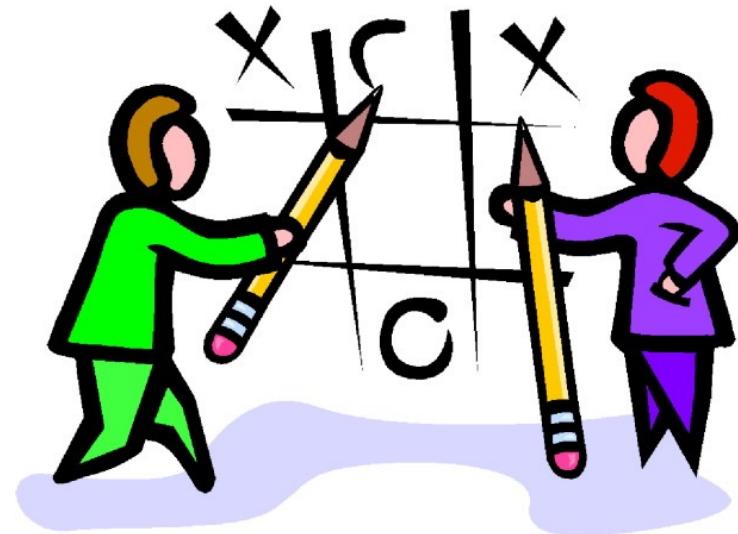
MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Outline

- Normal form games
- Nash equilibrium
- von Neumann's minimax theorem
- Correlated equilibrium
- Swap regret



Normal Form Games: Example

- Rock-Paper-Scissors.

	R	P	S
R	0,0	-1,1	1,-1
P	1,-1	0,0	-1,1
S	-1,1	1,-1	0,0

Be Truly Random

■ <http://goo.gl/3sVFzN>

Rock-Paper-Scissors: You vs. the Computer

Computers mimic human reasoning by building on simple rules and statistical averages. Test your strategy against the computer in this rock-paper-scissors game illustrating basic artificial intelligence. Choose from two different modes: novice, where the computer learns to play from scratch, and veteran, where the computer pits over 200,000 rounds of previous experience against you.

[TWITTER](#) [LINKEDIN](#) [SHARE](#)

Note: A truly random game of rock-paper-scissors would result in a statistical tie with each player winning, tying and losing one-third of the time. However, people are not truly random and thus can be studied and analyzed. While this computer won't win all rounds, over time it can exploit a person's tendencies and patterns to gain an advantage over its opponent.

HUMAN		
 Rock	 Paper	 Scissors



WINS	TIES	WINS
3	0	0

Round 4

 ✓	Round 3	
 ✓	Round 2	
 ✓	Round 1	

NOVICE COMPUTER		
Play at least five rounds to see what the computer is thinking.		



Normal Form Games

- p players.
- For each player $k \in [1, p]$:
 - set of actions (or **pure strategies**) \mathcal{A}_k .
 - payoff function $u_k : \prod_{k=1}^p \mathcal{A}_k \rightarrow \mathbb{R}$.
- Goal of each player: maximize his payoff in a repeated game.

Prisoner's Dilemma

Silence/Betrayal.

- for each player, the best action is B, regardless of the other player's action.
- but, with (B, B), both are worse off than (S, S).

payoff: reduction of the sentence

	S	B
S	2,2	0,3
B	3,0	1,1

Matching Pennies

- Player A wins when pennies match, player B otherwise.
 - other versions: penalty kick.
 - no pure strategy Nash equilibrium.

	H	T
H	1,-1	-1,1
T	-1,1	1,-1

Battle of The Sexes

Opera/Football.

- two pure strategy Nash equilibria.

	O	F
O	3,2	0,0
F	0,0	2,3

Mixed Strategies

■ Strategies:

- pure strategies: elements of $\prod_{k=1}^p \mathcal{A}_k$.
- mixed strategies: elements of $\prod_{k=1}^p \Delta_1(\mathcal{A}_k)$.

■ Payoff: for each player $k \in [1, p]$, when players play mixed strategies $(\mathbf{p}_1, \dots, \mathbf{p}_p)$,

$$\underset{a_j \sim \mathbf{p}_j}{\mathbb{E}} [u_k(\mathbf{a})] = \sum_{\mathbf{a}=(a_1, \dots, a_p)} \mathbf{p}_1(a_1) \cdots \mathbf{p}_p(a_p) u_k(\mathbf{a}).$$

Nash Equilibrium

- **Definition:** a mixed strategy (p_1, \dots, p_p) is a (mixed) Nash equilibrium if for all $k \in [1, p]$ and $q_k \in \Delta_1(\mathcal{A}_k)$,

$$u_k(q_k, p_{-k}) \leq u_k(p_k, p_{-k}).$$

- if for all k , p_k is a pure strategy, then (p_1, \dots, p_p) is said to be a pure Nash equilibrium.

Nash Equilibrium: Examples

- Prisoner's dilemma: (B, B) is a pure Nash equilibrium.
Dominant strategy: both better off playing B regardless of the other player's action.
- Matching Pennies: no pure Nash equilibrium; clear mixed Nash equilibrium: uniform probability for both.
- Battle of The Sexes:
 - pure Nash equilibria: both (O, O) and (F, F) .
 - mixed Nash equilibria: $((2/3, 1/3), (1/3, 2/3))$.
 - payoff of $2/3$ for both in mixed case: less than payoffs in pure cases!

Nash's Theorem

- **Theorem:** any normal form game with a finite set of players and finite set of actions admits a (mixed) Nash equilibrium.

Proof

- Define function $\Phi: \prod_{k=1}^p \Delta_1(\mathcal{A}_k) \rightarrow \prod_{k=1}^p \Delta_1(\mathcal{A}_k)$ by

$$\Phi(p_1, \dots, p_p) = (p'_1, \dots, p'_p)$$

with $\forall k \in [1, p], j \in [1, n_k]$, $p'^j_k = \frac{p_k^j + c_k^{j+}}{1 + \sum_{j=1}^{n_k} c_k^{j+}}$,

where $c_k^j = u_k(e_j, p_{-k}) - u_k(p_k, p_{-k})$, $c_k^{j+} = \max(0, c_k^j)$.

- Φ is a continuous function mapping from a non-empty compact convex set to itself, thus, by Brouwer's fixed-point theorem, there exists (p_1, \dots, p_p) such that

$$\Phi(p_1, \dots, p_p) = (p_1, \dots, p_p).$$

Proof

- Observe that for any $k \in [1, p]$,

$$\sum_{j=1}^{n_k} p_k^j c_k^j = \sum_{j=1}^{n_k} p_k^j u_k(\mathbf{e}_j, \mathbf{p}_{-k}) - u_k(\mathbf{p}_k, \mathbf{p}_{-k}) = 0.$$

- Thus, for any $k \in [1, p]$, there exists at least one j such that $c_k^j \leq 0$ with $p_k^j > 0$. For that j , $c_k^{j+} = 0$ and

$$\begin{aligned} p_k^j &= \frac{p_k^j}{1 + \sum_{j=1}^{n_k} c_k^{j+}} \Rightarrow 1 + \sum_{j=1}^{n_k} c_k^{j+} = 1 \\ &\Rightarrow c_k^{j+} = 0, \forall j \\ &\Rightarrow u_k(\mathbf{e}_j, \mathbf{p}_{-k}) \leq u_k(\mathbf{p}_k, \mathbf{p}_{-k}), \forall j \\ &\Rightarrow u_k(\mathbf{q}_k, \mathbf{p}_{-k}) \leq u_k(\mathbf{p}_k, \mathbf{p}_{-k}), \forall \mathbf{q}_k. \end{aligned}$$

Nash Equilibrium: Problems

- Different equilibria:
 - not clear which one will be selected.
 - different payoffs.
- Circular definition.
- Finding any Nash equilibrium is a PPAD-complete (polynomial parity argument on directed graphs) problem (Daskalakis et al., 2009).
- Not a natural model of rationality if computationally hard.

Zero-Sum Games: Order of Play

- If row player plays p then column player plays q solution of

$$\min_{q \in \Delta_1(\mathcal{A}_2)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{E} [u_1(\mathbf{a})].$$

- Thus, if row player starts, he plays p to maximize that quantity and the payoff is

$$\max_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{E} [u_1(\mathbf{a})].$$

- Similarly, if column player plays first, the expected payoff is

$$\min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{E} [u_1(\mathbf{a})].$$

von Neumann's Theorem

(von Neumann, 1928)

- **Theorem** (von Neumann's minimax theorem): for any two-player zero-sum game with finite action sets,

$$\max_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} E_{\substack{a_1 \sim p \\ a_2 \sim q}} [u_1(\mathbf{a})] = \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} E_{\substack{a_1 \sim p \\ a_2 \sim q}} [u_1(\mathbf{a})].$$

- common value called **value of the game**.
- mixed Nash equilibria coincide with maximizing and minimizing pairs and they all have the same payoff.

Proof

- Playing second is never worse:

$$\max_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})] \leq \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})].$$

- straightforward:

$$\forall p \in \Delta_1(\mathcal{A}_1), \forall q \in \Delta_1(\mathcal{A}_2), \quad \min_{q \in \Delta_1(\mathcal{A}_2)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})] \leq \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})]$$

$$\Rightarrow \quad \forall q \in \Delta_1(\mathcal{A}_2), \quad \max_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})] \leq \max_{p \in \Delta_1(\mathcal{A}_1)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})]$$

$$\Rightarrow \quad \max_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})] \leq \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} \underset{\substack{a_1 \sim p \\ a_2 \sim q}}{\mathbb{E}} [u_1(\mathbf{a})].$$

Proof

- Set-up: at each round,
 - column player selects q_t using RWM.
 - row player selects $p_t = \max_{p \in \Delta_1(\mathcal{A}_1)} p^\top \mathbf{U} q_t$.
- Thus, letting $T \rightarrow +\infty$ in the following completes the proof:

$$\begin{aligned}
 \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} \max_{\substack{a_1 \sim p \\ a_2 \sim q}} \mathbb{E}[u_1(a)] &= \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} p^\top \mathbf{U} q \\
 &\leq \max_{p \in \Delta_1(\mathcal{A}_1)} p^\top \mathbf{U} \left[\frac{1}{T} \sum_{t=1}^T q_t \right] = \max_{p \in \Delta_1(\mathcal{A}_1)} \frac{1}{T} \sum_{t=1}^T p^\top \mathbf{U} q_t \\
 &\leq \frac{1}{T} \sum_{t=1}^T \max_{p \in \Delta_1(\mathcal{A}_1)} p^\top \mathbf{U} q_t = \frac{1}{T} \sum_{t=1}^T p_t^\top \mathbf{U} q_t = \min_q \frac{1}{T} \sum_{t=1}^T p_t^\top \mathbf{U} q + \frac{R_T}{T} \\
 &= \min_q \left[\frac{1}{T} \sum_{t=1}^T p_t^\top \right] \mathbf{U} q + \frac{R_T}{T} \leq \max_p \min_q p^\top \mathbf{U} q + \frac{R_T}{T}.
 \end{aligned}$$

Proof

■ Let $p^* \in \operatorname{argmax}_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} u_1(p, q)$ and

$$q^* \in \operatorname{argmin}_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} u_1(p, q).$$

- p^* and q^* exist by the continuity of u_1 and the compactness of the simplices.
- By definition of p^* and q^* and the minmax theorem:

$$v = \min_q u_1(p^*, q) \leq u_1(p^*, q^*) \leq \max_p u_1(p, q^*) = v.$$

- Thus, (p^*, q^*) is a Nash equilibrium.

Proof

- Conversely, assume that (p^*, q^*) is a Nash equilibrium. Then,

$$u_1(p^*, q^*) = \max_p u_1(p, q^*) \geq \min_q \max_p u_1(p, q) = v$$

$$u_1(p^*, q^*) = \min_q u_1(p^*, q) \leq \max_p \min_q u_1(p, q) = v.$$

- This implies equalities and

$$u_1(p^*, q^*) = \max_p \min_q u_1(p, q) = \min_q \max_p u_1(p, q).$$

Notes

- Unique value: all Nash equilibria have the same payoff (less problematic than general case).
- Potentially several equilibria but no need to cooperate.
- Computationally efficient: convergence in $O\left(\sqrt{\frac{\log N}{T}}\right)$.
- Plausible explanation of how an equilibrium is reached — note that both players can play RWM.
- In general non-zero-sum games regret minimization does not lead to an equilibrium.

N: bound on the number of actions

Yao's Lemma

(Yao, 1977)

- **Theorem:** for any two-player zero-sum game with finite action sets,

$$\max_{p \in \Delta_1(\mathcal{A}_1)} \min_{a_2 \in \mathcal{A}_2} E_{a_1 \sim p}[u_1(a)] = \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{a_1 \in \mathcal{A}_1} E_{a_2 \sim q}[u_1(a)].$$

- consequence: for any distribution D over the inputs, the cost of a randomized algorithm is lower bounded by the minimum D-average cost of a deterministic algorithm.
- to determine a lower bound for the cost of a randomized algorithm, it suffices to inspect the complexity of deterministic algorithms with randomized inputs.

General Finite Games

- Regret notion not relevant: (external) regret minimization may not lead to a Nash equilibrium.
- Notion of equilibrium: several issues related to Nash equilibria.
→ new notion of equilibrium, new notion of regret.

Correlated Equilibrium - Tale

- There is an authority or a correlation mechanism device.
- The authority defines a probability distribution p over the p -tuple of the players' actions.
- The authority draws $(a_1, \dots, a_p) \sim p$ and reveals to each player k only his action a_k .
- The authority is a **correlated equilibrium** if player k has no incentive to deviate from the action recommended: the utility of any other action is lower than a_k , conditioned on the fact that he was told a_k , assuming that other players follow the recommendation they received.

Correlated Equilibrium

(Aumann, 1974)

- **Definition:** consider a normal form game with $p < +\infty$ players and finite action sets $\mathcal{A}_k, k \in [1, p]$. Then, a probability distribution p over $\prod_{k=1}^p \mathcal{A}_k$ is a **correlated equilibrium** if for all $k \in [1, p]$, for all $a_k \in \mathcal{A}_k$ with positive probability and all $a'_k \in \mathcal{A}_k$,

$$\underset{\mathbf{a} \sim p}{\text{E}}[u_k(a_k, a_{-k}) \mid a_k] \geq \underset{\mathbf{a} \sim p}{\text{E}}[u_k(a'_k, a_{-k}) \mid a_k].$$

Notes

- Think of the joint distribution as a correlation device.
- The set of all correlated equilibria is a convex set (it is a polyhedron): defined by a system of linear inequalities, including the simplex constraints. Solution via solving an LP problem.
- The set of Nash equilibria in general is not convex. It is defined by the intersection of the polyhedron of correlated equilibria and the constraints

$$p(\mathbf{a}) = p_1(a_1) \times \cdots \times p_p(a_p).$$

Traffic Lights

- Stop/Go.

	S	G
S	4,4	1,5
G	5,1	0,0

- Pure Nash equilibria: (S, G), (G, S). Mixed Nash equilibrium: $((1/2, 1/2), (1/2, 1/2))$.
- Correlated equilibria:

0	1/2
1/2	0
1/3	1/3
1/3	0

Internal Regret

- **Definition:** internal regret, $C_{a,b}$ functions $f: \mathcal{A} \rightarrow \mathcal{A}$ leaving all actions unchanged but a which is switched to b .

$$R_T = \sum_{t=1}^T \mathbb{E}_{a_t \sim p_t} [l(a_t)] - \min_{f \in C_{a,b}} \sum_{t=1}^T \mathbb{E}_{a_t \sim p_t} [l(f(a_t))].$$

- **Definition:** swap regret, C family of all functions $f: \mathcal{A} \rightarrow \mathcal{A}$.

$$R_T = \sum_{t=1}^T \mathbb{E}_{a_t \sim p_t} [l(a_t)] - \min_{f \in C} \sum_{t=1}^T \mathbb{E}_{a_t \sim p_t} [l(f(a_t))].$$

Swap Regret and Correlated Eq.

- **Theorem:** consider a finite normal form game played repeatedly. Assume that each player follows a swap regret minimizing strategy. Then, the empirical distribution of all plays converges to a correlated equilibrium.

Alternative Proof

- Let C be the convex set of correlated equilibria. If the sequence of empirical dist. $(\hat{p}_t)_{t \in \mathbb{N}}$ does not converge to C , it admits a subsequence in $C_\eta = \{p: d(p, C) \geq \eta\}$ (compact set), thus, it admits a subsequence converging to $\hat{p} \notin C$.
- Thus, there exist $\epsilon > 0$, $k \in [1, p]$, and $a_k, a'_k \in \mathcal{A}_k$ such that

$$\sum_{a_{-k} \in \mathcal{A}_{-k}} \hat{p}(\mathbf{a})[u_k(a'_k, a_{-k}) - u_k(a_k, a_{-k})] = \epsilon.$$

- Therefore, for t sufficiently large,

$$\sum_{a_{-k} \in \mathcal{A}_{-k}} \hat{p}_{\tau(t)}(\mathbf{a})[u_k(a'_k, a_{-k}) - u_k(a_k, a_{-k})] \geq \frac{\epsilon}{2}.$$

Alternative Proof

- Since $\widehat{p}_{\tau(t)}(\mathbf{a}) = \frac{1}{\tau(t)} \sum_{s=1}^{\tau(t)} 1_{\mathbf{a}_{-k,\tau(s)}=\mathbf{a}_{-k}} 1_{a_{k,\tau(s)}=a_k}$,
$$\frac{1}{\tau(t)} \sum_{s=1}^{\tau(t)} [u_k(a'_k, \mathbf{a}_{-k,\tau(s)}) - u_k(a_k, \mathbf{a}_{-k,\tau(s)})] 1_{a_{k,\tau(s)}=a_k} \geq \frac{\epsilon}{2}.$$
- Thus, the internal regret of player k for switching a_k to a'_k is lower bounded by $\frac{\epsilon}{2}$ at time $\tau(t)$ and later, which implies that the player is not following a swap regret minimization strategy.

Proof

- Define the instantaneous regret of player k at time t as

$$\widehat{r}_{k,t,j,j'} = \mathbf{1}_{a_{k,t}=j} [l_k(j, a_{-k,t}) - l_k(j', a_{-k,t})],$$

and $r_{k,t,j,j'} = p_{k,t,j} [l_k(j, a_{-k,t}) - l_k(j', a_{-k,t})]$.

- Then, $\mathbb{E}[\widehat{r}_{k,t,j,j'} | \text{past} \wedge \text{other players' actions}] = r_{k,t,j,j'}$.
- Thus, for any (j, j') , $(r_{k,t,j,j'} - \widehat{r}_{k,t,j,j'})$ is a bounded martingale difference. By Azuma's inequality and the Borell-Cantelli lemma, for all k and (j, j') ,

$$\limsup_{T \rightarrow +\infty} \frac{1}{T} \sum_{t=1}^T \widehat{r}_{k,t,j,j'} - r_{k,t,j,j'} = 0 \text{ (a.s.)}.$$

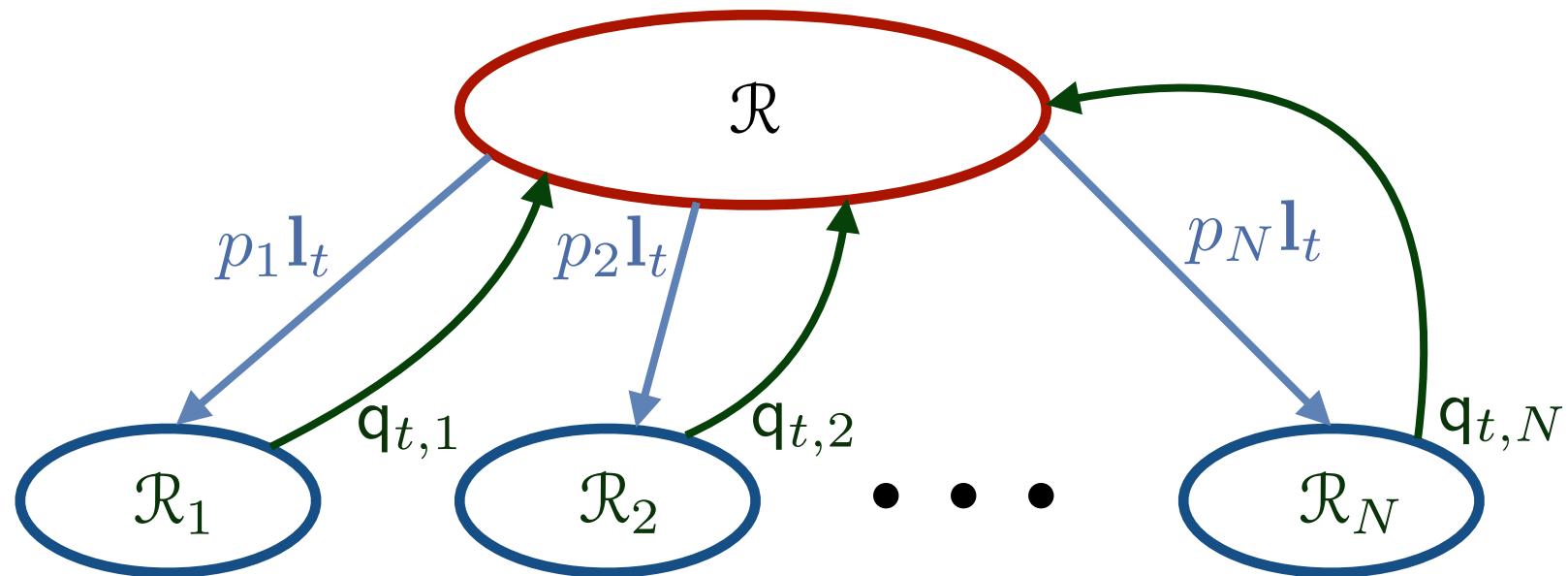
Therefore,

$$\forall k, \limsup_{T \rightarrow +\infty} \max_{j,j'} \frac{1}{T} \sum_{t=1}^T \widehat{r}_{k,t,j,j'} \leq 0 \text{ (a.s.)}.$$

Swap Regret Algorithm

(Blum and Mansour, 2007)

- **Theorem:** there exists an algorithm with $O(\sqrt{NT \log N})$ swap regret.



\mathcal{R}_i 's external regret minimization algorithms

Proof

- Define for all $t \in [1, T]$ the stochastic matrix

$$\mathbf{Q}_t = (q_{t,i,j})_{(i,j) \in [1, N]^2} = \begin{bmatrix} \mathbf{q}_{t,1}^\top \\ \vdots \\ \mathbf{q}_{t,N}^\top \end{bmatrix}.$$

- Since \mathbf{Q}_t is stochastic, it admits a stationary distribution \mathbf{p}_t :

$$\mathbf{p}_t^\top = \mathbf{p}_t^\top \mathbf{Q}_t \Leftrightarrow \forall j \in [1, N], p_{t,j} = \sum_{i=1}^N p_{t,i} q_{t,i,j}$$

- Thus,

$$\sum_{t=1}^T \mathbf{p}_t \cdot \mathbf{l}_t = \sum_{t=1}^T \sum_{j=1}^N p_{t,j} l_{t,j} = \sum_{t=1}^T \sum_{j=1}^N \sum_{i=1}^N p_{t,i} q_{t,i,j} l_{t,j} = \sum_{i=1}^N \sum_{t=1}^T \mathbf{q}_{t,i} \cdot (p_{t,i} \mathbf{l}_t) \leq \sum_{i=1}^N \min_j \sum_{t=1}^T p_{t,i} l_{t,j} + R_{T,i}.$$

Proof

- Thus, for any $f: \mathcal{A} \rightarrow \mathcal{A}$,

$$\sum_{t=1}^T \mathbf{p}_t \cdot \mathbf{l}_t \leq \sum_{i=1}^N \sum_{t=1}^T p_{t,i} l_{t,f(i)} + R_{T,i}.$$

- For RWM, $R_{T,i} = O(\sqrt{L_{\min,i} \log N})$. Thus, by Jensen's inequality,

$$\begin{aligned} \sum_{i=1}^N R_{T,i} &= N \frac{1}{N} \sum_{i=1}^N R_{T,i} \\ &\leq O \left(N \sqrt{\frac{1}{N} \sum_{i=1}^N L_{\min,i} \log N} \right) && \text{(Jensen's ineq.)} \\ &\leq O \left(N \sqrt{\frac{1}{N} T \log N} \right) = O \left(\sqrt{NT \log N} \right). && \left(\sum_{i=1}^N L_{\min,i} = \sum_{t=1}^T \sum_{i=1}^N p_{t,i} l_{t,j_i^*} \right) \end{aligned}$$

Notes

■ Surprising result:

- no explicit joint distribution in the game!
- correlation induced by the empirical sequence of plays by the players.

■ Game matrix:

- no need to know the full matrix (which could be huge with a lot of players).
- only need to know the loss or payoff for actions taken.

Coarse Correlated Equilibrium

- **Definition:** consider a normal form game with $p < +\infty$ players and finite action sets $\mathcal{A}_k, k \in [1, p]$. Then, a probability distribution \mathbf{p} over $\prod_{k=1}^p \mathcal{A}_k$ is a **coarse correlated equilibrium** if for all $k \in [1, p]$, for all $a_k \in \mathcal{A}_k$ and all $a'_k \in \mathcal{A}_k$,

$$\underset{\mathbf{a} \sim \mathbf{p}}{\text{E}} [u_k(a_k, a_{-k})] \geq \underset{\mathbf{a} \sim \mathbf{p}}{\text{E}} [u_k(a'_k, a_{-k})].$$

Notes

- Any correlated equilibrium is a coarse correlated equilibrium. Difference: realization a_k not known to player.
- Comparison with mixed Nash equilibria: (general) joint distribution vs. product distributions.
- Relationship with external regret, and external regret minimizers.

Conclusion

- Zero-sum finite games:
 - external regret minimization algorithms (e.g., RWM).
 - Nash equilibrium, value of the game reached.
- General finite games:
 - internal/swap regret minimization algorithms.
 - correlated equilibrium, can be learned.
- Questions:
 - Nash equilibria.
 - extensions: e.g., time selection functions ([Blum and Mansour, 2007](#)), conditional correlated equilibrium ([MM and Yang, 2014](#)).

References

- Robert Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.
- Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Constantinos Daskalakis, Paul W. Goldberg, Christos H. Papadimitriou. The complexity of computing a Nash equilibrium. *Commun. ACM* 52(2): 89-97 (2009).
- Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(12):40 – 55, 1997.
- Yoav Freund and Robert Schapire. Large margin classification using the perceptron algorithm. In *Proceedings of COLT 1998*. ACM Press, 1998.

References

- Ehud Lehrer. Correlated equilibria in two-player repeated games with nonobservable actions. *Mathematics of Operations Research*, 17(1), 1992.
- Nick Littlestone. From On-Line to Batch Learning. *COLT* 1989: 269-284.
- Nick Littlestone, Manfred K. Warmuth: The Weighted Majority Algorithm. *FOCS* 1989: 256-261.
- Mehryar Mohri and Scott Yang. Conditional swap regret and conditional correlated equilibrium. In *NIPS*. 2014.
- Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.
- John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

References

- Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1), 2007.
- Andrew Yao. Probabilistic computations: Toward a unified measure of complexity, in FOCS, pp. 222–227, 1977.

Advanced Machine Learning

Learning with Large Expert Spaces

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Problem

- Learning guarantees:

$$R_T = O(\sqrt{T \log N}).$$

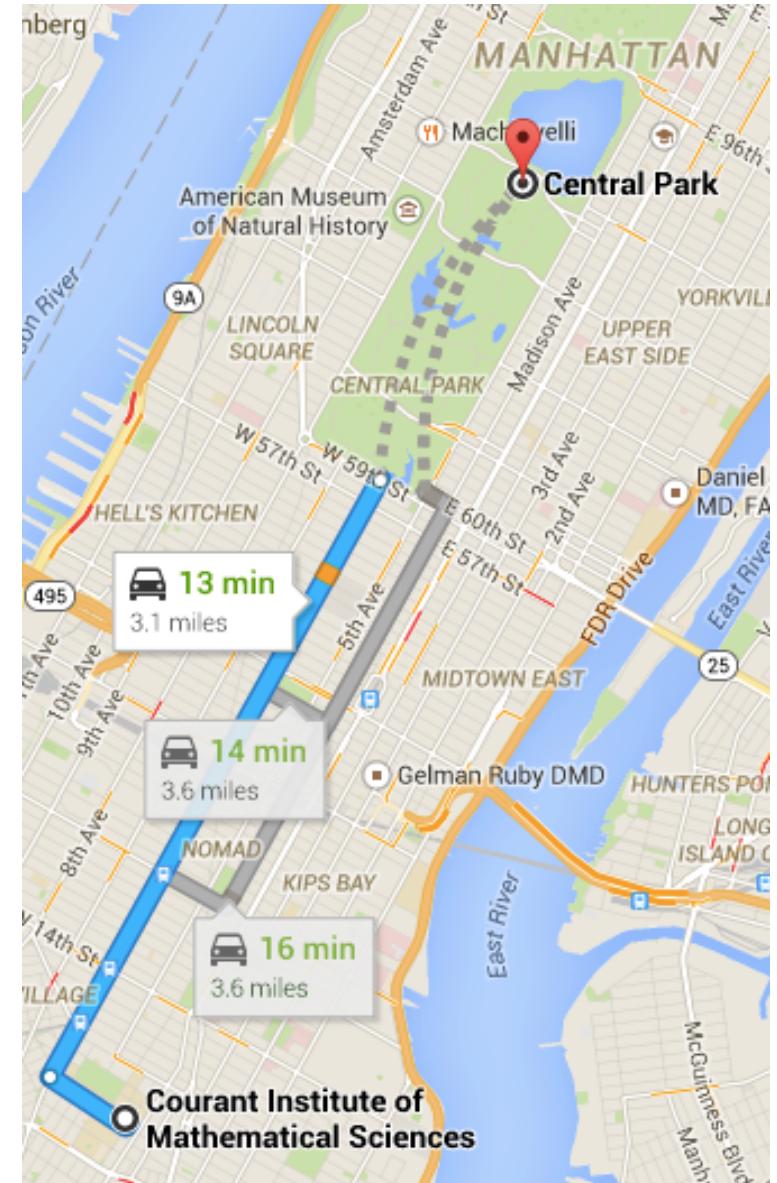
→ informative even for N very large.

- Problem: computational complexity of algorithm in $O(N)$.
Can we derive more efficient algorithms when experts admit some structure and when loss is decomposable?

Example: Online Shortest Path

■ Problems: path experts.

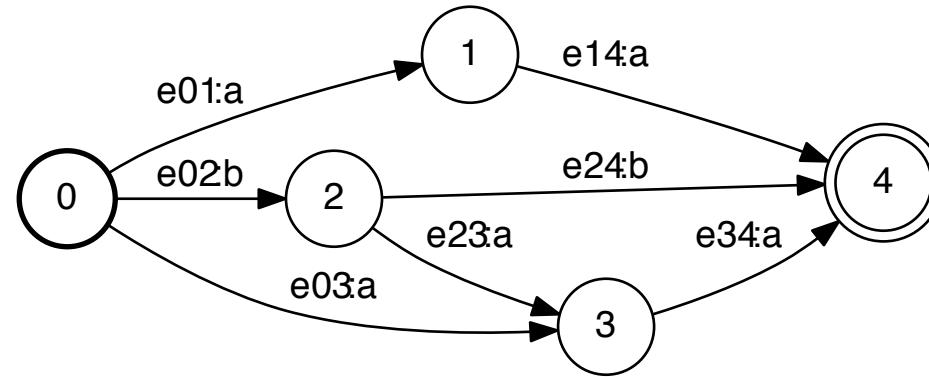
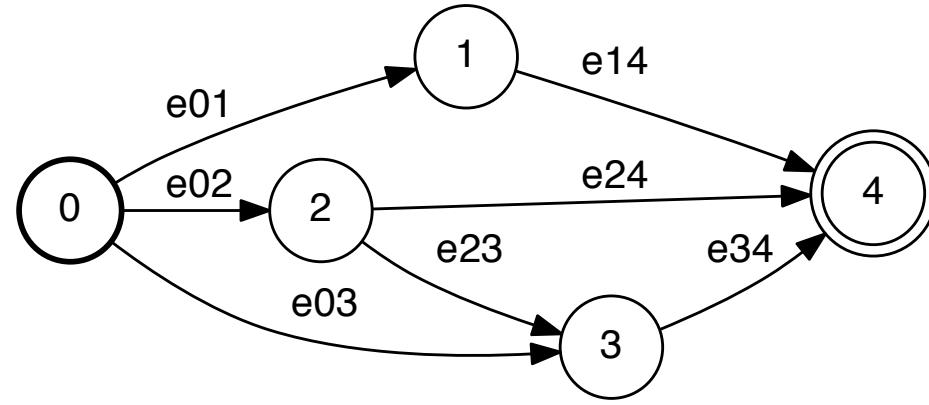
- sending packets along paths of a network with routers (vertices); delays (losses).
- car route selection in presence of traffic (loss).



Outline

- RWM with Path Experts
- FPL with Path Experts

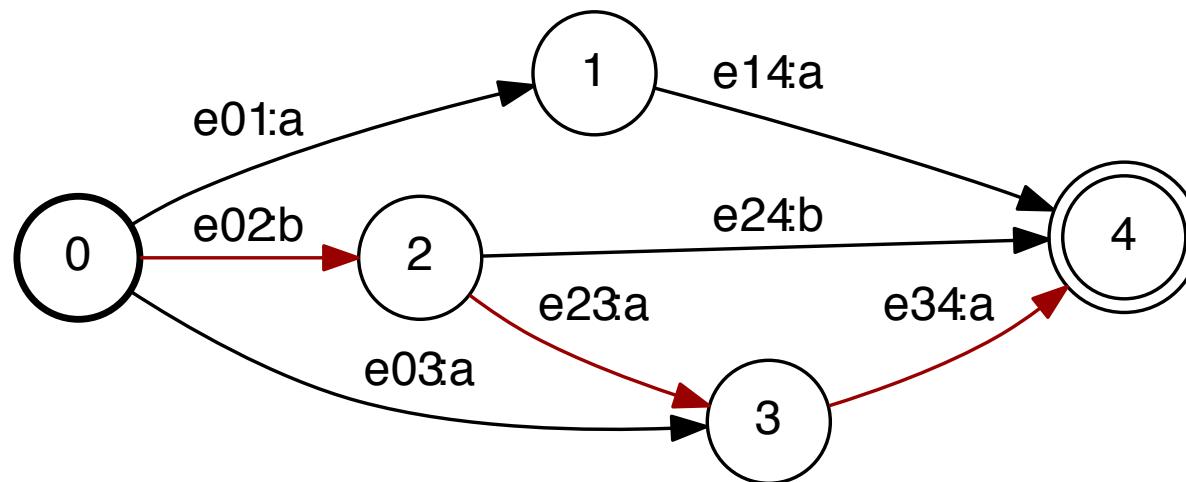
Path Experts



Additive Loss

- For path $\xi = e_{02}e_{23}e_{34}$,

$$l_t(\xi) = l_t(e_{02}) + l_t(e_{23}) + l_t(e_{34}).$$

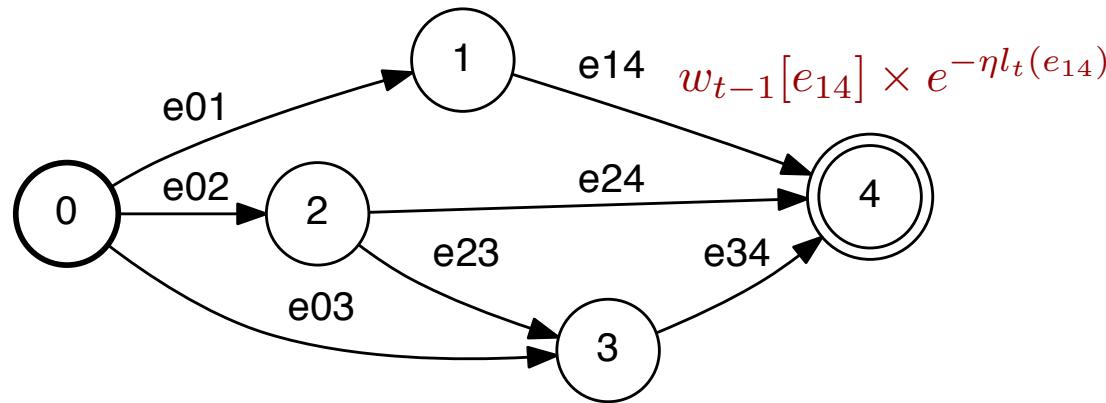


RWM + Path Experts

(Takimoto and Warmuth, 2002)

- **Weight update:** at each round t , update weight of path expert $\xi = e_1 \cdots e_n$:

- $w_t[\xi] \leftarrow w_{t-1}[\xi] e^{-\eta l_t(\xi)}$; equivalent to
- $w_t[e_i] \leftarrow w_{t-1}[e_i] e^{-\eta l_t(e_i)}$.



- **Sampling:** need to make graph/automaton stochastic.

Weight Pushing Algorithm

(MM 1997; MM, 2009)

- Weighted directed graph $G = (Q, E, w)$ with set of initial vertices $I \subseteq Q$ and final vertices $F \subseteq Q$:

- for any $q \in Q$,

$$d[q] = \sum_{\pi \in P(q, F)} w[\pi].$$

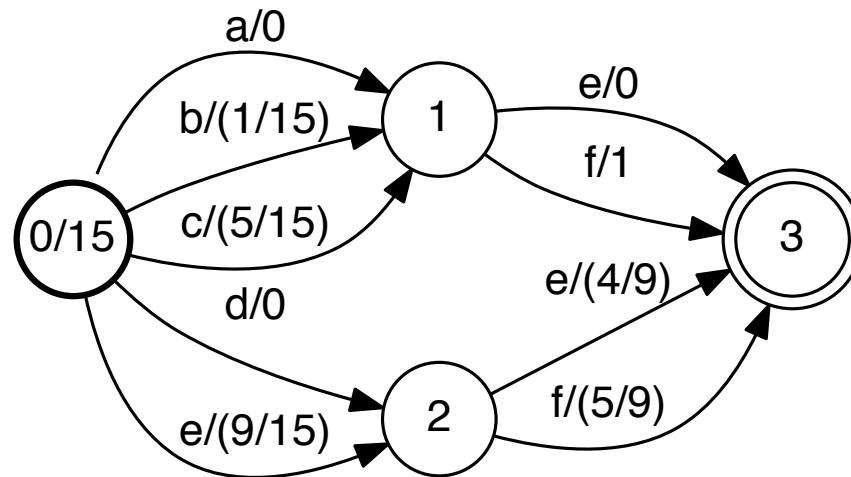
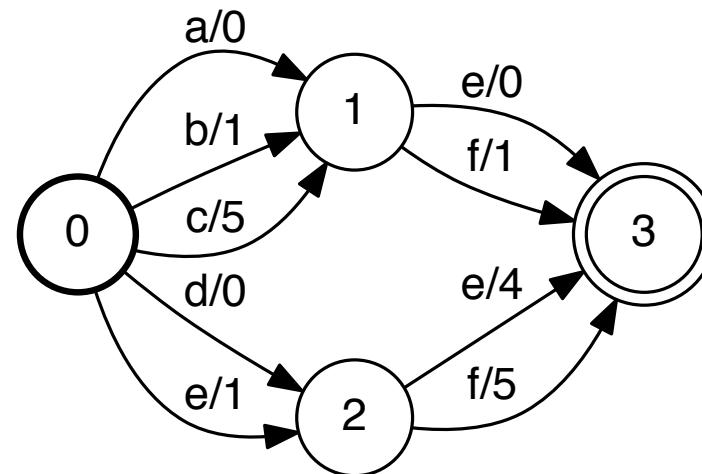
- for any $e \in E$ with $d[\text{orig}(e)] \neq 0$,

$$w[e] \leftarrow d[\text{orig}(e)]^{-1} \cdot w[e] \cdot d[\text{dest}(e)].$$

- for any $q \in I$, initial weight

$$\lambda(q) \leftarrow d(q).$$

Illustration



Properties

- **Stochasticity:** for any $q \in Q$ with $d[q] \neq 0$,

$$\sum_{e \in E[q]} w'[e] = \sum_{e \in E[q]} \frac{w[e] d[\text{dest}(e)]}{d[q]} = \frac{d[q]}{d[q]} = 1.$$

- **Invariance:** path weight preserved. Weight of path $\xi = e_1 \cdots e_n$ from I to F :

$$\begin{aligned} & \lambda(\text{orig}(e_1)) w'[e_1] \cdots w'[e_n] \\ &= d[\text{orig}(e_1)] \frac{w[e_1] d[\text{dest}(e_1)]}{d[\text{orig}(e_1)]} \frac{w[e_2] d[\text{dest}(e_2)]}{d[\text{dest}(e_1)]} \cdots \\ &= w[e_1] \cdots w[e_n] d[\text{dest}(e_n)] \\ &= w[e_1] \cdots w[e_n] = w[\xi]. \end{aligned}$$

Shortest-Distance Computation

■ Acyclic case:

- special instance of a generic single-source shortest-distance algorithm working with an arbitrary queue discipline and any k -closed semiring (MM, 2002).
- linear-time algorithm with the topological order queue discipline, $O(|Q| + |E|)$.

Generic Single-Source SD Algo.

(MM, 2002)

GEN-SINGLE-SOURCE(G, s)

```
1  for  $i \leftarrow 1$  to  $|Q|$  do
2       $d[i] \leftarrow r[i] \leftarrow \bar{0}$ 
3       $d[s] \leftarrow r[s] \leftarrow \bar{1}$ 
4       $\mathcal{Q} \leftarrow \{s\}$ 
5      while  $\mathcal{Q} \neq \emptyset$  do
6           $q \leftarrow \text{HEAD}(\mathcal{Q})$ 
7          DEQUEUE( $\mathcal{Q}$ )
8           $r' \leftarrow r[q]$ 
9           $r[q] \leftarrow \bar{0}$ 
10         for each  $e \in E[q]$  do
11             if  $d[n[e]] \neq d[n[e]] \oplus (r' \otimes w[e])$  then
12                  $d[n[e]] \leftarrow d[n[e]] \oplus (r' \otimes w[e])$ 
13                  $r[n[e]] \leftarrow r[n[e]] \oplus (r' \otimes w[e])$ 
14                 if  $n[e] \notin \mathcal{Q}$  then
15                     ENQUEUE( $\mathcal{Q}, n[e]$ )
```

Shortest-Distance Computation

■ General case:

- all-pairs shortest-distance algorithm in $(+, \times)$; for all pairs of vertices (p, q) ,

$$d[p, q] = \sum_{\pi \in P(p, q)} w[\pi].$$

- generalization of Floyd-Warshall algorithm to non-idempotent semirings ([MM, 2002](#)).
- time complexity in $O(|Q|^3)$, space complexity in $O(|Q|^2)$.
- alternative: approximation using generic single-source shortest-distance algorithm ([MM, 2002](#)).

Generic All-Pairs SD Algorithm

(MM, 2002)

GEN-ALL-PAIRS(G)

```
1  for  $i \leftarrow 1$  to  $|Q|$  do
2      for  $j \leftarrow 1$  to  $|Q|$  do
3           $d[i, j] \leftarrow \bigoplus_{e \in E \cap P(i, j)} w[e]$ 
4  for  $k \leftarrow 1$  to  $|Q|$  do
5      for  $i \leftarrow 1$  to  $|Q|, i \neq k$  do
6          for  $j \leftarrow 1$  to  $|Q|, j \neq k$  do
7               $d[i, j] \leftarrow d[i, j] \oplus (d[i, k] \otimes d[k, k]^* \otimes d[k, j])$ 
8      for  $i \leftarrow 1$  to  $|Q|, i \neq k$  do
9           $d[k, i] \leftarrow d[k, k]^* \otimes d[k, i]$ 
10          $d[i, k] \leftarrow d[i, k] \otimes d[k, k]^*$ 
11          $d[k, k] \leftarrow d[k, k]^*$ 
```

In-place version.

Learning Guarantee

- **Theorem:** let \mathcal{N} be total number of path experts and M an upper bound on the loss of a path expert. Then, the (expected) regret of RWM is bounded as follows:

$$\mathcal{L}_T \leq \mathcal{L}_T^{\min} + 2M\sqrt{T \log \mathcal{N}}.$$

Exponentiated Weighted Avg

- Computation of the prediction at each round:

$$\hat{y}_t = \frac{\sum_{\xi \in P(I, F)} w_t[\xi] y_{t,\xi}}{\sum_{\xi \in P(I, F)} w_t[\xi]}.$$

- Two single-source shortest-distance computations:
 - edge weight $w_t[e]$ (denominator).
 - edge weight $w_t[e]y_t[e]$ (numerator).

FPL + Path Experts

- Weight update: at each round, update weight of edge e ,

$$w_t[e] \leftarrow w_{t-1}[e] + l_t(e).$$

- Prediction: at each round, shortest path after perturbing each edge weight:

$$w'_t[e] \leftarrow w_t[e] + p_t(e),$$

where $p_t \sim U([0, 1/\epsilon]^{|E|})$

or $p_t \sim \text{Laplacian with density } f(\mathbf{x}) = \frac{\epsilon}{2} e^{-\epsilon \|\mathbf{x}\|_1}$.

Learning Guarantees

- **Theorem:** assume that edge losses are in $[0, 1]$. Let l_{\max} be the length of the longest path from I to F and M an upper bound on the loss of a path expert. Then,
 - the (expected) regret of FPL is bounded as follows:

$$\mathbb{E}[R_T] \leq 2\sqrt{l_{\max}M|E|T} \leq 2l_{\max}\sqrt{|E|T}.$$

- the (expected) regret of FPL* is bounded as follows:

$$\begin{aligned}\mathbb{E}[R_T] &\leq 4\sqrt{\mathcal{L}_T^{\min}|E|l_{\max}(1 + \log |E|)} + 4|E|l_{\max}(1 + \log |E|) \\ &\leq 4l_{\max}\sqrt{T|E|(1 + \log |E|)} + 4|E|l_{\max}(1 + \log |E|) \\ &= O(l_{\max}\sqrt{T|E|\log |E|}).\end{aligned}$$

Proof

- For FPL, use bound of previous lectures with

$$X_1 = |E| \quad W_1 = l_{\max} \quad R = M \leq l_{\max}.$$

- For FPL*, use bound of previous lecture with

$$X_1 = |E| \quad W_1 = l_{\max} \quad N = |E|.$$

Computational Complexity

- For an acyclic graph:
 - T updates of all edge weights.
 - T runs of a linear-time single-source shortest-path.
 - overall $O(T(|Q| + |E|))$.

Extensions

- Component hedge algorithm (Koolen, Warmuth, and Kivinen, 2010):
 - optimal regret complexity: $R_T = O(M\sqrt{T \log |E|})$.
 - special instance of mirror descent.
- Non-additive losses (Cortes, Kuznetsov, MM, Warmuth, 2015):
 - extensions of RWM and FPL.
 - rational and tropical losses.

References

- Corinna Cortes, Vitaly Kuznetsov, and Mehryar Mohri. Ensemble methods for structured prediction. In ICML. 2014.
- Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Manfred K. Warmuth. On-line learning algorithms for path experts with non-additive losses. In COLT, 2015.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- T. van Erven, W. Kotlowski, and Manfred K. Warmuth. Follow the leader with dropout perturbations. In COLT, 2014.
- Adam T. Kalai, Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71(3): 291-307. 2005.
- Wouter M. Koolen, Manfred K. Warmuth, and Jyrki Kivinen. Hedging structured concepts. In COLT, pages 93-105, 2010.

References

- Nick Littlestone, Manfred K. Warmuth: The Weighted Majority Algorithm. *FOCS* 1989: 256-261.
- Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, 23:2, 1997.
- Mehryar Mohri. Semiring Frameworks and Algorithms for Shortest-Distance Problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321-350, 2002.
- Mehryar Mohri. Weighted automata algorithms. *Handbook of Weighted Automata*, Monographs in Theoretical Computer Science, pages 213-254. Springer, 2009.
- Eiji Takimoto and Manfred K. Warmuth. Path kernels and multiplicative updates. *JMLR*, 4:773-818, 2003.

Advanced Machine Learning

Online Convex Optimization

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Outline

- Online projected sub-gradient descent.
- Exponentiated Gradient (EG).
- Mirror descent.
- Dual Averaging.

Set-Up

- Convex set C .
- For $t = 1$ to T do
 - predict $\mathbf{w}_t \in C$.
 - receive convex loss function $f_t : C \rightarrow \mathbb{R}$.
 - incur loss $f_t(\mathbf{w}_t)$.
- Regret of algorithm \mathcal{A} :

$$R_T(\mathcal{A}) = \sum_{t=1}^T f_t(\mathbf{w}_t) - \inf_{\mathbf{w} \in C} \sum_{t=1}^T f_t(\mathbf{w}).$$

Online Projected Subgrad. Desc.

■ Algorithm:

- $\mathbf{w}_1 \in C$ arbitrary.
- $\mathbf{w}_{t+1} = \Pi_C[\mathbf{w}_t - \eta \delta f_t(\mathbf{w}_t)]$, where
 - Π_C is the projection over C .
 - $\delta f_t(\mathbf{w}_t) \in \partial f_t(\mathbf{w}_t)$ (sub-gradient of f_t at \mathbf{w}_t).
 - $\eta > 0$ parameter.

Analysis

(Zinkevich, 2009)

■ Assumptions:

- $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq R$ where $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w} \in C} \sum_{t=1}^T f_t(\mathbf{w})$.
- $\|\delta f_t(\mathbf{w}_t)\| \leq G$.

■ Theorem: the regret of online projected sub-gradient descent (PSGD) is bounded as follows

$$R_T(\text{PSGD}) \leq \frac{R^2}{2\eta} + \frac{\eta G^2 T}{2}.$$

Choosing η to minimize the bound gives

$$R_T(\text{PSGD}) \leq RG\sqrt{T}.$$

Proof

- The proof uses the definition of subgradient and the property of projection:

$$\begin{aligned} R_T(\text{PSGD}) &= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\ &\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) && \text{(def. of subgrad.)} \\ &= \sum_{t=1}^T \frac{1}{2\eta} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta^2 \|\delta f_t(\mathbf{w}_t)\|^2 - \|\mathbf{w}_t - \eta \delta f_t(\mathbf{w}_t) - \mathbf{w}^*\|^2 \right] \\ &\leq \sum_{t=1}^T \frac{1}{2\eta} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta^2 G^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right] && \text{(prop. of proj.)} \\ &\leq \frac{1}{2\eta} \left[\|\mathbf{w}_1 - \mathbf{w}^*\|^2 + \eta^2 G^2 T - \|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2 \right] && \text{(telescop. sum)} \\ &\leq \frac{1}{2\eta} \left[\|\mathbf{w}_1 - \mathbf{w}^*\|^2 + \eta^2 G^2 T \right] \leq \frac{1}{2\eta} \left[R^2 + \eta^2 G^2 T \right]. \end{aligned}$$

Convex Optimization

■ Application: $\min_{\mathbf{w} \in C} f(\mathbf{w})$.

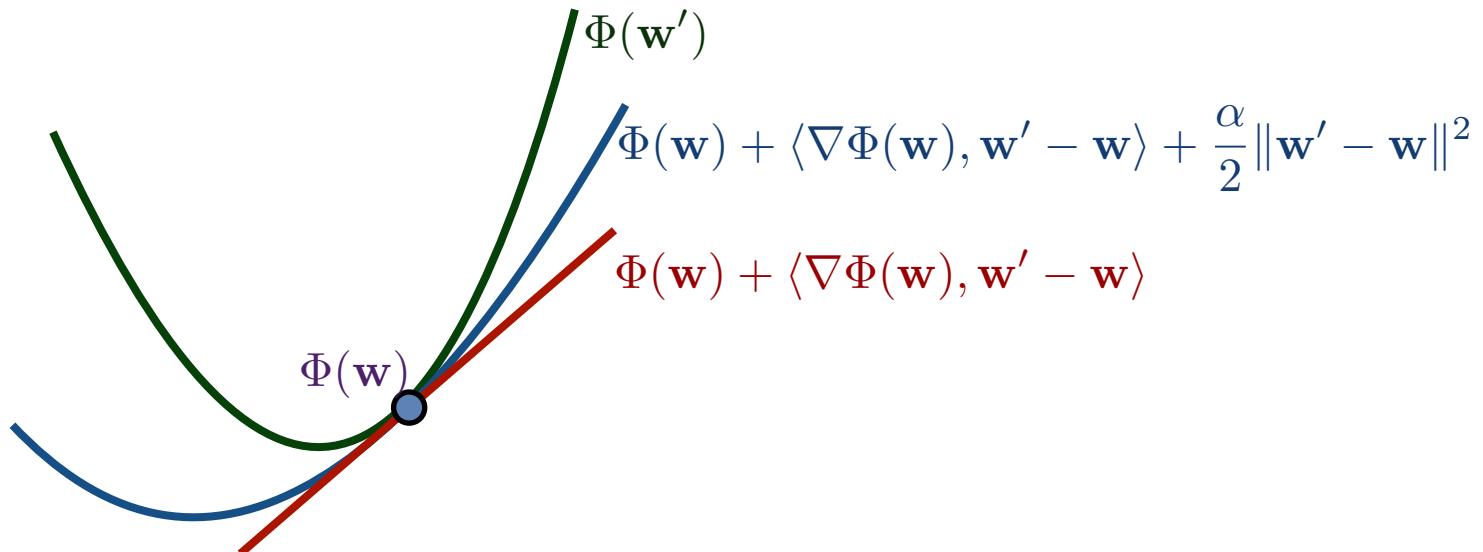
- fixed loss function: $f_t = f$.
- guarantee for average weight vector:

$$\begin{aligned} f\left(\frac{1}{T} \sum_{t=1}^T \mathbf{w}_t\right) - f(\mathbf{w}^*) &\leq \frac{1}{T} \sum_{t=1}^T f(\mathbf{w}_t) - f(\mathbf{w}^*) \\ &= \frac{R_T(\mathcal{A})}{T} = O\left(\frac{1}{\sqrt{T}}\right). \end{aligned}$$

- thus, convergence in $O\left(\frac{1}{\epsilon^2}\right)$.

Strong Convexity

- **Definition:** a convex function Φ defined over a convex set C is α -strongly convex with respect to norm $\|\cdot\|$ if the function $\mathbf{w} \mapsto \Phi(\mathbf{w}) - \frac{\alpha}{2}\|\mathbf{w}\|^2$ is convex or, equivalently,
 - for all \mathbf{w}, \mathbf{w}' in C and $\delta\Phi(\mathbf{w}) \in \partial\Phi(\mathbf{w})$,
$$\Phi(\mathbf{w}') \geq \Phi(\mathbf{w}) + \delta\Phi(\mathbf{w}) \cdot (\mathbf{w}' - \mathbf{w}) + \frac{\alpha}{2}\|\mathbf{w}' - \mathbf{w}\|^2.$$



Strongly Convex Objectives

(Hazan et al., 2007)

- **Theorem:** assume that the functions f_t are α -strongly convex and $\|\delta f_t(\mathbf{w})\| \leq G$ for all \mathbf{w} and $\delta f_t \in \partial f_t(\mathbf{w})$. Then, the regret of online projected sub-gradient descent (PSGD) with parameter $\eta_{t+1} = \frac{1}{\alpha t}$ is bounded as follows

$$R_T(\text{PSGD}) \leq \frac{G^2}{2\alpha} (1 + \log T).$$

Proof

$$R_T(\text{PSGD})$$

$$\begin{aligned}
&= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\
&\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 && \text{(strong convexity)} \\
&= \sum_{t=1}^T \frac{1}{2\eta_{t+1}} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_{t+1}^2 \|\delta f_t(\mathbf{w}_t)\|^2 - \|\mathbf{w}_t - \eta_{t+1} \delta f_t(\mathbf{w}_t) - \mathbf{w}^*\|^2 \right] - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 \\
&\leq \sum_{t=1}^T \frac{1}{2\eta_{t+1}} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_{t+1}^2 G^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right] - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 && \text{(prop. of proj.)} \\
&\leq \frac{\alpha}{2} \sum_{t=1}^T \left[(t-1) \|\mathbf{w}_t - \mathbf{w}^*\|^2 - t \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right] + \frac{G^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} && \text{(def. of } \eta_{t+1} \text{)} \\
&= \frac{\alpha}{2} \left[-T \|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2 \right] + \frac{G^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} \leq \frac{G^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} \leq \frac{G^2}{2\alpha} (1 + \log T). && \text{(telescoping sum)}
\end{aligned}$$

Smoothness

- **Definition:** a continuously differentiable function f is β -smooth if its gradient is β -Lipschitz:

$$\|\nabla f(\mathbf{w}') - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{w}' - \mathbf{w}\|,$$

for all \mathbf{w}, \mathbf{w}' .

- **Property:** if f is convex and β -smooth, then, for all \mathbf{w}, \mathbf{w}' ,

$$0 \leq f(\mathbf{w}) - f(\mathbf{w}') - \nabla f(\mathbf{w}') \cdot (\mathbf{w} - \mathbf{w}') \leq \frac{\beta}{2} \|\mathbf{w} - \mathbf{w}'\|^2.$$

Exponentiated Gradient (EG)

(Kivinen and Warmuth, 1997)

- Convex set: simplex $C = \{\mathbf{w} \in \mathbb{R}^N : \mathbf{w} \geq 0 \wedge \|\mathbf{w}\|_1 = 1\}$.
- Algorithm:

- $\mathbf{w}_1 = (\frac{1}{N}, \dots, \frac{1}{N})^\top$.

- $\mathbf{w}_{t+1,i} = \frac{\mathbf{w}_{t,i} \exp(-\eta [\delta f_t(\mathbf{w}_t)]_i)}{Z_t}$ where

$$Z_t = \sum_{i=1}^N \mathbf{w}_{t,i} e^{-\eta [\delta f_t(\mathbf{w}_t)]_i}.$$

Analysis

■ Assumption:

- $\|\delta f_t(\mathbf{w}_t)\|_\infty \leq G_\infty$.

■ Theorem: the regret of the Exponentiated Gradient (EG) algorithm is bounded as follows

$$R_T(\text{EG}) \leq \frac{\log N}{\eta} + \frac{\eta G_\infty^2 T}{2}.$$

Choosing η to minimize the bound gives

$$R_T(\text{EG}) \leq 2G_\infty \sqrt{T \log N}.$$

Proof

- Potential: $\Phi_t = D(\mathbf{w}^* \parallel \mathbf{w}_t) = \sum_{i=1}^N \mathbf{w}_i^* \log \frac{\mathbf{w}_i^*}{\mathbf{w}_{t,i}}.$
- $\Phi_{t+1} - \Phi_t = \sum_{i=1}^N \mathbf{w}_i^* \log \frac{\mathbf{w}_{t,i}}{\mathbf{w}_{t+1,i}}$
 $= \sum_{i=1}^N \mathbf{w}_i^* [\log Z_t + \eta[\delta f_t(\mathbf{w}_t)]_i] = \log Z_t + \eta \mathbf{w}^* \cdot \delta f_t(\mathbf{w}_t).$
- $\log Z_t = \log \left[\sum_{i=1}^N \mathbf{w}_{t,i} e^{-\eta[\delta f_t(\mathbf{w}_t)]_i} \right]$
 $= \log \mathbb{E}_{i \sim \mathbf{w}_t} \left[e^{-\eta[\delta f_t(\mathbf{w}_t)]_i} \right]$
 $= \log \mathbb{E}_{i \sim \mathbf{w}_t} \left[e^{-\eta([\delta f_t(\mathbf{w}_t)]_i - \mathbb{E}[[\delta f_t(\mathbf{w}_t)]_i]) - \eta \mathbb{E}[[\delta f_t(\mathbf{w}_t)]_i]} \right]$
 $\leq \eta^2 \frac{4G_\infty^2}{8} - \eta \mathbf{w}_t \cdot \delta f_t(\mathbf{w}_t). \quad (\text{Hoeffding's lemma})$

Proof

■ Combining equality and inequality:

$$\begin{aligned}\Phi_{t+1} - \Phi_t &\leq \frac{\eta^2 G_\infty^2}{2} - \eta(\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) \\ \Leftrightarrow \eta(\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) &\leq \frac{\eta^2 G_\infty^2}{2} + (\Phi_t - \Phi_{t+1}) \\ \Rightarrow \sum_{t=1}^T (\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) &\leq \frac{\eta^2 G_\infty^2 T}{2} + \frac{\Phi_1 - \Phi_{T+1}}{\eta} \\ \Rightarrow \sum_{t=1}^T (\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) &\leq \frac{\eta^2 G_\infty^2 T}{2} + \frac{\Phi_1}{\eta}. \quad (\text{Rel. Ent. non-neg.})\end{aligned}$$

$$\begin{aligned}R_T(\text{EG}) &= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\ &\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) \\ &\leq \frac{\eta G_\infty^2 T}{2} + \frac{\Phi_1}{\eta} = \frac{\eta G_\infty^2 T}{2} + \frac{D(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} \leq \frac{\eta G_\infty^2 T}{2} + \frac{\log N}{\eta}.\end{aligned}$$

Generalization

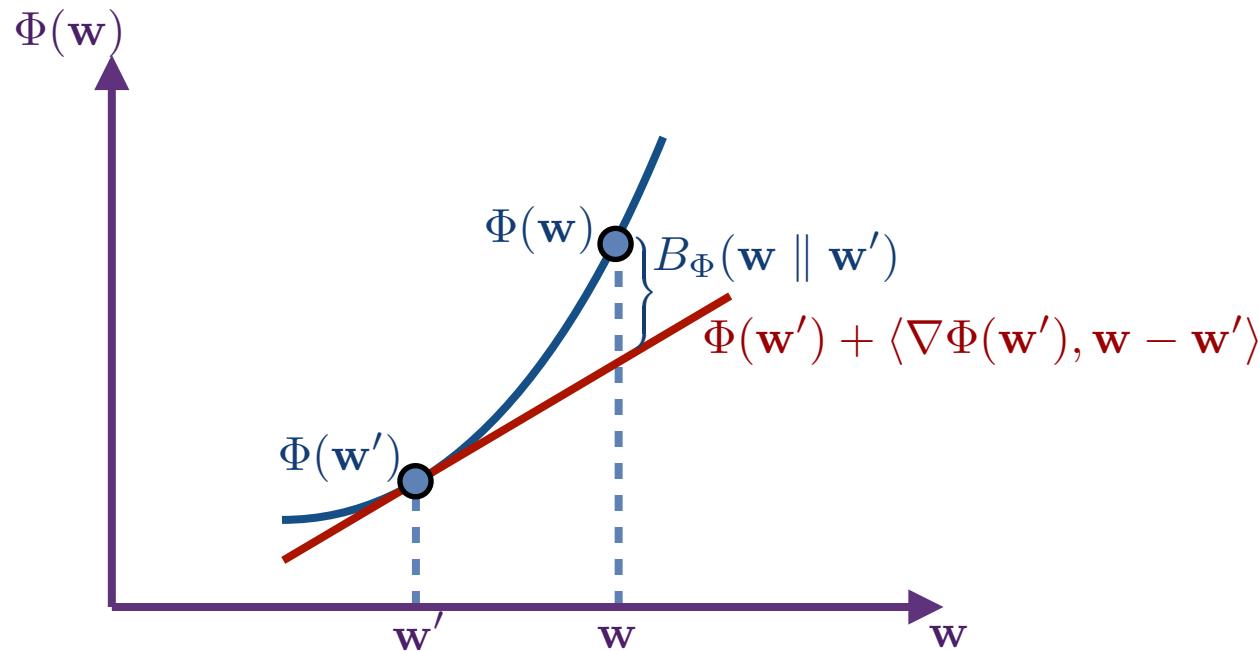
- PSGD and EG both special instances of a more general algorithm: Mirror Descent.
- Mirror Descent is based on a Bregman divergence:
 - PSGD: $B(\mathbf{w} \parallel \mathbf{w}') = \frac{1}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2$.
 - EG: unnormalized relative entropy;

$$B(\mathbf{w} \parallel \mathbf{w}') = \sum_{i=1}^N \left[w_i \log \left[\frac{w_i}{w'_i} \right] - w_i + w'_i \right].$$

Bregman Divergence

- **Definition:** Φ convex differentiable over open convex set C .
The Bregman divergence associated to Φ is defined by

$$B_{\Phi}(\mathbf{w} \parallel \mathbf{w}') = \Phi(\mathbf{w}) - \Phi(\mathbf{w}') - \langle \nabla \Phi(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle.$$



Properties

■ **Proposition:** the following properties hold for a Bregman divergence.

- non-negativity: $\forall \mathbf{w}, \mathbf{w}' \in C, B_\Phi(\mathbf{w} \parallel \mathbf{w}') \geq 0$.
- linearity: $B_{\alpha\Phi+\beta\Psi} = \alpha B_\Phi + \beta B_\Psi$.
- projection: for any closed convex set $K \subseteq \overline{C}$, the projection of B_Φ -projection of \mathbf{w}' over K is unique:

$$P_K(\mathbf{w}') = \operatorname{argmin}_{\mathbf{w} \in K} B_F(\mathbf{w} \parallel \mathbf{w}').$$

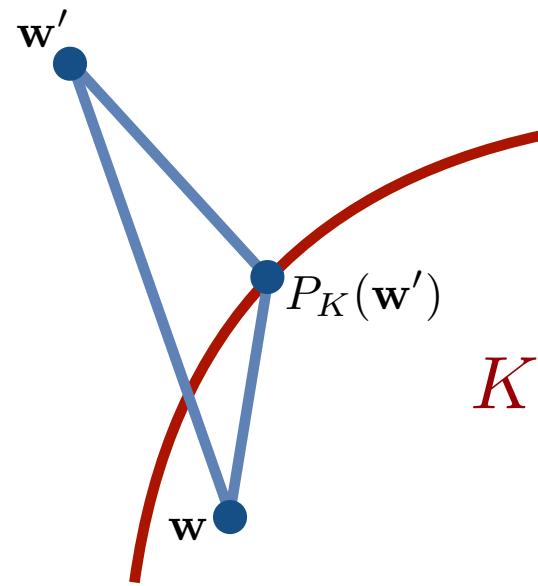
- Triangular identity:

$$(\nabla\Phi(\mathbf{w}) - \nabla\Phi(\mathbf{v})) \cdot (\mathbf{w} - \mathbf{u}) = B(\mathbf{u} \parallel \mathbf{w}) + B(\mathbf{w} \parallel \mathbf{v}) - B(\mathbf{u} \parallel \mathbf{v}).$$

- Pythagorean theorem:

$$B_\Phi(\mathbf{w} \parallel \mathbf{w}') \geq B_\Phi(\mathbf{w} \parallel P_K(\mathbf{w}')) + B_\Phi(P_K(\mathbf{w}') \parallel \mathbf{w}').$$

Pythagorean theorem



$$B_\Phi(\mathbf{w} \parallel \mathbf{w}') \geq B_\Phi(\mathbf{w} \parallel P_K(\mathbf{w}')) + B_\Phi(P_K(\mathbf{w}') \parallel \mathbf{w}').$$

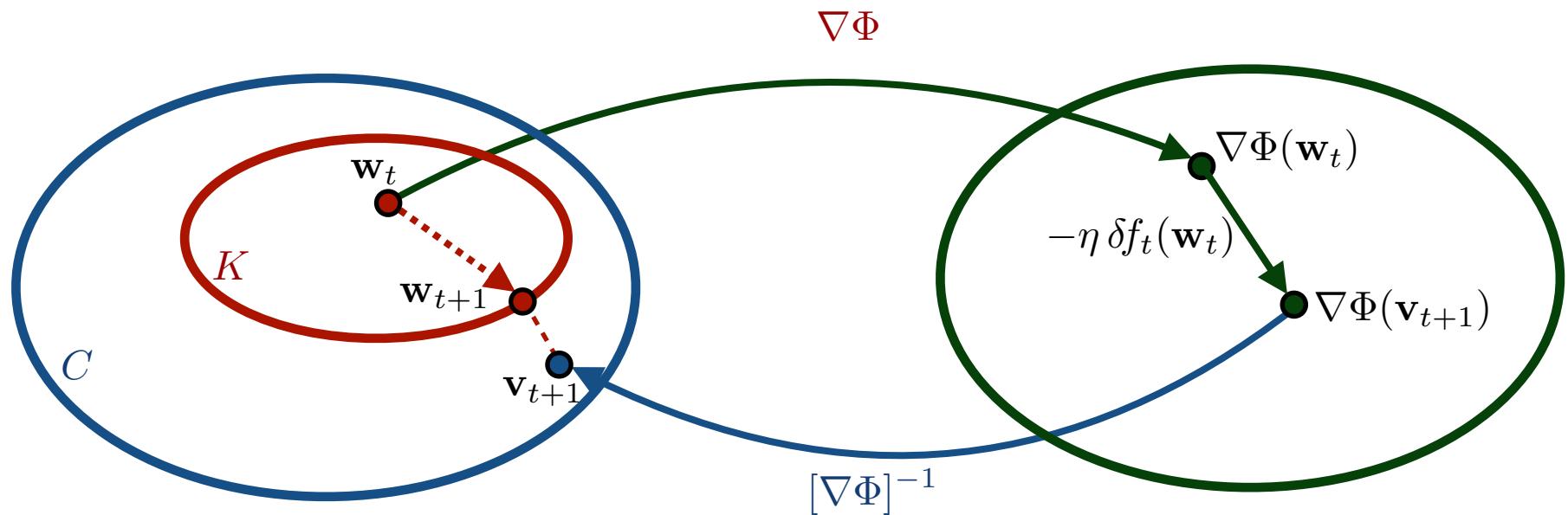
Legendre Type Functions

(Rockafellar, 1970)

- **Definition:** a real-valued function Φ defined over a non-empty open convex set C is said to be of **Legendre type** if it is proper closed convex and differentiable over C and if one of the following equivalent conditions holds:
 - $\nabla\Phi$ is one-to-one mapping from C to $\nabla\Phi(C)$.
 - $\lim_{\mathbf{w} \rightarrow \partial C} \|\nabla\Phi(\mathbf{w})\| = +\infty$.
- *proper*: $(\forall x \in C, \Phi(x) > -\infty) \wedge (\exists x_0 \in C, \Phi(x_0) < +\infty)$.
- *closed*: sublevel set $\{x \in C : \Phi(x) \leq t\}$ closed for any $t \in \mathbb{R}$.

Mirror Descent

(Nemirovski and Yudin, 1983)



Mirror Descent

MIRROR-DESCENT(Φ)

```
1    $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w})$ 
2   for  $t \leftarrow 1$  to  $T$  do
3        $\mathbf{v}_{t+1} \leftarrow [\nabla \Phi]^{-1} (\nabla \Phi(\mathbf{w}_t) - \eta \delta f_t(\mathbf{w}_t))$ 
4        $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1})$ 
```

K is a subset of C

MD Guarantee

- **Theorem:** let C be a non-empty open convex set and $K \subset \overline{C}$ a compact convex set. Assume that $\Phi: C \rightarrow \mathbb{R}$ is of Legendre type and α -strongly convex with respect to $\|\cdot\|$ and f_t 's convex and G_* -Lipschitz with respect to $\|\cdot\|_*$. Then, the regret of Mirror Descent can be bounded as follows:

$$R_T(\text{MD}) \leq \frac{B(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \frac{\eta G_*^2 T}{2\alpha}.$$

Choosing η to minimize the bound gives

$$R_T(\text{MD}) \leq D_\Phi G_* \sqrt{\frac{2T}{\alpha}},$$

with $B(\mathbf{w}^* \parallel \mathbf{w}_1) \leq D_\Phi^2$.

Proof

$$R_T(\text{MD})$$

$$\begin{aligned}
&= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\
&\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) && \text{(def. of subgrad.)} \\
&= \frac{1}{\eta} \sum_{t=1}^T [\nabla \Phi(\mathbf{w}_t) - \nabla \Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_t - \mathbf{w}^*) && \text{(def. of } \mathbf{v}_t) \\
&= \frac{1}{\eta} \sum_{t=1}^T [B(\mathbf{w}^* \parallel \mathbf{w}_t) - B(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] && \text{(Triang. Identity)} \\
&\leq \frac{1}{\eta} \sum_{t=1}^T [B(\mathbf{w}^* \parallel \mathbf{w}_t) - B(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] \\
&&& \text{(Pythagorean ineq.)} \\
&= \frac{1}{\eta} [B(\mathbf{w}^* \parallel \mathbf{w}_1) - B(\mathbf{w}^* \parallel \mathbf{w}_{T+1})] + \frac{1}{\eta} \sum_{t=1}^T [-B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] \\
&&& \text{(Telescoping sum)} \\
&\leq \frac{B(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T [B(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1})]. \\
&&& \text{(Non-negativity of Breg. div.)}
\end{aligned}$$

Proof

$$\begin{aligned} & \left[B(\mathbf{w}_t \| \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \| \mathbf{v}_{t+1}) \right] \\ &= \Phi(\mathbf{w}_t) - \Phi(\mathbf{w}_{t+1}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) \\ &\leq (\nabla \Phi(\mathbf{w}_t) - \nabla \Phi(\mathbf{v}_{t+1})) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 && (\text{α-strong convexity}) \\ &= -\eta \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 && (\text{def. of } \mathbf{v}_{t+1}) \\ &\leq \eta G_* \|\mathbf{w}_t - \mathbf{w}_{t+1}\| - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 && (G_*\text{-Lipschitzness}) \\ &\leq \frac{(\eta G_*)^2}{2\alpha}. && (\text{max. of 2nd deg. eq.}) \end{aligned}$$

Example: PSGD

- Mirror map: $\Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, clearly strongly convex with respect to $\|\cdot\|_2$.
- Bregman divergence: $B(\mathbf{w} \parallel \mathbf{w}') = \frac{1}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2$.

Example: EG

- Mirror map: $\Phi(\mathbf{w}) = \sum_{j=1}^N w_j \log w_j$, defined over \mathbb{R}_+^N , differentiable over $(\mathbb{R}_+^*)^N$.
 - thus, the negative entropy function.
 - 1-strongly convex with respect to $\|\cdot\|_1$ on the simplex:

$$\begin{aligned}\sum_{j=1}^N \left[w_j \log \frac{w_j}{w'_j} + w'_j - w_j \right] &= \sum_{j=1}^N \left[w_j \log \frac{w_j}{w'_j} \right] && (\mathbf{w} \text{ and } \mathbf{w}' \text{ in simplex}) \\ &\geq \frac{1}{2} \|\mathbf{w} - \mathbf{w}'\|_1^2. && (\text{Schützenberger-Pinsker ineq.})\end{aligned}$$

- Bregman divergence: unnormalized relative entropy defined over $(\mathbb{R}_+^*)^N$,

$$B(\mathbf{w} \parallel \mathbf{w}') = \sum_{i=1}^N \left[w_i \log \left[\frac{w_i}{w'_i} \right] - w_i + w'_i \right].$$

Example: Spectrahedron

- Mirror map: $\Phi(\mathbf{M}) = \sum_{j=1}^N \lambda_j(\mathbf{M}) \log \lambda_j(\mathbf{M})$, defined over the set of semi-definite positive symmetric matrices \mathbb{S}_+^N :
 - thus, negative von Neumann entropy.
 - $\frac{1}{2}$ -strongly convex with respect to the Shatten 1-norm

$$\|\mathbf{M}\|_{(1)} = \sum_{j=1}^N s_j(\mathbf{M}) = \sum_{j=1}^N \lambda_j(\mathbf{M}).$$

Strongly Convex Objectives

- **Theorem:** assume additionally that f_t s are σ -strongly convex with respect to Φ . Then, the regret of Mirror Descent with parameter $\eta_{t+1} = \frac{1}{\sigma t}$ can be bounded as follows:

$$R_T(\text{MD}) \leq \frac{G_*^2}{2\sigma\alpha}(1 + \log T).$$

Conjugate Functions

- **Definition:** let $\Phi: C \rightarrow \mathbb{R}$ be a convex function defined over a subset $C \subseteq \mathbb{R}^N$. Then, the conjugate function Φ^* is defined by:

$$\Phi^*(u) = \sup_{x \in C} (\langle x, u \rangle - \Phi(x)).$$

- For a Legendre function Φ , $(\nabla \Phi)^{-1} = \Phi^*$.
- For a convex function Φ taking value $+\infty$ outside a convex and compact set K , Φ not necessarily Legendre but Φ^* differentiable, a variant of MD can be used.

Proof

$$R_T(\text{MD})$$

$$\begin{aligned}
&= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\
&\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) - \sigma B(\mathbf{w}^* \parallel \mathbf{w}_t) \quad (\Phi\text{-strong convexity}) \\
&= \sum_{t=1}^T \frac{1}{\eta_{t+1}} [\nabla \Phi(\mathbf{w}_t) - \nabla \Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_t - \mathbf{w}^*) - \sigma B(\mathbf{w}^* \parallel \mathbf{w}_t) \quad (\text{Def. of } \mathbf{v}_t) \\
&= \sum_{t=1}^T \frac{1}{\eta_{t+1}} [B(\mathbf{w}^* \parallel \mathbf{w}_t) - B(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] - \sigma B(\mathbf{w}^* \parallel \mathbf{w}_t) \\
&\quad (\text{Breg. div. Identity}) \\
&\leq \sum_{t=1}^T \frac{1}{\eta_{t+1}} [B(\mathbf{w}^* \parallel \mathbf{w}_t) - B(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] - \sigma B(\mathbf{w}^* \parallel \mathbf{w}_t) \\
&\quad (\text{Pyth. ineq.}) \\
&= \sigma \sum_{t=1}^T [(t-1)B(\mathbf{w}^* \parallel \mathbf{w}_t) - tB(\mathbf{w}^* \parallel \mathbf{w}_{t+1})] + \sum_{t=1}^T \frac{1}{\eta_{t+1}} [-B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] \\
&\quad (\text{Def. of } \eta_{t+1}) \\
&\leq -\sigma T B(\mathbf{w}^* \parallel \mathbf{w}_{T+1}) + \sum_{t=1}^T \frac{1}{\eta_{t+1}} [B(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1})] \\
&\quad (\text{Telescoping sum}) \\
&\leq \sum_{t=1}^T \frac{1}{\eta_{t+1}} [B(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1})]. \quad (\text{Non-negativity of Breg. div.})
\end{aligned}$$

Proof

$$\begin{aligned}
& \left[B(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) \right] \\
&= \Phi(\mathbf{w}_t) - \Phi(\mathbf{w}_{t+1}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) \quad (\text{Def. of Breg. div.}) \\
&\leq (\nabla \Phi(\mathbf{w}_t) - \nabla \Phi(\mathbf{v}_{t+1})) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \\
&\hspace{400pt} (\alpha\text{-strong convexity}) \\
&= -\eta_{t+1} \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \quad (\text{Def. of } \mathbf{v}_{t+1}) \\
&\leq \eta_{t+1} G_* \|\mathbf{w}_t - \mathbf{w}_{t+1}\| - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \quad (G_*\text{-Lipschitzness}) \\
&\leq \frac{(\eta_{t+1} G_*)^2}{2\alpha}. \quad (\text{Max. of 2nd deg. polynomial})
\end{aligned}$$

Thus,

$$\sum_{t=1}^T \frac{1}{\eta_{t+1}} \left[B(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) \right] \leq \frac{G_*^2}{2\alpha\sigma} \sum_{t=1}^T \frac{1}{t} \leq \frac{G_*^2}{2\alpha\sigma} (1 + \log T).$$

Mirror Descent (duplicate)

MIRROR-DESCENT(Φ)

```
1  $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w})$ 
2 for  $t \leftarrow 1$  to  $T$  do
3    $\mathbf{v}_{t+1} \leftarrow [\nabla \Phi]^{-1} (\nabla \Phi(\mathbf{w}_t) - \eta \delta f_t(\mathbf{w}_t))$ 
4    $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1})$ 
```

Equivalent Description

MIRROR-DESCENT(Φ)

```
1    $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w})$ 
2   for  $t \leftarrow 1$  to  $(T - 1)$  do
3        $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \delta f_t(\mathbf{w}_t) \cdot \mathbf{w} + \frac{1}{\eta} B(\mathbf{w} \parallel \mathbf{w}_t)$ 
```

linearization of f_t regularization

■ Proof:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot \mathbf{w} \quad (\text{def. of Breg. div.}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - (\nabla \Phi(\mathbf{w}_t) - \eta \delta f_t(\mathbf{w}_t)) \cdot \mathbf{w} \quad (\text{def. of } \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \eta \delta f_t(\mathbf{w}_t) \cdot \mathbf{w} + B(\mathbf{w} \parallel \mathbf{w}_t). \quad (\text{def. of Breg. div.})\end{aligned}$$

Dual Averaging

(Iouditski and Nesterov, 2010)

DUAL-AVERAGING(Φ)

```
1    $\mathbf{v}_1 \leftarrow 0$ 
2    $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_1)$ 
3   for  $t \leftarrow 1$  to  $T$  do
4        $\mathbf{v}_{t+1} \leftarrow [\nabla \Phi]^{-1} (\underline{\nabla \Phi(\mathbf{v}_t)} - \eta \delta f_t(\mathbf{w}_t))$ 
5        $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1})$ 
```

Equivalently:

$$\mathbf{v}_{t+1} \leftarrow [\nabla \Phi]^{-1} \left(\nabla \Phi(\mathbf{w}_1) - \eta \sum_{s=1}^t \delta f_s(\mathbf{w}_s) \right)$$

Equivalent Description

■ Equivalent form:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \| \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot \mathbf{w} \quad (\text{def. of Breg. div.}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - (\nabla \Phi(\mathbf{v}_t) - \eta \delta f_t(\mathbf{w}_t)) \cdot \mathbf{w} \quad (\text{def. of } \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \eta \sum_{s=1}^t \delta f_t(\mathbf{w}_s) + \Phi(\mathbf{w}). \quad (\text{recurrence})\end{aligned}$$

■ In particular, for linear losses, $f_t(\mathbf{w}) = \mathbf{a}_t \cdot \mathbf{w}$, Dual Averaging coincides with **regularized FL (FTRL)**:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in K \cap C} \sum_{s=1}^t \mathbf{a}_s \cdot \mathbf{w} + \frac{1}{\eta} \Phi(\mathbf{w}).$$

FTRL: Follow the Regularized Leader

MD Guarantee (duplicate)

- **Theorem:** let C be a non-empty open convex set and $K \subset \overline{C}$ a compact convex set. Assume that $\Phi: C \rightarrow \mathbb{R}$ is of Legendre type and α -strongly convex with respect to $\|\cdot\|$ and f_t 's convex and G_* -Lipschitz with respect to $\|\cdot\|_*$. Then, the regret of Mirror Descent can be bounded as follows:

$$R_T(\text{MD}) \leq \frac{B(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \frac{\eta G_*^2 T}{2\alpha}.$$

Choosing η to minimize the bound gives

$$R_T(\text{MD}) \leq D_\Phi G_* \sqrt{\frac{2T}{\alpha}},$$

with $B(\mathbf{w}^* \parallel \mathbf{w}_1) \leq D_\Phi^2$.

DA Guarantee

- **Theorem:** under the same assumptions as for MD, the following holds for the regret of Dual Averaging,

$$R_T(\text{DA}) \leq \frac{\Phi(\mathbf{w}^*) - \Phi(\mathbf{w}_1)}{\eta} + \frac{2\eta G_*^2 T}{\alpha}.$$

Choosing η to minimize the bound gives

$$R_T(\text{DA}) \leq 2D_\Phi G_* \sqrt{\frac{2T}{\alpha}},$$

with $\Phi(\mathbf{w}^*) - \Phi(\mathbf{w}_1) \leq D_\Phi^2$.

References

- Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. On-line convex optimization in the bandit setting: gradient descent without a gradient. In SODA, pages 385–394. SIAM, 2005.
- Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. Comput. Syst. Sci.*, 74(1):97–114, 2008.
- Amir Beck and Marc Teboulle. Mirror Descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- A .J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

References

- Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
- Anatoli Iouditski, Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. 2010. <hal-00508933v1>
- Adam T. Kalai, Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71(3): 291-307. 2005.
- Jyrki Kivinen and Manfred K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, 1997.
- Jyrki Kivinen and Manfred K. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45(3):301–329, 2001.
- Yurii Nesterov. Introductory lectures on convex optimization: A basic course. Kluwer Academic Publishers, 2004a.
- R. Tyrrell Rockafellar. Convex Analysis. Princeton University Press, 1970.

References

- Arkadii Semenovich Nemirovski, David Berkovich Yudin. Problem complexity and Method Efficiency in Optimization, Wiley, New York, 1983.
- Eiji Takimoto and Manfred K. Warmuth. Path kernels and multiplicative updates. JMLR, 4:773–818, 2003.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In ICML, pages 928–936, 2009.

Advanced Machine Learning

Bandit Problems

MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Multi-Armed Bandit Problem

- **Problem:** which arm of a K -slot machine should a gambler pull to maximize his cumulative reward over a sequence of trials?
 - stochastic setting.
 - adversarial setting.



Motivation

- Clinical trials: potential treatments for a disease to select from, new patient or category at each round (Thompson, 1933).
- Ads placement: selection of ad to display out of a finite set (which could vary with time though) for each new web page visitor.
- Adaptive routing: alternative paths for routing packets through a “series of tubes” or alternative roads for driving from a source to a destination.
- Games: different moves at each round of a game such as chess, or Go.

Key Problem

- Exploration vs exploitation dilemma (or trade-off):
 - inspect new arms with possibly better rewards.
 - use existing information to select best arm.

Outline

- Stochastic bandits
- Adversarial bandits

Stochastic Model

- K arms: for each arm $i \in \{1, \dots, K\}$,
 - reward distribution P_i .
 - reward mean μ_i .
 - gap to best: $\Delta_i = \mu^* - \mu_i$, where $\mu^* = \max_{i \in [1, K]} \mu_i$.

Bandit Setting

- For $t = 1$ to T do
 - player selects action $I_t \in \{1, \dots, K\}$ (randomized).
 - player receives reward $X_{I_t, t} \sim P_{I_t}$.
- Equivalent descriptions:
 - on-line learning with partial information (\neq full).
 - one-state MDPs (Markov Decision Processes).

Objectives

■ Expected regret

$$\mathbb{E}[R_T] = \mathbb{E} \left[\max_{i \in [1, K]} \sum_{t=1}^T X_{i,t} - \sum_{t=1}^T X_{I_t,t} \right].$$

■ Pseudo-regret

$$\begin{aligned} \bar{R}_T &= \max_{i \in [1, K]} \mathbb{E} \left[\sum_{t=1}^T X_{i,t} - \sum_{t=1}^T X_{I_t,t} \right]. \\ &= \mu^* T - \mathbb{E} \left[\sum_{t=1}^T X_{I_t,t} \right]. \end{aligned}$$

■ By Jensen's inequality, $\bar{R}_T \leq \mathbb{E}[R_T]$.

Expected Regret

- If $(X_{i,t} - \mu_i)$ s take values in $[-r, +r]$, then

$$\mathbb{E} \left[\max_{i \in [1, K]} \sum_{t=1}^T (X_{i,t} - \mu^*) \right] \leq r \sqrt{2T \log K}.$$

- The $O(\sqrt{T})$ dependency cannot be improved;
→ better guarantees can be achieved for pseudo-regret.

Pseudo-Regret

- Expression in terms of Δ_i s:

$$\bar{R}_T = \sum_{i=1}^K \mathbb{E}[T_i(T)]\Delta_i ,$$

where $T_i(t)$ denotes the number of times arm i was pulled up to time t , $T_i(t) = \sum_{s=1}^t 1_{I_s=i}$.

- Proof:
$$\begin{aligned} \bar{R}_T &= \mu^* T - \mathbb{E} \left[\sum_{t=1}^T X_{I_t,t} \right] = \mathbb{E} \left[\sum_{t=1}^T (\mu^* - X_{I_t,t}) \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^K (\mu^* - X_{i,t}) 1_{I_t=i} \right] = \sum_{t=1}^T \sum_{i=1}^K \mathbb{E}[(\mu^* - X_{i,t})] \mathbb{E}[1_{I_t=i}] \\ &= \sum_{i=1}^K (\mu^* - \mu_i) \mathbb{E} \left[\sum_{t=1}^T 1_{I_t=i} \right] = \sum_{i=1}^K \mathbb{E}[T_i(T)]\Delta_i. \end{aligned}$$

ϵ -Greedy Strategy

(Auer et al. 2002a)

- At time t ,
 - with probability $1 - \epsilon_t$, select arm i with best emp. mean.
 - with probability ϵ_t , select random arm.
- For $\epsilon_t = \min\left(\frac{6K}{\Delta^2 t}, 1\right)$, with $\Delta = \min_{i: \Delta_i > 0} \Delta_i$,
 - for $t \geq \frac{6K}{\Delta^2}$, $\Pr[I_t \neq i^*] \leq \frac{C}{\Delta^2 t}$ for some $C > 0$.
 - thus, $E[T_i(T)] \leq \frac{C}{\Delta^2} \log T$ and $\bar{R}_T \leq \sum_{i: \Delta_i > 0} \frac{C \Delta_i}{\Delta^2} \log T$.
- Logarithmic regret but,
 - requires knowledge of Δ .
 - sub-optimal arms treated similarly (naive search).

UCB Strategy

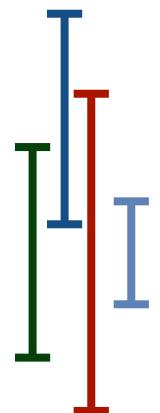
(Lai and Robbins, 1985; Agrawal 1995; Auer et al. 2002a)

■ Optimism in face of uncertainty:

- at each time $t \in [1, T]$ compute upper confidence bound (UCB) on the expected reward of each arm $i \in [1, K]$.
- select arm with largest UCB.

■ Idea: wrong arm i cannot be selected for too long.

- by definition, $\mu_i \leq \mu^* \leq \text{UCB}_i$.
- pulling i often → UCB closer to μ_i .



Upper Confidence Bound

Note on Concentration Ineqs

- Let X be a random variable such that for all $t \geq 0$,

$$\log \mathbb{E} [e^{t(X - \mathbb{E}[X])}] \leq \Psi(t),$$

where Ψ is a convex function. For Hoeffding's inequality and $X \in [a, b]$, $\Psi(t) = \frac{t^2(b-a)^2}{8}$.

- Then, $\Pr[X - \mathbb{E}[X] > \epsilon] = \Pr[e^{t(X - \mathbb{E}[X])} > e^{t\epsilon}]$
$$\leq \inf_{t>0} e^{-t\epsilon} \mathbb{E}[e^{t(X - \mathbb{E}[X])}]$$
$$\leq \inf_{t>0} e^{-t\epsilon} e^{\Psi(t)}$$
$$= e^{-\sup_{t>0} (t\epsilon - \Psi(t))}$$
$$= e^{-\Psi^*(\epsilon)} .$$

UCB Strategy

- Average reward estimate for arm i by time t :

$$\hat{\mu}_{i,t} = \frac{1}{T_i(t)} \sum_{s=1}^t X_{i,s} 1_{I_s=i}.$$

- Concentration inequality (e.g., Hoeffding's ineq.):

$$\Pr[\mu_i - \frac{1}{t} \sum_{s=1}^t X_{i,s} > \epsilon] \leq e^{-t\psi^*(\epsilon)}.$$

- Thus, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mu_i < \frac{1}{t} \sum_{s=1}^t X_{i,s} + \psi^{*-1}\left(\frac{1}{t} \log \frac{1}{\delta}\right).$$

(α, ψ) -UCB Strategy

- Parameter $\alpha > 0$; (α, ψ) -UCB strategy consists of selecting at time t

$$I_t \in \operatorname{argmax}_{i \in [1, K]} \left[\hat{\mu}_{i,t-1} + \psi^{*-1} \left(\frac{\alpha \log t}{T_i(t-1)} \right) \right].$$

(α, ψ) -UCB Guarantee

- **Theorem:** for $\alpha > 2$, the pseudo-regret of (α, ψ) -UCB satisfies

$$\overline{R}_T \leq \sum_{i: \Delta_i > 0} \left(\frac{\alpha \Delta_i}{\psi^*(\frac{\Delta_i}{2})} \log T + \frac{\alpha}{\alpha - 2} \right).$$

- for Hoeffding's lemma, α -UCB, $\psi^*(\epsilon) = 2\epsilon^2$ (Auer et al. 2002a),

$$\overline{R}_T \leq \sum_{i: \Delta_i > 0} \left(\frac{2\alpha}{\Delta_i} \log T + \frac{\alpha}{\alpha - 2} \right).$$

Proof

■ **Lemma:** for any $s \geq 0$,

$$\sum_{t=1}^T 1_{I_t=i} \leq s + \sum_{t=s+1}^T 1_{I_t=i} 1_{T_i(t-1) \geq s}.$$

■ **Proof:** observe that

$$\sum_{t=1}^T 1_{I_t=i} = \sum_{t=1}^T 1_{I_t=i} 1_{T_i(t-1) < s} + \sum_{t=1}^T 1_{I_t=i} 1_{T_i(t-1) \geq s}.$$

- Now, for $t^* = \max \{t \leq T : 1_{T_i(t-1) < s} \neq 0\}$,

$$\sum_{t=1}^T 1_{I_t=i} 1_{T_i(t-1) < s} = \sum_{t=1}^{t^*} 1_{I_t=i} 1_{T_i(t-1) < s}.$$

- By definition of t^* , the number of non-zero terms in the sum is at most s .

Proof

- For any i and t define $\eta_{i,t-1} = \psi^{*-1}\left(\frac{\alpha \log t}{T_i(t-1)}\right)$. At time t , if i is selected, then

$$(\hat{\mu}_{i,t-1} + \eta_{i,t-1}) - (\hat{\mu}_{i^*,t} + \eta_{i^*,t-1}) \geq 0$$

$$\Leftrightarrow [\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1}] + [2\eta_{i,t-1} - \Delta_i] + [\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1}] \geq 0.$$

Thus, at least one of these three terms is non-negative. Also, if one is non-positive, at least one of the other two is non-negative.

Proof

- To bound the pseudo-regret, we bound $E[T_i(T)]$. But, observe first that

$$T_i(t-1) \geq s = \left\lceil \frac{\alpha \log T}{\psi^*(\frac{\Delta_i}{2})} \right\rceil \geq \frac{\alpha \log t}{\psi^*(\frac{\Delta_i}{2})} \Rightarrow \Delta_i - 2\eta_{i,t-1} \geq 0.$$

- Thus,

$$\begin{aligned} E[T_i(T)] &= E \left[\sum_{t=1}^T 1_{I_t=i} \right] \\ &\leq s + E \left[\sum_{t=s+1}^T 1_{I_t=i} 1_{T_i(t-1) \geq s} \right] \\ &\leq s + \sum_{t=s+1}^T \Pr[\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1} \geq 0] + \Pr[\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1} \geq 0]. \end{aligned}$$

Proof

- Each of the two probability terms can be bounded as follows using the union bound:

$$\begin{aligned} & \Pr[\mu^* - \hat{\mu}_{i^*, t-1} - \eta_{i^*, t-1} \geq 0] \\ & \leq \Pr \left[\exists s \in [1, t]: \mu^* - \frac{1}{s} \sum_{k=1}^s X_{i,k} - \psi^{*-1} \left(\frac{\alpha \log t}{s} \right) \geq 0 \right] \\ & \leq \sum_{s=1}^t \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}. \end{aligned}$$

- Final constant of the bound obtained by further simple calculations.

Lower Bound

(Lai and Robbins, 1985)

- **Theorem:** for any strategy such that $E[T_i(T)] = o(T^\beta)$ for any arm i and any $\beta > 0$ for any set of Bernoulli reward distributions, the following holds for all Bernoulli reward distributions:

$$\liminf_{T \rightarrow +\infty} \frac{\bar{R}_T}{\log T} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{D(\mu_i \parallel \mu^*)}.$$

- a more general result holds for general distributions.

Notes

■ Observe that

$$\sum_{i: \Delta_i > 0} \frac{\Delta_i}{D(\mu_i \| \mu^*)} \geq \mu^*(1 - \mu^*) \sum_{i: \Delta_i > 0} \frac{1}{\Delta_i},$$

$$\begin{aligned} \text{since } D(\mu_i \| \mu^*) &= \mu_i \log \frac{\mu_i}{\mu^*} + (1 - \mu_i) \log \frac{1 - \mu_i}{1 - \mu^*} \\ &\leq \mu_i \frac{\mu_i - \mu^*}{\mu^*} + (1 - \mu_i) \frac{\mu^* - \mu_i}{1 - \mu^*} \\ &= \frac{(\mu_i - \mu^*)^2}{\mu^*(1 - \mu^*)} = \frac{\Delta_i^2}{\mu^*(1 - \mu^*)}. \end{aligned}$$

Outline

- Stochastic bandits
- Adversarial bandits

Adversarial Model

- K arms: for each arm $i \in \{1, \dots, K\}$,
 - no stochastic assumption.
 - rewards in $[0, 1]$.

Bandit Setting

- For $t=1$ to T do
 - player selects action $I_t \in \{1, \dots, K\}$ (randomized).
 - player receives reward $x_{I_t, t}$.
- Notes:
 - rewards $x_{i,t}$ for all arms determined by adversary simultaneously with the selection I_t of an arm by player.
 - adversary **oblivious** or **nonoblivious** (or **adaptive**).
 - strategies: deterministic, regret of at least $\frac{T}{2}$ for some (bad) sequences, thus must consider randomization.

Scenarios

■ Oblivious case:

- adversary rewards selected independently of the player's actions; thus, reward vector at time t only a function of t .

■ Non-oblivious case:

- adversary rewards at time t function of the player's past actions I_1, \dots, I_{t-1} .
- notion of regret problematic: cumulative reward compared to a quantity that depends on the player's actions! (single best action in hindsight function of actions I_1, \dots, I_T played; playing that single "best" action could have resulted in different rewards.) Policy Regret to help!

Objectives

- Minimize regret ($\ell_{i,t} = 1 - x_{i,t}$), expectation or high prob.:

$$R_T = \max_{i \in [1, K]} \sum_{t=1}^T x_{i,t} - \sum_{t=1}^T x_{I_t,t} = \sum_{t=1}^T \ell_{I_t,t} - \min_{i \in [1, K]} \sum_{t=1}^T \ell_{i,t}.$$

- Pseudo-regret:

$$\bar{R}_T = \mathbb{E} \left[\sum_{t=1}^T \ell_{I_t,t} \right] - \min_{i \in [1, K]} \mathbb{E} \left[\sum_{t=1}^T \ell_{i,t} \right].$$

- By Jensen's inequality, $\bar{R}_T \leq \mathbb{E}[R_T]$.

Importance Weighting

- In the bandit setting, the cumulative loss of each arm is not observed, so how should we update the probabilities?
- Estimates via surrogate loss:

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} \mathbf{1}_{I_t=i},$$

where $\mathbf{p}_t = (p_{1,t}, \dots, p_{K,t})$ is the probability distribution the player uses at time t to draw an arm ($p_{i,t} > 0$).

- Unbiased estimate: for any i ,

$$\mathbb{E}_{I_t \sim \mathbf{p}_t} [\tilde{\ell}_{i,t}] = \sum_{j=1}^K p_{j,t} \frac{\ell_{i,t}}{p_{i,t}} \mathbf{1}_{j=i} = \ell_{i,t}.$$

EXP3

EXP3(K)

(Auer et al. 2002b)

```
1    $\mathbf{p}_1 \leftarrow (\frac{1}{K}, \dots, \frac{1}{K})$ 
2    $(\tilde{L}_{1,0}, \dots, \tilde{L}_{K,0}) \leftarrow (0, \dots, 0)$ 
3   for  $t \leftarrow 1$  to  $T$  do
4       SAMPLE( $I_t \sim \mathbf{p}_t$ )
5       RECEIVE( $\ell_{I_t, t}$ )
6       for  $i \leftarrow 1$  to  $K$  do
7            $\tilde{\ell}_{i,t} \leftarrow \frac{\ell_{i,t}}{p_{i,t}} 1_{I_t=i}$ 
8            $\tilde{L}_{i,t} \leftarrow \tilde{L}_{i,t-1} + \tilde{\ell}_{i,s}$ 
9       for  $i \leftarrow 1$  to  $K$  do
10           $p_{i,t+1} \leftarrow \frac{e^{-\eta \tilde{L}_{i,t}}}{\sum_{j=1}^K e^{-\eta \tilde{L}_{j,t}}}$ 
11  return  $\mathbf{p}_{T+1}$ 
```

EXP3 (Exponential weights for Exploration and Exploitation)

EXP3 Guarantee

- **Theorem:** the pseudo-regret of EXP3 can be bounded as follows:

$$\bar{R}_T \leq \frac{\log K}{\eta} + \frac{\eta KT}{2}.$$

Choosing η to minimize the bound gives

$$\boxed{\bar{R}_T \leq \sqrt{2KT \log K}.}$$

- **Proof:** similar to that of EG, but we cannot use Hoeffding's inequality since $\tilde{\ell}_{i,t}$ is unbounded.

Proof

■ Potential: $\Phi_t = \log \sum_{i=1}^K e^{-\eta \tilde{L}_{i,t}}$.

■ Upper bound:

$$\begin{aligned}\Phi_t - \Phi_{t-1} &= \log \frac{\sum_{i=1}^K e^{-\eta \tilde{L}_{i,t}}}{\sum_{i=1}^N e^{-\eta \tilde{L}_{i,t-1}}} = \log \frac{\sum_{i=1}^K e^{-\eta \tilde{L}_{i,t-1}} e^{-\eta \tilde{\ell}_{i,t}}}{\sum_{i=1}^N e^{-\eta \tilde{L}_{i,t-1}}} \\ &= \log \left[\mathbb{E}_{i \sim \mathbf{p}_t} [e^{-\eta \tilde{\ell}_{i,t}}] \right] \\ &\leq \mathbb{E}_{i \sim \mathbf{p}_t} [e^{-\eta \tilde{\ell}_{i,t}}] - 1 \quad (\log x \leq x - 1) \\ &\leq \mathbb{E}_{i \sim \mathbf{p}_t} \left[-\eta \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \tilde{\ell}_{i,t}^2 \right] \quad (e^{-x} \leq 1 - x + \frac{x^2}{2}) \\ &= -\eta \mathbb{E}_{i \sim \mathbf{p}_t} [\tilde{\ell}_{i,t}] + \frac{\eta^2}{2} \mathbb{E}_{i \sim \mathbf{p}_t} \left[\frac{l_{i,t}^2 \mathbf{1}_{I_t=i}}{p_{i,t}^2} \right] \\ &= -\eta \ell_{I_t,t} + \frac{\eta^2}{2} \frac{l_{I_t,t}^2}{p_{I_t,t}} \leq -\eta \ell_{I_t,t} + \frac{\eta^2}{2} \frac{1}{p_{I_t,t}}.\end{aligned}$$

Proof

- Upper bound: summing up the inequalities yields

$$\mathbb{E}[\Phi_T - \Phi_0] \leq -\eta \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\sum_{t=1}^T \ell_{I_t, t} \right] + \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\sum_{t=1}^T \frac{\eta^2}{2p_{I_t, t}} \right] = -\eta \mathbb{E} \left[\sum_{t=1}^T \ell_{I_t, t} \right] + \frac{\eta^2 KT}{2}.$$

- Lower bound: for all $j \in [1, K]$,

$$\begin{aligned} \mathbb{E}[\Phi_T - \Phi_0] &= \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\log \left[\sum_{i=1}^K e^{-\eta \tilde{L}_{i,T}} \right] - \log K \right] \\ &\geq -\eta \mathbb{E}_{I_t \sim \mathbf{p}_t} [\tilde{L}_{j,T}] - \log K = -\eta \mathbb{E}_{I_t \sim \mathbf{p}_t} [L_{j,T}] - \log K. \end{aligned}$$

- Comparison:

$$\begin{aligned} \forall j \in [1, K], \quad \eta \mathbb{E} \left[\sum_{t=1}^T \ell_{I_t, t} \right] - \eta \mathbb{E}[L_{j,T}] &\leq \log K + \frac{\eta^2}{2} KT \\ \Rightarrow \bar{R}_T &\leq \frac{\log K}{\eta} + \frac{\eta KT}{2}. \end{aligned}$$

Notes

■ When T is not known:

- standard doubling trick.
- or, use $\eta_t = \sqrt{\frac{\log K}{Kt}}$, then $\bar{R}_T \leq 2\sqrt{KT \log K}$.

■ High probability bounds:

- importance weighting problem: unbounded second moment (see [\(Cortes, Mansour, MM, 2010\)](#)), $E_{i \sim p_t} [\tilde{\ell}_{i,t}^2] = \frac{\ell_{I_{t,t}}^2}{p_{I_{t,t}}}$.
- [\(Auer et al., 2002b\)](#): mixing probability with a uniform distribution to ensure a lower bound on $p_{i,t}$; but not sufficient for high probability bound.
- solution: biased estimate $\tilde{\ell}_{i,t} = \frac{\ell_{i,t} 1_{I_t=i} + \beta}{p_{i,t}}$ with $\beta > 0$ a parameter to tune.

Lower Bound

(Bubeck and Cesa-Bianchi, 2012)

- Sufficient lower bound in a stochastic setting for the pseudo-regret (and therefore for the expected regret).
- **Theorem:** for any $T \geq 1$ and any player strategy, there exists a distribution of losses in $\{0, 1\}$ for which

$$\overline{R}_T \geq \frac{1}{20} \sqrt{KT}.$$

Notes

- Bound of EXP3 matching lower bound modulo Log term.
- Log-free bound: $p_{i,t+1} = \psi(C_t - \tilde{L}_{i,t})$ where C_t is a constant ensuring $\sum_{i=1}^K p_{i,t+1} = 1$ and ψ increasing, convex, twice differentiable over \mathbb{R}^* (Audibert and Bubeck, 2010).
 - EXP3 coincides with $\psi(x) = e^{\eta x}$.
 - log-free bound with $\psi(x) = (-\eta x)^{-q}$ and $q = 2$.
 - formulation as mirror descent.
 - only in oblivious case.

References

- R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Mathematics*, vol. 27, pp. 1054–1078, 1995.
- Jean-Yves Audibert and Sebastian Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, vol. 11, pp. 2635– 2686, 2010.
- Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi- armed bandit problem, *Machine Learning Journal*, vol. 47, no. 2–3, pp. 235– 256, 2002a.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002b.
- Sébastien Bubeck, Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5, 1-122, 2012.

References

- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In NIPS, 2010.
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.
- Gilles Stoltz. Incomplete information and internal regret in prediction of individual sequences. *Ph.D. thesis, Universite Paris-Sud*, 2005.
- R. Tyrrell Rockafellar. Convex Analysis. *Princeton University Press*, 1970.
- W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, vol. 25, pp. 285–294, 1933.

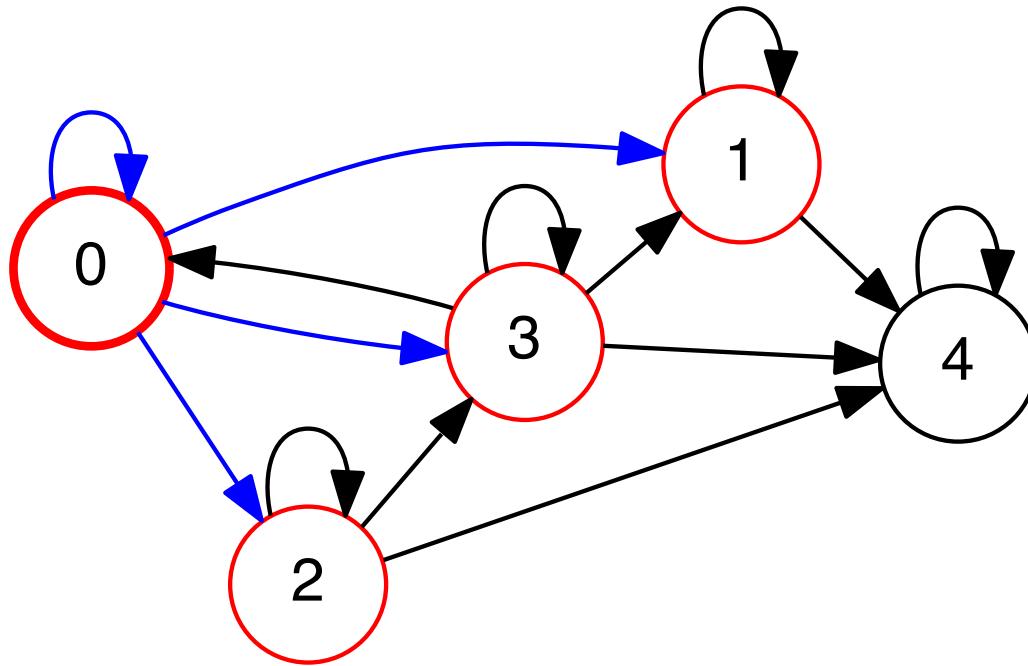
Online Learning with Feedback Graphs

MEHRYAR MOHRI MOHRI@
GOOGLE RESEARCH & COURANT INSTITUTE

Motivation

- Online learning with side observation ([Mannor and Shamir, 2011](#)):
 - side observation modeled as feedback graph.
 - full information and bandit: special cases.
 - intermediate regret guarantees expressed in terms of graph properties (mas-number, independence number, domination number).

Feedback Graph

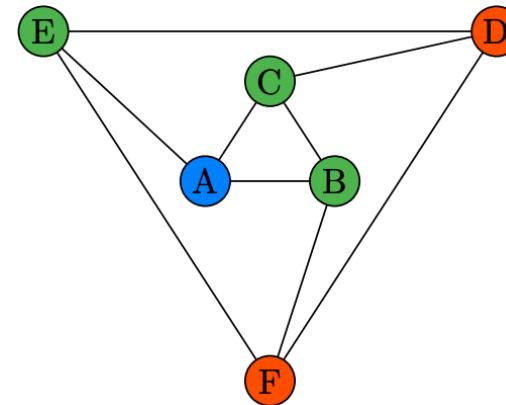
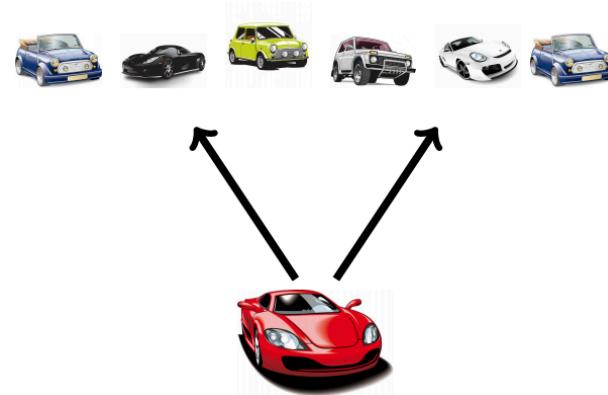


- If arm 0 is selected, then the losses of arms 0, 1, 2, and 3 are observed (but not the loss of arm 4).

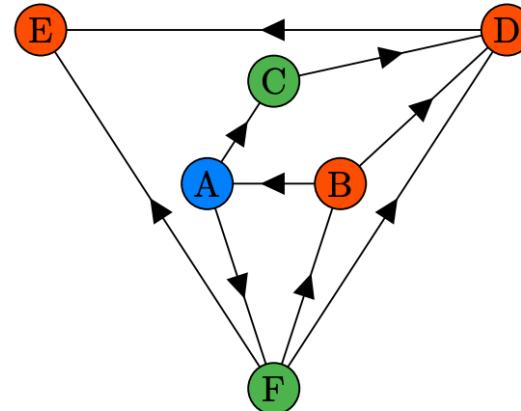
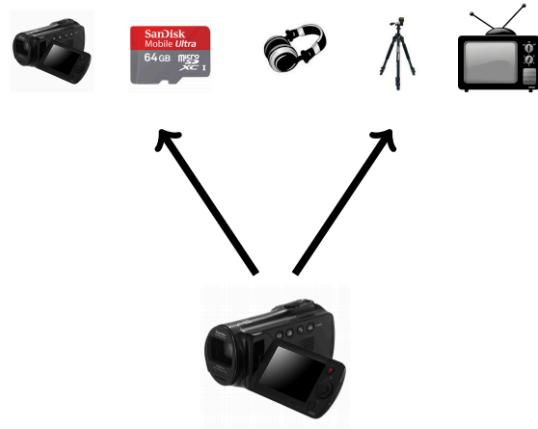
Applications

(Valko, 2016)

■ Undirected graph:



■ Directed graph:



Graph Theory Notions

(Goddard and Henning, 2013)

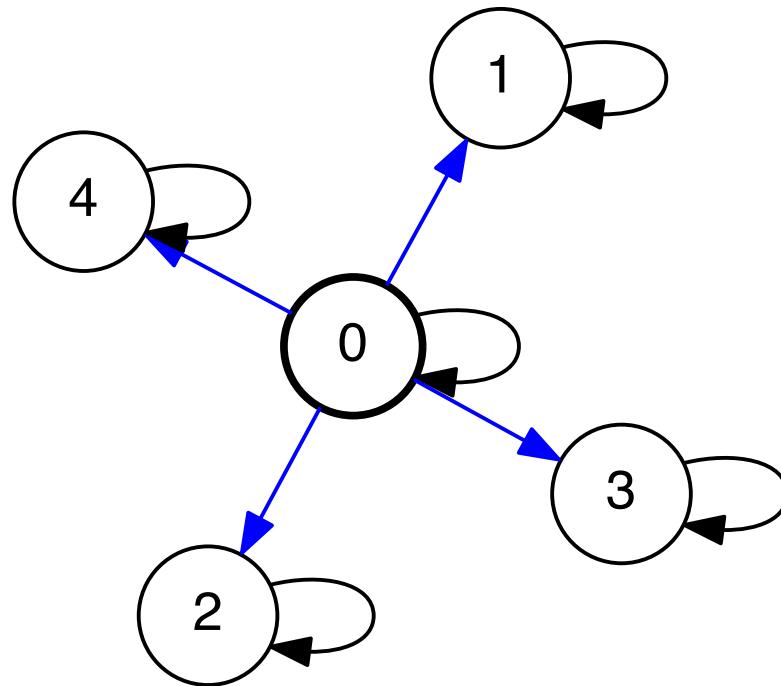
- Given a directed graph $G = (V, E)$ (self-loops ignored),
 - the **mas-number** of G , $\mu(G)$, is the size of the maximum acyclic subgraph of G .
 - a subset of the vertices is **independent** if no two vertices in it are adjacent; the **independence number** of G , $\alpha(G)$, is the size of the maximum independent set in G .
 - a **dominating set** of G is a subset $S \subseteq V$ such that every vertex not in S is adjacent to S ; the **domination number** of G , $\gamma(G)$, is the minimum size of a dominating set.
 - it follows that for any graph: $\gamma(G) \leq \alpha(G) \leq \mu(G)$.

Graph Theory Notions

- Computing domination number is NP-hard since it is equivalent to the minimum vertex cover problem. But, it can be approximated modulo logarithmic factor via greedy set cover:
 - at each round select vertex with largest uncovered adjacent set.
- When G is undirected (symmetric edges), then,

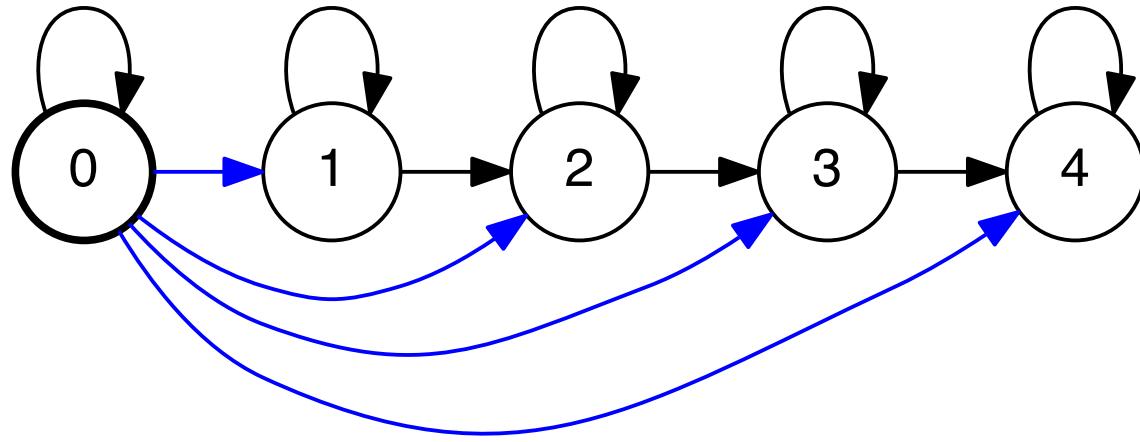
$$\alpha(G) = \mu(G).$$

Examples



- Star graph: $\gamma(G) = 1, \alpha(G) = n - 1, \mu(G) = n.$

Example



- Auction graph: $\gamma(G) = 1, \alpha(G) = (n - 1)/2, \mu(G) = n.$

Adversarial Setting

Protocols

- Graph information:
 - **pre-informed setting**: feedback graph received before selecting arm.
 - **uninformed setting**: feedback graph received after selecting arm.
- Time-dependent or fixed feedback graph.

EXP3-SET Algorithm

■ (Alon et al., 2013): variant of EXP3;

- uninformed setting.
- directed feedback graphs $G_t = (V, E_t)$.
- surrogate loss:

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{q_{i,t}} \mathbb{I}\{(I_t, i) \in E_t\}$$

$$q_{i,t} = \sum_{j: (j,i) \in E_t} p_{j,t}$$

Probability of observing i .

I_t is an action we choose at step t , while it is not necessarily i now.

EXP3-SET

EXP3-SET(η)

```
1   $\forall i \in V, w_{i,1} \leftarrow 1$ 
2  for  $t \leftarrow 1$  to  $T$  do
3       $\forall i \in V, p_{i,t} \leftarrow \frac{w_{i,t}}{\sum_{j \in V} w_{j,t}}$ 
4      SAMPLE( $I_t \sim p_t$ )
5      RECEIVE( $\{(j, \ell_{j,t}) : (I_t, j) \in E_t\}$ )
6      RECEIVE( $G_t$ )
7      for  $i \leftarrow 1$  to  $|V|$  do
8           $q_{i,t} \leftarrow \sum_{j : (j,i) \in E_t} p_{j,t}$   $\triangleright$  probability of observing  $i$ .
9           $\tilde{\ell}_{i,t} \leftarrow \frac{\ell_{i,t}}{q_{i,t}} \mathbb{I}\{(I_t, i) \in E_t\}$ 
10          $w_{i,t+1} \leftarrow e^{-\eta \tilde{\ell}_{i,t}} w_{i,t}$ 
```

EXP3-SET Guarantee

- **Theorem:** the pseudo-regret of EXP3-SET can be bounded as follows:

$$\overline{\text{Reg}}(\text{EXP3-SET}) \leq \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}[Q_t],$$

- where $Q_t = \sum_{i \in V} \frac{p_{i,t}}{q_{i,t}}$.

Proof

- Potential: $\Phi_{t+1} = \log \sum_{i=1}^K e^{-\eta \tilde{L}_{i,t}}$, with $\tilde{L}_{i,t} = \sum_{s=1}^t \tilde{\ell}_{i,s}$.
- Upper bound:

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= \log \frac{\sum_{i=1}^K e^{-\eta \tilde{L}_{i,t}}}{\sum_{i=1}^N e^{-\eta \tilde{L}_{i,t-1}}} = \log \frac{\sum_{i=1}^K e^{-\eta \tilde{L}_{i,t-1}} e^{-\eta \tilde{\ell}_{i,t}}}{\sum_{i=1}^N e^{-\eta \tilde{L}_{i,t-1}}} \\ &= \log \left[\mathbb{E}_{i \sim \mathbf{p}_t} [e^{-\eta \tilde{\ell}_{i,t}}] \right] \\ &\leq \mathbb{E}_{i \sim \mathbf{p}_t} [e^{-\eta \tilde{\ell}_{i,t}}] - 1 \quad (\log x \leq x - 1) \\ &\leq \mathbb{E}_{i \sim \mathbf{p}_t} \left[-\eta \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \tilde{\ell}_{i,t}^2 \right] \quad (e^{-x} \leq 1 - x + \frac{x^2}{2}).\end{aligned}$$

- Summing up:

$$\Phi_{T+1} - \Phi_1 \leq -\eta \sum_{t=1}^T \sum_{i \in V} p_{i,t} \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \sum_{t=1}^T \sum_{i \in V} p_{i,t} \tilde{\ell}_{i,t}^2.$$

Proof

■ Lower bound:

$$\Phi_{T+1} - \Phi_1 = \log \left[\sum_{i=1}^K e^{-\eta \tilde{L}_{i,T}} - \log K \right] \geq -\eta \tilde{L}_{j,T} - \log K = -\eta \sum_{t=1}^T \tilde{\ell}_{j,t} - \log K.$$

■ Comparison:

$$\sum_{t=1}^T \sum_{i \in V} p_{i,t} \tilde{\ell}_{i,t} \leq \sum_{t=1}^T \tilde{\ell}_{j,t} + \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{i \in V} p_{i,t} \tilde{\ell}_{i,t}^2.$$

■ Using conditional expectation $\mathbb{E}_t = \mathbb{E}_{I_t \sim \mathbf{p}_t} [\cdot | I_1, \dots, I_{t-1}]$:

$$\sum_{t=1}^T \sum_{i \in V} \mathbb{E} \left[p_{i,t} \mathbb{E}_t [\tilde{\ell}_{i,t}] \right] \leq \sum_{t=1}^T \mathbb{E} \left[\mathbb{E}_t [\tilde{\ell}_{j,t}] \right] + \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{i \in V} \mathbb{E} \left[p_{i,t} \mathbb{E}_t [\tilde{\ell}_{i,t}^2] \right].$$

Proof

■ Observe that:

$$\begin{aligned}\mathbb{E}_t \left[\tilde{\ell}_{i,t} \right] &= \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\frac{\ell_{i,t}}{q_{i,t}} \mathbb{I}\{(I_t, i) \in E\} \right] \\ &= \sum_{j \in V} p_{j,t} \frac{\ell_{i,t}}{q_{i,t}} \mathbb{I}\{(j, i) \in E\} = q_{i,t} \frac{\ell_{i,t}}{q_{i,t}} = \ell_{i,t}.\end{aligned}$$

■ Similarly,

$$\begin{aligned}\mathbb{E}_t \left[\tilde{\ell}_{i,t}^2 \right] &= \mathbb{E}_{I_t \sim \mathbf{p}_t} \left[\frac{\ell_{i,t}^2}{q_{i,t}^2} \mathbb{I}\{(I_t, i) \in E\} \right] \\ &= \sum_{j \in V} p_{j,t} \frac{\ell_{i,t}^2}{q_{i,t}^2} \mathbb{I}\{(j, i) \in E\} = q_{i,t} \frac{\ell_{i,t}^2}{q_{i,t}^2} = \frac{\ell_{i,t}^2}{q_{i,t}} \leq \frac{1}{q_{i,t}}.\end{aligned}$$

Proof

■ Thus,

$$\sum_{t=1}^T \sum_{i \in V} \mathbb{E}[p_{i,t} \ell_{i,t}] \leq \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}_t[\tilde{\ell}_{j,t}]\right] + \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{i \in V} \mathbb{E}\left[\frac{p_{i,t}}{q_{i,t}}\right],$$

- and,

$$\overline{\text{Reg}}(\text{EXP3-SET}) \leq \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}[Q_{i,t}].$$

EXP3-SET Guarantee

- **Theorem:** the pseudo-regret of EXP3-SET can be bounded as follows for **directed graphs**:

$$\overline{\text{Reg}}(\text{EXP3-SET}) \leq \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}[\mu(G_t)].$$

- for $\mathbb{E}[\mu(G_t)] \leq \mu_t$,

$$\overline{\text{Reg}}(\text{EXP3-SET}) \leq \sqrt{2(\log K) \sum_{t=1}^T \mu_t}.$$

Proof

- For any graph (dropping time indices),

$$\sum_{i=1}^K \frac{p_i}{\sum_{j \in \text{IN}(i)} p_j} \leq \mu(G).$$

- Construct subset of vertices V' inducing acyclic graph such that $\sum_{i=1}^K \frac{p_i}{\sum_{j \in \text{IN}(i)} p_j} \leq |V'|$.
- define $i_1 = \operatorname{argmin}_{i \in V} \sum_{j \in \text{IN}(i)} p_j$ and remove that vertex from the graph as well as all $j \in \text{IN}(i_1)$ and all edges entering or leaving these vertices.
- Observe that:

$$\sum_{k \in \text{IN}(i_1)} \frac{p_k}{\sum_{j \in \text{IN}(k)} p_j} \leq \sum_{k \in \text{IN}(i_1)} \frac{p_k}{\sum_{j \in \text{IN}(i_1)} p_j} = 1.$$

Proof

■ Thus, $\sum_{i=1}^K \frac{p_i}{\sum_{j \in \text{IN}(i)} p_j} \leq \sum_{i \notin \text{IN}(i_1)} \frac{p_i}{\sum_{j \in \text{IN}(i)} p_j} + 1.$

- reiterating until no vertex is left, with $V' = \{i_1, \dots, i_k\}$,

$$\sum_{k \in \text{IN}(i_1)} \frac{p_k}{\sum_{j \in \text{IN}(k)} p_j} \leq |V'|.$$

- the graph induced by V' cannot contain cycles since at each step all incoming edges of i_r and source vertices of those edges are removed.

EXP3-SET Guarantee

- **Corollary:** the pseudo-regret of EXP3-SET can be bounded as follows for **undirected graphs**:

$$\overline{\text{Reg}}(\text{EXP3-SET}) \leq \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathbb{E}[\alpha(G_t)].$$

- for $\alpha(G_t) \leq \alpha_t$,

$$\overline{\text{Reg}}(\text{EXP3-SET}) \leq \sqrt{2(\log K) \sum_{t=1}^T \alpha_t}.$$

Adversarial Setting

- [\(Mannor and Shamir, 2011\)](#): introduced online learning with side information modeled as feedback graph.
- [\(Alon et al., 2013\)](#): directed feedback graphs, variants of EXP3.
- [\(Alon et al., 2015\)](#): algorithm with $O(T^{\frac{2}{3}})$ regret for weakly observable graphs (vertex with no self-loop or no entering edge from all other vertices).
- [\(Alon et al., 2014\)](#): high probability bounds based on mas-number.
- [\(Neu 2015\)](#): high probability bounds based on independence number, *implicit exploration*.

Stochastic Setting

UCB-N

- (Caron et al., 2012): UCB-type algorithm;

- number of observations of arm i up to time $(t - 1)$, $O_{i,t-1}$.
- average reward of arm i up to time $(t - 1)$, $\bar{X}_{i,t-1}$.

$$\bar{X}_{i,t-1} = \frac{1}{O_{i,t-1}} \sum_{s=1}^{t-1} X_{i,s} 1_{i \in N(I_s)}.$$

- arm selected at time t : $I_t = \operatorname{argmax}_{i \in [K]} \bar{X}_{i,t-1} + \sqrt{\frac{2 \log t}{O_{i,t-1}}}$.

UCB-N

UCB-N(G)

```
1    $\forall i, \bar{X}_i, O_i \leftarrow 0$ 
2   for  $t \leftarrow 1$  to  $T$  do
3        $I_t \leftarrow \operatorname{argmax}_{i \in [K]} \bar{X}_i + \sqrt{\frac{2 \log T}{O_i}}$ 
4       for  $k \in N(I_t)$  do
5            $O_k \leftarrow O_k + 1$ 
6            $\bar{X}_k \leftarrow \frac{1}{O_k} X_k + (1 - \frac{1}{O_k}) \bar{X}_k$ 
```

Graph Theory Notions

- A **clique** in an undirected graph $G = (V, E)$: subset of V with any two vertices being adjacent.
- A **clique covering** \mathcal{C} of G is a set of cliques such that

$$V = \bigcup_{C \in \mathcal{C}} C.$$

UCB-N Guarantee

- **Theorem:** the pseudo-regret of UCB-N can be bounded as follows for **undirected graphs**:

$$\overline{\text{Reg}}(\text{UCB-N}) \leq \inf_{\mathcal{C}} \left\{ 8 \sum_{C \in \mathcal{C}} \frac{\max_{i \in C} \Delta_i}{\min_{i \in C} \Delta_i^2} \log T \right\} + \left(1 + \frac{\pi^2}{3} \right) \sum_{i=1}^K \Delta_i.$$

Proof

■ **Lemma:** for any $s \geq 0$, for $T_C(t-1) = \sum_{i \in C} T_i(t-1)$,

$$\sum_{t=1}^T \sum_{i \in C} 1_{I_t=i} \Delta_i \leq s \left(\max_{i \in C} \Delta_i \right) + \sum_{t=s+1}^T \sum_{i \in C} 1_{I_t=i} 1_{T_C(t-1) \geq s} \Delta_i.$$

■ **Proof:** observe that

$$\sum_{t=1}^T \sum_{i \in C} 1_{I_t=i} \Delta_i = \sum_{t=1}^T \sum_{i \in C} 1_{I_t=i} 1_{T_C(t-1) < s} \Delta_i + \sum_{t=1}^T \sum_{i \in C} 1_{I_t=i} 1_{T_C(t-1) \geq s} \Delta_i.$$

- Now, for $t^* = \max \{t \leq T : 1_{T_C(t-1) < s} \neq 0\}$,

$$\sum_{t=1}^T \sum_{i \in C} 1_{I_t=i} 1_{T_C(t-1) < s} \Delta_i \leq \left(\max_{i \in C} \Delta_i \right) \sum_{t=1}^{t^*} \sum_{i \in C} 1_{I_t=i} 1_{T_C(t-1) < s}.$$

- By definition of t^* , the number of non-zero terms in the sum is at most s .

Proof

- For any i and t define $\eta_{i,t-1} = \sqrt{\frac{2 \log t}{O_i(t-1)}}$. At time t , if i is selected, then

$$(\hat{\mu}_{i,t-1} + \eta_{i,t-1}) - (\hat{\mu}_{i^*,t} + \eta_{i^*,t-1}) \geq 0$$

$$\Leftrightarrow [\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1}] + [2\eta_{i,t-1} - \Delta_i] + [\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1}] \geq 0.$$

Thus, at least one of these three terms is non-negative. Also, if one is non-positive, at least one of the other two is non-negative.

Proof

- To bound the pseudo-regret, we bound $\sum_{i \in C} \mathbb{E}[T_i(T)]$. Observe first that

$$O_i(t-1) \geq s_C = \max_{i \in C} \left\lceil \frac{8 \log T}{\Delta_i^2} \right\rceil \geq \frac{8 \log T}{\min_{i \in C} \Delta_i^2} \Rightarrow \forall i \in C, \Delta_i - 2\eta_{i,t-1} \geq 0.$$

- Thus,

$$\begin{aligned} \sum_{i \in C} \mathbb{E}[T_i(T)] \Delta_i &= \mathbb{E} \left[\sum_{t=1}^T \sum_{i \in C} 1_{I_t=i} \right] \\ &\leq s_C \left(\max_{i \in C} \Delta_i \right) + \mathbb{E} \left[\sum_{t=s_C+1}^T \sum_{i \in C} 1_{I_t=i} 1_{T_C(t-1) \geq s_C} \Delta_i \right] \\ &\leq s_C \left(\max_{i \in C} \Delta_i \right) + \mathbb{E} \left[\sum_{t=s_C+1}^T \sum_{i \in C} 1_{I_t=i} 1_{O_i(t-1) \geq s_C} \Delta_i \right] \\ &\leq s_C \left(\max_{i \in C} \Delta_i \right) + \sum_{t=s_C+1}^T \sum_{i \in C} \Delta_i \mathbb{P}[\hat{\mu}_{i,t-1} - \mu_{i,t-1} - \eta_{i,t-1} \geq 0] + \Delta_i \mathbb{P}[\mu^* - \hat{\mu}_{i^*,t-1} - \eta_{i^*,t-1} \geq 0] \\ &\leq s_C \left(\max_{i \in C} \Delta_i \right) + \sum_{t=s_C+1}^T \sum_{i \in C} \Delta_i \sum_{t=1}^T \frac{2}{t^4}. \end{aligned}$$

Proof

- The pseudo-regret of the algorithm can thus be upper-bounded as follows:

$$\begin{aligned}\overline{\text{Reg}}(\text{UCB-N}) &= \sum_{C \in \mathcal{C}} \sum_{i \in C} \mathbb{E}[T_i(T)] \Delta_i \\ &\leq \sum_{C \in \mathcal{C}} s_C \left(\max_{i \in C} \Delta_i \right) + \sum_{C \in \mathcal{C}} \sum_{t=s_C+1}^T \sum_{i \in C} \Delta_i \sum_{t=1}^T \frac{2}{t^4} \\ &\leq \sum_{C \in \mathcal{C}} \left(\max_{i \in C} \Delta_i \right) \frac{8 \log T}{\min_{i \in C} \Delta_i^2} + \sum_{C \in \mathcal{C}} \sum_{i \in C} \Delta_i \left(1 + \frac{\pi^2}{3} \right).\end{aligned}$$

Stochastic Setting

- [\(Caron et al., 2012\)](#): UCB-type algorithm for undirected feedback graphs in stochastic setting; guarantees in terms of graph clique structure.
- [\(Cohen et al., 2016\)](#): full feedback graph never revealed, regret guarantee based on independence number, contrast with adversarial setting where bandit bound remains optimal.
- [\(Buccapatnam et al., 2014\)](#): LP-based solution, regret guarantee based on domination number, lower bound.

Stochastic Bandits with Side Observations on Networks

Stochastic Setting

- [\(Buccapatnam et al., 2017\)](#): more general setting covering [\(Cohen et al., 2016\)](#).
- [\(Lykouris et al., 2020\)](#): analysis of algorithms using *layering technique*; e.g. independence number guarantee for UCB-N.
- [\(Cortes et al., 2019\)](#): sleeping experts with dependent losses and awake sets.
- [\(Cortes et al., 2020\)](#): dependent losses and feedback graphs varying stochastically.
- [\(Marinov, MM, Zimmert, 2022\)](#): notion of optimal finite-time regret not uniquely defined in this context! Algorithm with quasi-optimal pseudo-regret for a *meaningful* notion.

Extensions

- [\(Valko, 2016\)](#): general survey of feedback graphs.
- [\(Kocak et al., 2016\)](#): online learning with noisy side information.
- [\(Arora et al., 2019\)](#): adversarial setting with feedback graphs and switching costs.
- [\(Dann et al., 2020\)](#): reinforcement learning with feedback graphs.

Advanced Machine Learning

Bandit Convex Optimization

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Set-Up

- Convex set C .
- For $t = 1$ to T do
 - predict $\mathbf{x}_t \in C$.
 - receive convex loss function $f_t: C \rightarrow \mathbb{R}$.
 - incur loss $f_t(\mathbf{x}_t)$.
- Bandit setting: only loss revealed, no gradient information.
- Regret of algorithm \mathcal{A} :

$$R_T(\mathcal{A}) = \sum_{t=1}^T f_t(\mathbf{x}_t) - \inf_{\mathbf{x} \in C} \sum_{t=1}^T f_t(\mathbf{x}).$$

Single-Point Gradient Estimate

(Flaxman et al., 2005)

■ Definitions:

- $\mathbb{B} = \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\| \leq 1\}$.
- $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\| = 1\}$.
- $\hat{f}(\mathbf{x}) = \underset{\mathbf{v} \in \mathbb{B}}{\text{E}} [f(\mathbf{x} + \delta \mathbf{v})]$: smoothed version of $f(\mathbf{x})$.

■ Lemma: fix $\delta > 0$. Then, the following equality holds:

$$\underset{\mathbf{u} \in \mathbb{S}}{\text{E}} [f(\mathbf{x} + \delta \mathbf{u}) \mathbf{u}] = \frac{\delta}{N} \nabla \hat{f}(\mathbf{x}).$$

Proof

■ By Stokes' theorem,

$$\nabla \int_{\delta\mathbb{B}} f(\mathbf{x} + \mathbf{v}) d\mathbf{v} = \int_{\delta\mathbb{S}} f(\mathbf{x} + \mathbf{u}) \frac{\mathbf{u}}{\|\mathbf{u}\|} d\mathbf{u}.$$

■ Thus,

$$\begin{aligned}\nabla \hat{f}(\mathbf{x}) &= \nabla \left[\frac{\int_{\delta\mathbb{B}} f(\mathbf{x} + \mathbf{v}) d\mathbf{v}}{\text{vol}_N(\delta\mathbb{B})} \right] = \frac{\int_{\delta\mathbb{S}} f(\mathbf{x} + \mathbf{v}) d\mathbf{v}}{\text{vol}_N(\delta\mathbb{B})} \\ &= \frac{\int_{\delta\mathbb{S}} f(\mathbf{x} + \mathbf{v}) d\mathbf{v}}{\text{vol}_{N-1}(\delta\mathbb{S})} \frac{\text{vol}_{N-1}(\delta\mathbb{S})}{\text{vol}_N(\delta\mathbb{B})} \\ &= \underset{\mathbf{u} \in \mathbb{S}}{\text{E}} [f(\mathbf{x} + \delta\mathbf{u}) \mathbf{u}] \frac{N}{\delta}.\end{aligned}$$

Algorithm

(Flaxman et al., 2005)

- Assume that C centered in the origin and let $C_\delta = \frac{1}{1-\delta}C$.

FKM(T)

```
1    $\mathbf{y}_1 \leftarrow \mathbf{0}$ 
2   for  $t \leftarrow 1$  to  $T$  do
3        $\mathbf{u}_t \leftarrow \text{SAMPLE}(\mathbb{S})$ 
4        $\mathbf{x}_t \leftarrow \mathbf{y}_t + \delta \mathbf{u}_t$ 
5       LOSS  $\leftarrow \text{RECEIVE}(f_t(\mathbf{x}_t))$ 
6        $\mathbf{g}_t \leftarrow \frac{N}{\delta} f_t(\mathbf{x}_t) \mathbf{u}_t$ 
7        $\mathbf{y}_{t+1} \leftarrow \Pi_{C_\delta}(\mathbf{y}_t - \eta \mathbf{g}_t)$ 
```

Analysis

■ Assumptions:

- $\text{diam}(C) \leq D$.
- f_t bounded by M and G -Lipschitz.

■ Theorem: the regret of the FKM algorithm is bounded by

$$\frac{D^2}{2\eta} + \frac{\eta M^2 N^2 T}{2\delta^2} + \delta(D+2)GT.$$

- choosing $\eta = \frac{\delta D}{MN\sqrt{T}}$ and $\delta = \sqrt{\frac{DMN}{(D+2)G}} \frac{1}{T^{\frac{1}{4}}}$ yields the upper bound

$$2\sqrt{D(D+2)GMN} T^{\frac{3}{4}} = O(\sqrt{N}T^{\frac{3}{4}}).$$

Proof

- Let x_δ^* be the projection of x^* on C_δ , then $\|x^* - x_\delta^*\| \leq \delta D$.
Thus, since f_t s are G -Lipschitz,

$$\begin{aligned} & \sum_{t=1}^T (\mathbb{E}[f_t(\mathbf{x}_t)] - f_t(\mathbf{x}^*)) \\ &= \sum_{t=1}^T (\mathbb{E}[f_t(\mathbf{x}_t)] - \mathbb{E}[\hat{f}_t(\mathbf{x}_t)] + \mathbb{E}[\hat{f}_t(\mathbf{x}_t)] - \hat{f}_t(\mathbf{x}_\delta^*) + \hat{f}_t(\mathbf{x}_\delta^*) - f_t(\mathbf{x}_\delta^*) + f_t(\mathbf{x}_\delta^*) - f_t(\mathbf{x}^*)) \\ &\leq \sum_{t=1}^T (\mathbb{E}[\hat{f}_t(\mathbf{x}_t)] - \hat{f}_t(\mathbf{x}_\delta^*)) + 2\delta GT + \delta DGT \\ &\leq \sum_{t=1}^T (\mathbb{E}[\hat{f}_t(\mathbf{x}_t)] - \hat{f}_t(\mathbf{x}_\delta^*)) + \delta(D+2)GT. \end{aligned}$$

Proof

■ **Lemma:** fix a sequence of convex and differentiable functions $u_1, \dots, u_T: C \rightarrow \mathbb{R}$ and $\eta > 0$. Let $\mathbf{z}_0, \dots, \mathbf{z}_T \in C$ be defined by $\mathbf{z}_0 = 0$ and $\mathbf{z}_{t+1} = \Pi_C(\mathbf{z}_t - \eta \mathbf{g}_t)$, where \mathbf{g}_t s are random variables such that

- $E[\mathbf{g}_t | \mathbf{z}_t] = \nabla u_t(\mathbf{z}_t)$ and $\|\mathbf{g}_t\| \leq G$; then,

$$E \left[\sum_{t=1}^T u_t(\mathbf{z}_t) \right] - \min_{\mathbf{z} \in C} \sum_{t=1}^T u_t(\mathbf{z}) \leq E[R_T(\text{PSGD}, \mathbf{g}_1, \dots, \mathbf{g}_T)].$$

■ **Proof:** define h_t by $h_t(\mathbf{z}) = u_t(\mathbf{z}) + [\mathbf{g}_t - \nabla u_t(\mathbf{z}_t)] \cdot \mathbf{z}$. Then, $\nabla h_t(\mathbf{z}_t) = \mathbf{g}_t$, $E[h_t(\mathbf{z}_t)] = E[u_t(\mathbf{z}_t)]$ since $E[\mathbf{g}_t | \mathbf{z}_t] = \nabla u_t(\mathbf{z}_t)$ and for any fixed \mathbf{z} , $E[h_t(\mathbf{z})] = E[u_t(\mathbf{z})]$. Thus, running deterministic PSGD on h_t s is equivalent to expected PSGD on the fixed functions u_t s.

Proof

■ Regret bound for online projected gradient descent:

$$\begin{aligned} & \sum_{t=1}^T (\mathbb{E}[\hat{f}_t(\mathbf{x}_t)] - \hat{f}_t(\mathbf{x}_\delta^*)) \\ & \leq \sum_{t=1}^T \mathbb{E} [\mathbf{g}_t \cdot (\mathbf{x}_t - \mathbf{x}_\delta^*)] \\ & = \sum_{t=1}^T \frac{1}{2\eta} \mathbb{E} \left[\|\mathbf{x}_t - \mathbf{x}_\delta^*\|^2 + \eta^2 \|\mathbf{g}_t\|^2 - \|\mathbf{x}_t - \eta\mathbf{g}_t - \mathbf{x}_\delta^*\|^2 \right] \\ & \leq \sum_{t=1}^T \frac{1}{2\eta} \mathbb{E} \left[\|\mathbf{x}_t - \mathbf{x}_\delta^*\|^2 + \eta^2 M^2 \frac{N^2}{\delta^2} - \|\mathbf{x}_{t+1} - \mathbf{x}_\delta^*\|^2 \right] \quad (\text{prop. of proj.}) \\ & \leq \frac{1}{2\eta} \mathbb{E} \left[\|\mathbf{x}_1 - \mathbf{x}_\delta^*\|^2 + \eta^2 M^2 \frac{N^2}{\delta^2} - \|\mathbf{x}_{T+1} - \mathbf{x}_\delta^*\|^2 \right] \\ & \leq \frac{1}{2\eta} \left[\|\mathbf{x}_1 - \mathbf{x}_\delta^*\|^2 + \eta^2 M^2 \frac{N^2}{\delta^2} T \right] \leq \frac{1}{2\eta} \left[D^2 + \eta^2 M^2 \frac{N^2}{\delta^2} T \right]. \end{aligned}$$

References

- J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In COLT, pp. 263–274, 2008.
- Agarwal, A., O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In COLT, pp. 28–40, 2010.
- Bubeck, S. and R. Eldan. Multi-scale exploration of convex functions and bandit convex optimization. CoRR abs/1507.06580, 2015.
- V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In NIPS, 2008.
- Ofer Dekel, Ronen Eldan, Tomer Koren. Bandit Smooth Convex Optimization: Improving the Bias-Variance Tradeoff. NIPS 2015: 2926-2934.

References

- A. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In SODA, pp. 385–394, 2005.
- Hazan, E. and K. Y. Levy (2014). Bandit convex optimization: Towards tight bounds. In NIPS, pp. 784–792.
- Saha, A. and A. Tewari (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In AISTATS, pp. 636–642.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In ICML, pages 928–936, 2009.

Bandit Convex Optimization

Scott Yang

March 7, 2017

Table of Contents

1 Bandit Convex Optimization (BCO)

2 Projection Methods

3 Barrier Methods

4 Variance reduction

5 Other methods

6 Conclusion

Learning scenario

- Compact convex action set $\mathcal{K} \subset \mathbb{R}^d$.
- For $t = 1$ to T :
 - Predict $x_t \in \mathcal{K}$.
 - Receive convex loss function $f_t : \mathcal{K} \rightarrow \mathbb{R}$.
 - Incur loss $f_t(x_t)$.
- *Bandit setting*: only loss revealed, no other information.
- Regret of algorithm \mathcal{A} :

$$\text{Reg}_T(\mathcal{A}) = \sum_{t=1}^T f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^T f_t(x).$$

Related settings

- Online convex optimization: ∇f_t (and maybe $\nabla^2 f_t, \nabla^3 f_t, \dots$) known at each round.
- Multi-armed bandit: $\mathcal{K} = \{1, 2, \dots, K\}$ discrete.
- Zero-th order optimization: $f_t = f$.
- Stochastic bandit convex optimization: $f_t(x) = f(x) + \epsilon_t$, $\epsilon_t \sim \mathcal{D}$ noisy estimate.
- Multi-point bandit convex optimization: Query f_t at points $(x_{t,i})_{i=1}^m$, $m \geq 2$.

Table of Contents

1 Bandit Convex Optimization (BCO)

2 Projection Methods

3 Barrier Methods

4 Variance reduction

5 Other methods

6 Conclusion

“Gradient descent without a gradient”

- Online gradient descent algorithm:

$$x_{t+1} \leftarrow x_t - \eta \nabla f_t(x_t).$$

- BCO setting: $\nabla f_t(x_t)$ is not known!

- BCO idea:

- Find \hat{g}_t such that $\hat{g}_t \approx \nabla f_t(x_t)$.
 - Update

$$x_{t+1} \leftarrow x_t - \eta \hat{g}_t.$$

- Question: how do we pick \hat{g}_t ?

Single-point gradient estimates (one dimension)

- By the fundamental theorem of calculus:

$$\begin{aligned} f'(x) &\approx \frac{1}{2\delta} \int_{-\delta}^{\delta} f'(x+y) dy = \frac{1}{2\delta} [f(x+\delta) - f(x-\delta)] \\ &= \mathbb{E}_{z \sim \mathcal{D}} \left[\frac{1}{\delta} f(x+z) \frac{z}{|z|} \right] \\ &\text{where } \mathcal{D}(z) = \delta \text{ w.p. } \frac{1}{2} \text{ and } = -\delta \text{ w.p. } \frac{1}{2}. \end{aligned}$$

- With enough regularity (e.g. f Lipschitz),

$$\frac{d}{dx} \frac{1}{2\delta} \int_{-\delta}^{\delta} f(x+y) dy = \frac{1}{2\delta} \int_{-\delta}^{\delta} f'(x+y) dy.$$

Single-point gradient estimates (higher dimensions)

- $\mathbb{B}_1 = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$.
- $\mathbb{S}_1 = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$.
- $\int_A dy = |A|$.
- By Stokes' theorem,

$$\begin{aligned}\nabla f(x) &\approx \frac{1}{|\delta\mathbb{B}_1|} \int_{\delta\mathbb{B}_1} \nabla f(x + y) dy = \frac{1}{|\delta\mathbb{B}_1|} \int_{\delta\mathbb{S}_1} f(x + z) \frac{z}{|z|} dz \\ &= \frac{1}{|\delta\mathbb{B}_1|} \int_{\mathbb{S}_1} f(x + \delta z) z dz = \frac{|\mathbb{S}_1|}{|\mathbb{S}_1||\delta\mathbb{B}_1|} \int_{\mathbb{S}_1} f(x + \delta z) z dz \\ &= \frac{|\mathbb{S}_1|}{|\delta\mathbb{B}_1|} \mathbb{E}_{z \sim U(\mathbb{S}_1)} [f(x + \delta z)] = \frac{d}{\delta} \mathbb{E}_{z \sim U(\mathbb{S}_1)} [f(x + \delta z)].\end{aligned}$$

- With enough regularity on f ,

$$\nabla \frac{1}{|\delta\mathbb{B}_1|} \int_{\delta\mathbb{B}_1} f(x + y) dy = \frac{1}{|\delta\mathbb{B}_1|} \int_{\delta\mathbb{B}_1} \nabla f(x + y) dy.$$

Projection method [Flaxman et al, 2005]

- Let $\hat{f}(x) = \frac{1}{|\delta\mathbb{B}_1|} \int_{\delta\mathbb{B}_1} f(x + y) dy$.
- Estimate $\nabla \hat{f}(x)$ by sampling on $\delta\mathbb{S}_1(x)$
- Project gradient descent update to keep samples inside \mathcal{K} :
 $K_\delta = \frac{1}{1-\delta} K$.

BANDITPGD(T, η, δ):

- $x_1 \leftarrow 0$.
- For $t = 1, 2, \dots, T$:
 - $u_t \leftarrow \text{SAMPLE}(\mathbb{S}_1)$
 - $y_t \leftarrow x_t + \delta u_t$
 - $\text{PLAY}(y_t)$
 - $f_t(y_t) \leftarrow \text{RECEIVELOSS}(y_t)$
 - $\hat{g}_t \leftarrow \frac{d}{\delta} f_t(y_t) u_t$
 - $x_{t+1} \leftarrow \Pi_{K_\delta}(x_t - \eta \hat{g}_t)$.

Analysis of BANDITPGD

Theorem (Flaxman et al, 2005)

Assume $\text{diam}(\mathcal{K}) \leq D$, $|f_t| \leq C$, and $\|\nabla f_t\| \leq L$. Then after T rounds, the (expected) regret of the BANDITPGD algorithm is bounded by:

$$\frac{D^2}{2\eta} + \frac{\eta C^2 d^2 T}{2\delta^2} + \delta(D+2)L T.$$

In particular, by setting $\eta = \frac{\delta D}{Cd\sqrt{T}}$ and $\delta = \sqrt{\frac{DCd}{(D+2)L T^{1/2}}}$, the regret is upper bounded by: $\mathcal{O}(d^{1/2} T^{3/4})$.

Proof of BANDITPGD regret

- For any $x \in \mathcal{K}$, let $x_\delta = \Pi_{\mathcal{K}_\delta}(x)$.
- $\widehat{f}_t(z) \geq f_t(z)$.
- Then

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}[f_t(y_t) - f_t(x^*)] \\ &= \sum_{t=1}^T \mathbb{E} \left[f_t(y_t) - f_t(x_t) + f_t(x_t) - \widehat{f}_t(x_t) + \widehat{f}_t(x_t) - \widehat{f}_t(x_\delta^*) \right. \\ &\quad \left. + \widehat{f}_t(x_\delta^*) - f_t(x_\delta^*) + f_t(x_\delta^*) - f_t(x^*) \right] \\ &\leq \sum_{t=1}^T \mathbb{E} \left[\widehat{f}_t(x_t) - \widehat{f}_t(x_\delta^*) \right] + [2\delta LT + \delta DLT] \\ &\leq \sum_{t=1}^T \mathbb{E} \left[\widehat{f}_t(x_t) - \widehat{f}_t(x_\delta^*) \right] + \delta(D+2)LT. \end{aligned}$$

Proof of BANDITPGD regret

- $\mathbb{E} [\|\hat{g}_t\|_2] \leq \frac{C^2 d^2}{\delta^2}.$
- Thus,

$$\begin{aligned}& \sum_{t=1}^T \mathbb{E} [\hat{f}_t(x_t) - \hat{f}_t(x_\delta^*)] \leq \sum_{t=1}^T \mathbb{E} [\nabla \hat{f}_t(x_t) \cdot (x_t - x_\delta^*)] \\&= \sum_{t=1}^T \mathbb{E} [\hat{g}_t \cdot (x_t - x_\delta^*)] \\&= \sum_{t=1}^T \frac{1}{2\eta} \mathbb{E} [\eta^2 \|\hat{g}_t\|^2 + \|x_t - x_\delta^*\|^2 - \|x_t - \eta \hat{g}_t - x_\delta^*\|^2] \\&\leq \sum_{t=1}^T \frac{1}{2\eta} \mathbb{E} \left[\eta^2 \frac{C^2 d^2}{\delta^2} + \|x_t - x_\delta^*\|^2 - \|x_{t+1} - x_\delta^*\|^2 \right] \\&\leq \frac{1}{2\eta} \mathbb{E} \left[\|x_1 - x_\delta^*\|^2 + \eta^2 \frac{C^2 d^2}{\delta^2} \right] \leq \frac{1}{2\eta} \left[D^2 + \eta^2 \frac{C^2 d^2}{\delta^2} \right].\end{aligned}$$

Table of Contents

1 Bandit Convex Optimization (BCO)

2 Projection Methods

3 Barrier Methods

4 Variance reduction

5 Other methods

6 Conclusion

Revisiting projection

- Goal of projection: Keep $x_t \in \mathcal{K}_\delta$ so that $y_t \in \mathcal{K}$.
- Total “cost” of projection: δDLT .
- Deficiency: completely separate from gradient descent update.
- Question: is there a better way to ensure that $y_t \in \mathcal{K}$?

Gradient Descent to Follow-the-Regularized-Leader

- Let $\hat{g}_{1:t} = \sum_{s=1}^t \hat{g}_s$.
- Proximal form of gradient descent:

$$x_{t+1} \leftarrow x_t - \eta \hat{g}_t$$

$$x_{t+1} \leftarrow \operatorname{argmin}_{x \in \mathbb{R}^d} \eta \hat{g}_{1:t} \cdot x + \|x\|^2$$

- Follow-the-Regularized-Leader: for $\mathcal{R} : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$x_{t+1} \leftarrow \operatorname{argmin}_{x \in \mathbb{R}^d} \hat{g}_{1:t} \cdot x + \mathcal{R}(x),$$

- BCO “wishlist” for \mathcal{R} :
 - Want to ensure that x_{t+1} stays inside \mathcal{K} .
 - Want enough “room” so that $y_{t+1} \in \mathcal{K}$ as well.

Self-concordant barriers

Definition (Self-concordant barrier (SCB))

Let $\nu \geq 0$. A C^3 function $\mathcal{R} : \text{int}(\mathcal{K}) \rightarrow \mathbb{R}$ is a ν -self-concordant barrier for \mathcal{K} if for any sequence $(z_s)_{s=1}^\infty \subset \text{int}(\mathcal{K})$, with $z_s \rightarrow \partial\mathcal{K}$, we have $\mathcal{R}(z_s) \rightarrow \infty$, and for all $x \in \mathcal{K}$ and $y \in \mathbb{R}^n$, the following inequalities hold:

$$|\nabla^3 \mathcal{R}(x)[y, y, y]| \leq 2\|y\|_x^3, \quad |\nabla \mathcal{R}(x) \cdot y| \leq \nu^{1/2} \|y\|_x,$$

where $\|z\|_x^2 = \|z\|_{\nabla^2 \mathcal{R}(x)}^2 = z^\top \nabla^2 \mathcal{R}(x) z$.

Examples of barriers

- $\mathcal{K} = \mathbb{B}_1$:

$$\mathcal{R}(x) = -\log(1 - \|x\|^2)$$

is 1-self-concordant.

- $\mathcal{K} = \{x : a_i^\top x \leq b_i\}_{i=1}^m$:

$$\mathcal{R}(x) = \sum_{i=1}^m -\log(b_i - a_i^\top x)$$

is m -self-concordant.

- Existence of “universal barrier” [Nesterov & Nemirovski, 1994]: every closed convex domain \mathcal{K} admits a $\mathcal{O}(d)$ -self-concordant barrier.

Properties of self-concordant-barriers

- *Translation invariance:*

for any constant $c \in \mathbb{R}$, $\mathcal{R} + z$ is also a SCB (so wlog, we assume $\min_{z \in \mathcal{K}} \mathcal{R}(z) = 0$.)

- *Dikin ellipsoid contained in interior:*

let $\mathcal{E}(x) = \{y \in \mathbb{R}^n : \|y\|_x \leq 1\}$. Then for any $x \in \text{int}(\mathcal{K})$, $\mathcal{E}(x) \subset \text{int}(\mathcal{K})$.

- *Logarithmic growth away from boundary:*

for any $\epsilon \in (0, 1]$, let $y = \operatorname{argmin}_{z \in \mathcal{K}} \mathcal{R}(z)$ and $\mathcal{K}_{y,\epsilon} = \{y + (1 - \epsilon)(x - y) : x \in \mathcal{K}\}$. Then for all $x \in \mathcal{K}_{y,\epsilon}$,

$$\mathcal{R}(x) \leq \nu \log(1/\epsilon).$$

- *Proximity to minimizer:* If $\|\nabla \mathcal{R}(x)\|_{x,*} \leq \frac{1}{2}$, then

$$\|x - \operatorname{argmin} \mathcal{R}\|_x \leq 2\|\nabla \mathcal{R}(x)\|_{x,*}.$$

Adjusting to the local geometry

- Let $A \succ 0$ SPD matrix.
- Sampling around A instead of Euclidean ball:

$$u \sim \text{SAMPLE}(\mathbb{S}_1), \quad x \leftarrow y + \delta A u$$

- Smoothing over A instead of Euclidean ball:

$$\hat{f}(x) = \mathbb{E}_{u \sim U(\mathbb{S}_1)}[f(x + \delta A u)].$$

- One-point gradient estimate based on A :

$$\hat{g} = \frac{d}{\delta} f(x + \delta A u) A^{-1} u, \quad \mathbb{E}_{u \sim U(\mathbb{S}_1)} [\hat{g}] = \nabla \hat{f}(x).$$

- Local norm bound:

$$\|\hat{g}\|_{A^2}^2 \leq \frac{d^2}{\delta^2} C^2.$$

BANDITFTRL [Abernethy et al, 2008; Saha and Tewari 2011]

BANDITFTRL($\mathcal{R}, \delta, \eta, T, x_1$)

- For $t \leftarrow 1$ to T :

- $u_t \leftarrow \text{SAMPLE}(U(\mathbb{S}_1)).$
- $y_t \leftarrow x_t + \delta(\nabla^2 \mathcal{R}(x_t))^{-1/2} u_t.$
- $\text{PLAY}(y_t).$
- $f_t(y_t) \leftarrow \text{RECEIVELOSS}(y_t).$
- $\hat{g}_t \leftarrow \frac{d}{\delta} f_t(y_t) \nabla^2 \mathcal{R}(x_t) u_t.$
- $x_{t+1} \leftarrow \operatorname{argmin}_{x \in \mathbb{R}^d} \eta \hat{g}_{1:t}^\top x + \mathcal{R}(x).$

Analysis of BANDITFTRL

Theorem (Abernethy et al, 2008; Saha and Tewari, 2011)

Assume $\text{diam}(\mathcal{K}) \leq D$. Let \mathcal{R} be a self-concordant-barrier for \mathcal{K} , $|f_t| \leq C$, and $\|\nabla f_t\| \leq L$. Then the regret of BANDITFTRL is upper bounded as follows:

- If $(f_t)_{t=1}^T$ are linear functions, then
 $\text{Reg}_T(\text{BANDITFTRL}) = \tilde{\mathcal{O}}(T^{1/2})$.
- If $(f_t)_{t=1}^T$ have Lipschitz gradients, then
 $\text{Reg}_T(\text{BANDITFTRL}) = \tilde{\mathcal{O}}(T^{2/3})$.

Proof of BANDITFTRL regret: linear case

Approximation error of smoothed losses:

- $x^* = \operatorname{argmin}_{x \in \mathcal{K}} \sum_{t=1}^T f_t(x)$
- $x_\epsilon^* \in \operatorname{argmin}_{y \in \mathcal{K}, \operatorname{dist}(y, \partial \mathcal{K}) > \epsilon} \|y - x^*\|$
- Because f_t are linear,

$$\begin{aligned}\text{Reg}_T(\text{BANDITFTRL}) &= \mathbb{E} \left[\sum_{t=1}^T f_t(y_t) - f_t(x^*) \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T f_t(y_t) - \hat{f}_t(y_t) + \hat{f}_t(y_t) - \hat{f}_t(x_t) + \hat{f}_t(x_t) - \hat{f}_t(x_\epsilon^*) + \hat{f}_t(x_\epsilon^*) \right. \\ &\quad \left. - f_t(x_\epsilon^*) + f_t(x_\epsilon^*) - f_t(x^*) \right] \\ &\leq \mathbb{E} \left[\sum_{t=1}^T \hat{f}_t(x_t) - \hat{f}_t(x_\epsilon^*) \right] + \epsilon LT = \mathbb{E} \left[\sum_{t=1}^T \hat{g}_t^\top (x_t - x_\epsilon^*) \right] + \epsilon LT.\end{aligned}$$

Proof of BANDITFTRL regret: linear case

Claim: for any $z \in \mathcal{K}$,

$$\sum_{t=1}^T \hat{g}_t^\top (x_{t+1} - z) \leq \frac{1}{\eta} \mathcal{R}(z).$$

- $T = 1$ case is true by definition of x_2 .
- Assuming statement is true for $T - 1$:

$$\begin{aligned} \sum_{t=1}^T \hat{g}_t^\top x_{t+1} &= \sum_{t=1}^{T-1} \hat{g}_t^\top x_{t+1} + \hat{g}_T^\top x_{T+1} \leq \frac{1}{\eta} \mathcal{R}(x_T) + \sum_{t=1}^{T-1} \hat{g}_t^\top x_T + \hat{g}_T^\top x_{T+1} \\ &\leq \frac{1}{\eta} \mathcal{R}(x_{T+1}) + \sum_{t=1}^{T-1} \hat{g}_t^\top x_{T+1} + \hat{g}_T^\top x_{T+1} \leq \frac{1}{\eta} \mathcal{R}(z) + \sum_{t=1}^T \hat{g}_t^\top z. \end{aligned}$$

Proof of BANDITFTRL regret: linear case

$$\begin{aligned}\mathbb{E} \left[\sum_{t=1}^T \hat{f}_t(x_t) - \hat{f}_t(x_\epsilon^*) \right] &\leq \sum_{t=1}^T \mathbb{E} \left[\hat{f}_t(x_t) - \hat{f}_t(x_{t+1}) + \hat{f}_t(x_{t+1}) - \hat{f}_t(x_\epsilon^*) \right] \\ &\leq \sum_{t=1}^T \mathbb{E} \left[\|\hat{g}_t\|_{x_t,*} \|x_t - x_{t+1}\|_{x_t} \right] + \frac{1}{\eta} \mathcal{R}(x_\epsilon^*)\end{aligned}$$

Proximity to minimizer for SCB:

- Recall:
 - $x_{t+1} = \operatorname{argmin}_{x \in \mathcal{K}} \eta \hat{g}_{1:t}^\top x + \mathcal{R}(x)$
 - $F_t(x) = \eta \hat{g}_{1:t}^\top x + \mathcal{R}(x)$ is a SCB.
- Proximity bound: $\|x_t - x_{t+1}\|_{x_t} \leq \|\nabla F_t(x_t)\|_{x_t,*} = \eta \|\hat{g}_t\|_{x_t,*}$

Proof of BANDITFTRL regret: linear case

$$\text{Reg}_T(\text{BANDITFTRL}) \leq \epsilon LT + \sum_{t=1}^T \mathbb{E} [\eta \|\hat{g}_t\|_{x_t,*}^2] + \frac{1}{\eta} \mathcal{R}(x_\epsilon^*)$$

- By the local norm bound: $\mathbb{E} [\eta \|\hat{g}_t\|_{x_t,*}^2] \leq \frac{C^2 d^2}{\delta^2}$.
- By the logarithmic growth of the SCB: $\frac{1}{\eta} \mathcal{R}(x_\epsilon^*) \leq \nu \log \left(\frac{1}{\epsilon} \right)$.

$$\Rightarrow \text{Reg}_T(\text{BANDITFTRL}) \leq \epsilon LT + T \eta \frac{C^2 d^2}{\delta^2} + \frac{\nu}{\eta} \log \left(\frac{1}{\epsilon} \right).$$

Table of Contents

1 Bandit Convex Optimization (BCO)

2 Projection Methods

3 Barrier Methods

4 Variance reduction

5 Other methods

6 Conclusion

Issues with BANDITFTRL in the non-linear case

- Approximation error of $f \sim \hat{f}$:
 - $\sim \delta T$ for $\mathcal{C}^{0,1}$ functions
 - $\sim \delta^2 T$ for $\mathcal{C}^{1,1}$ functions
- Variance of gradient estimates: $\mathbb{E} [\eta \|\hat{g}_t\|_{x_t,*}^2] \leq \frac{C^2 d^2}{\delta^2}$
- Regret for non-linear loss functions:
 - $\mathcal{O}(T^{3/4})$ for $\mathcal{C}^{0,1}$ functions
 - $\mathcal{O}(T^{2/3})$ for $\mathcal{C}^{1,1}$ functions
- Question: can we reduce the variance of the gradient estimates to improve the regret?

Variance reduction

- Observation [Dekel et al, 2015]: If $\bar{g}_t = \frac{1}{k+1} \sum_{i=0}^k \hat{g}_{t-i}$, then

$$\|\bar{g}_t\|_{x_t,*}^2 = \mathcal{O}\left(\frac{C^2 d^2}{\delta^2(k+1)}\right).$$

- Note: averaged gradient \bar{g}_t is no longer an unbiased estimate of $\nabla \hat{f}_t$.
- Idea: If f_t is sufficiently regular, then the bias will still be manageable.

Improving variance reduction via “optimism”

- Optimistic FTRL [Rakhlin and Sridharan, 2013]:

$$x_{t+1} \leftarrow \operatorname{argmin}_{x \in \mathcal{K}} (g_{1:t} + \tilde{g}_{t+1})^\top x + \mathcal{R}(x)$$

$$\sum_{t=1}^T f_t(x_t) - f_t(x^*) \leq \eta \sum_{t=1}^T \|g_t - \tilde{g}_t\|_{x_t,*} + \frac{1}{\eta} \mathcal{R}(x^*)$$

- By re-centering the averaged gradient at each step, we can further reduce the variance:

$$\tilde{g}_t = \frac{1}{k+1} \sum_{i=1}^k \hat{g}_{t-i}.$$

- Variance of re-centered averaged gradients:

$$\|\bar{g}_t - \tilde{g}_t\|_{x_t,*}^2 = \frac{1}{(k+1)^2} \|\hat{g}_t\|_{x_t,*}^2 = \mathcal{O}\left(\frac{C^2 d^2}{\delta^2 (k+1)^2}\right).$$

BANDITFTRL-VR

BANDITFTRL-VR($\mathcal{R}, \delta, \eta, k, T, x_1$)

- For $t \leftarrow 1 \rightarrow T$:

- $u_t \leftarrow \text{SAMPLE}(U(\mathbb{S}_1))$
- $y_t \leftarrow x_t + \delta(\nabla^2 \mathcal{R}(x_t))^{-\frac{1}{2}} u_t$
- $\text{PLAY}(y_t)$
- $f_t(y_t) \leftarrow \text{RECEIVELOSS}(y_t)$
- $\hat{g}_t \leftarrow \frac{d}{\delta} f_t(y_t)(\nabla^2 \mathcal{R}(x_t))^{-\frac{1}{2}} u_t$
- $\bar{g}_t \leftarrow \frac{1}{k+1} \sum_{i=0}^k \hat{g}_{t-i}$
- $\tilde{g}_{t+1} \leftarrow \frac{1}{k+1} \sum_{i=1}^k \hat{g}_{t+1-i}$
- $x_{t+1} \leftarrow \underset{x \in \mathbb{R}^d}{\text{argmin}} \eta(\bar{g}_{1:t} + \tilde{g}_{t+1})^\top x + \mathcal{R}(x)$

Analysis of BANDITFTRL-VR

Theorem (Mohri & Y., 2016)

Assume $\text{diam}(\mathcal{K}) \leq D$. Let \mathcal{R} be a self-concordant-barrier for \mathcal{K} , $|f_t| \leq C$, and $\|\nabla f_t\| \leq L$. Then the regret of BANDITFTRL is upper bounded as follows:

- If $(f_t)_{t=1}^T$ are Lipschitz, then $\text{Reg}_T(\text{BANDITFTRL-VR}) = \tilde{\mathcal{O}}(T^{\frac{11}{16}})$.
- If $(f_t)_{t=1}^T$ have Lipschitz gradients, then
 $\text{Reg}_T(\text{BANDITFTRL-VR}) = \tilde{\mathcal{O}}(T^{\frac{8}{13}})$.

Proof of BANDITFTRL-VR regret: Lipschitz case

Approximation: real to smoothed losses

- Relate global optimum x^* to projected optimum x_ϵ^* .
- Use Lipschitz property of losses to relate y_t to x_t and f_t to \hat{f}_t .

$$\begin{aligned}\text{Reg}_T(\text{BANDITFTRL-VR}) &= \mathbb{E} \left[\sum_{t=1}^T f_t(y_t) - f_t(x^*) \right] \\ &\leq \epsilon LT + 2L\delta DT + \sum_{t=1}^T \mathbb{E} \left[\hat{f}_t(x_t) - \hat{f}(x_\epsilon^*) \right].\end{aligned}$$

Proof of BANDITFTRL-VR regret: Lipschitz case

Approximation: smoothed to averaged losses

$$\begin{aligned} \sum_{t=1}^T \mathbb{E} \left[\hat{f}_t(x_t) - \hat{f}(x_\epsilon^*) \right] &= \sum_{t=1}^T \mathbb{E} \left[\frac{1}{k+1} \sum_{i=0}^k \left(\hat{f}_t(x_t) - \hat{f}_{t-i}(x_{t-i}) \right) \right. \\ &\quad \left. + \frac{1}{k+1} \sum_{i=0}^k \left(\hat{f}_{t-i}(x_{t-i}) - \bar{f}_t(x_\epsilon^*) \right) + \frac{1}{k+1} \sum_{i=0}^k \left(\bar{f}_t(x_\epsilon^*) - \hat{f}_t(x_\epsilon^*) \right) \right] \\ &\leq \frac{Ck}{2} + LT \sup_{\substack{t \in [1, T] \\ i \in [0, k \wedge t]}} \mathbb{E} [\|x_{t-i} - x_t\|_2] + \sum_{t=1}^T \mathbb{E} [\bar{g}_t^\top (x_t - x_\epsilon^*)]. \end{aligned}$$

Proof of BANDITFTRL-VR regret: Lipschitz case

FTRL analysis on averaged gradients with re-centering:

$$\sum_{t=1}^T \mathbb{E} [\bar{g}_t^\top (x_t - x_\epsilon^*)] \leq \frac{2C^2 d^2 \eta T}{\delta^2 (k+1)^2} + \frac{1}{\eta} \mathcal{R}(x_\epsilon^*).$$

Cumulative analysis:

$$\begin{aligned} \text{Reg}_T(\text{BANDITFTRL-VR}) &\leq \epsilon LT + 2L\delta DT + \frac{Ck}{2} + \frac{2C^2 d^2 \eta T}{\delta^2 (k+1)^2} + \frac{1}{\eta} \mathcal{R}(x_\epsilon^*) \\ &+ LT \sup_{\substack{t \in [1, T] \\ i \in [0, k \wedge t]}} \mathbb{E} [\|x_{t-i} - x_t\|_2]. \end{aligned}$$

Proof of BANDITFTRL-VR regret: Lipschitz case

Stability estimate for the actions

- Want to bound: $\sup_{\substack{t \in [1, T] \\ i \in [0, k \wedge t]}} \mathbb{E} [\|x_{t-i} - x_t\|_2]$.
- Fact: Let D be the diameter of \mathcal{K} . For any $x \in \mathcal{K}$ and $z \in \mathbb{R}^d$,

$$D^{-1}\|z\|_{x,*} \leq \|z\|_2 \leq D\|z\|_x.$$

- By triangle inequality and equivalence of norms,

$$\begin{aligned} \mathbb{E} [\|x_{t-i} - x_t\|_2] &\leq \sum_{s=t-i}^{t-1} \mathbb{E} [\|x_s - x_{s+1}\|_2] \\ &\leq D \sum_{s=t-i}^{t-1} \mathbb{E} [\|x_s - x_{s+1}\|_{x_s}] \leq D \sum_{s=t-i}^{t-1} 2\eta \mathbb{E} [\|\bar{g}_s + \tilde{g}_{s+1} - \tilde{g}_s\|_{x_s,*}]. \end{aligned}$$

Proof of BANDITFTRL-VR regret: Lipschitz case

- $\bar{g}_s + \tilde{g}_{s+1} - \tilde{g}_s = \frac{1}{k+1} \sum_{i=0}^k \hat{g}_{s-i} + \frac{1}{k+1} \hat{g}_s$
- Thus,

$$\begin{aligned}& \mathbb{E} [\|\bar{g}_s + \tilde{g}_{s+1} - \tilde{g}_s\|_{x_s,*}^2] \\& \leq \frac{3}{k^2} \left\| \sum_{i=0}^{k-1} \mathbb{E}_{s-i} [\hat{g}_{s-i}] \right\|_{x_s,*}^2 + \frac{3}{k^2} \mathbb{E} \left[\left\| \sum_{i=0}^{k-1} \hat{g}_{s-i} - \mathbb{E}_{s-i} [\hat{g}_{s-i}] \right\|_{x_s,*}^2 \right] + \frac{3}{k^2} L \\& \leq \frac{3}{k^2} L + 2D^2L^2 + \frac{3}{k^2} \mathbb{E} \left[\left\| \sum_{i=0}^{k-1} \hat{g}_{s-i} - \mathbb{E}_{s-i} [\hat{g}_{s-i}] \right\|_{x_s,*}^2 \right].\end{aligned}$$

Proof of BANDITFTRL-VR regret: Lipschitz case

- Fact: $\forall 0 \leq i \leq k$ such that $t - i \geq 1$,

$$\frac{1}{2} \|z\|_{x_{t-i},*} \leq \|z\|_{x_t,*} \leq 2 \|z\|_{x_{t-i},*}.$$

- Because the terms in the sum make up martingale difference,

$$\begin{aligned} & \mathbb{E} \left[\left\| \sum_{i=0}^{k-1} \hat{g}_{s-i} - \mathbb{E}_{s-i}[\hat{g}_{s-i}] \right\|_{x_s,*}^2 \right] \leq 4 \mathbb{E} \left[\left\| \sum_{i=0}^{k-1} \hat{g}_{s-i} - \mathbb{E}_{s-i}[\hat{g}_{s-i}] \right\|_{x_{s-k},*}^2 \right] \\ & \leq 4 \sum_{i=0}^{k-1} \mathbb{E} \left[\left\| \hat{g}_{s-i} - \mathbb{E}_{s-i}[\hat{g}_{s-i}] \right\|_{x_{s-k},*}^2 \right] \\ & \leq 16 \sum_{i=0}^{k-1} \mathbb{E} \left[\left\| \hat{g}_{s-i} - \mathbb{E}_{s-i}[\hat{g}_{s-i}] \right\|_{x_{s-i},*}^2 \right] \\ & \leq 16 \sum_{i=0}^{k-1} \mathbb{E} \left[\left\| \hat{g}_{s-i} \right\|_{x_{s-i},*}^2 \right] \leq 16 \sum_{i=0}^{k-1} \frac{C^2 d^2}{\delta^2} = 16k \frac{C^2 d^2}{\delta^2}. \end{aligned}$$

Proof of BANDITFTRL-VR regret: Lipschitz case

- By combining the components of the stability estimate,

$$\mathbb{E} [\|x_{t-i} - x_t\|_2] \leq 2\eta D \sum_{s=t-i}^{t-1} \sqrt{\frac{3}{k^2} L + 2D^2L^2 + \frac{3}{k^2} 16k \frac{C^2 d^2}{\delta^2}}.$$

- By the previous calculations,

$$\begin{aligned} \text{Reg}_T(\text{BANDITFTRL-VR}) &\leq \epsilon LT + 2L\delta DT + \frac{Ck}{2} + \frac{2C^2d^2\eta T}{\delta^2(k+1)^2} \\ &+ \frac{1}{\eta} \log(1/\epsilon) + LTD2\eta k \sqrt{\frac{3}{k^2} L + 2D^2L^2 + \frac{48}{k^2} \frac{C^2 d^2}{\delta^2}}. \end{aligned}$$

- Now set $\eta = T^{-11/16}d^{-3/8}$, $\delta = T^{-5/16}d^{3/8}$, $k = T^{1/8}d^{1/4}$.

Discussion of BANDITFTRL-VR regret: Lipschitz gradient case

- Approximation of real to smoothed losses incurs a $\delta^2 D^2 T$ penalty instead of δDT .
- Rest of analysis also leads to some changes in constants.
- General regret bound:

$$\begin{aligned}\text{Reg}_T(\text{BANDITFTRL-VR}) &\leq \epsilon LT + H\delta^2 D^2 T + Ck + \frac{2C^2 d^2 \eta T}{\delta^2(k+1)^2} \\ &+ \frac{1}{\eta} \log(1/\epsilon) + (TL + DHT)2\eta kD \sqrt{\frac{3}{k^2} L + 2D^2 L^2} + \frac{48}{k^2} \frac{C^2 d^2}{\delta^2}.\end{aligned}$$

- Now set $\eta = T^{-8/13}d^{-5/6}$, $\delta = T^{-5/26}d^{1/3}$, $k = T^{1/13}d^{5/3}$.

Table of Contents

1 Bandit Convex Optimization (BCO)

2 Projection Methods

3 Barrier Methods

4 Variance reduction

5 Other methods

6 Conclusion

Other BCO methods

Strongly convex loss functions:

- Augment \mathcal{R} in BANDITFTRL with additional regularization.
- $\mathcal{C}^{0,1}$ [Agarwal et al, 2010]: $\mathcal{O}(T^{2/3})$ regret
- $\mathcal{C}^{1,1}$ [Hazan & Levy, 2014]: $\mathcal{O}(T^{1/2})$ regret

Other types of algorithms:

- Ellipsoid method-based algorithm [Hazan and Li, 2016]:
 $\mathcal{O}(2^{d^4} \log(T)^{2d} T^{1/2})$.
- Kernel-based algorithm [Bubeck et al, 2017]: $\mathcal{O}(d^{9.5} T^{1/2})$

Table of Contents

1 Bandit Convex Optimization (BCO)

2 Projection Methods

3 Barrier Methods

4 Variance reduction

5 Other methods

6 Conclusion

Conclusion

- BCO is a flexible framework for modeling learning problems with sequential data and very limited feedback.
- BCO generalizes many existing models of online learning and optimization.
- State-of-the-art algorithms leverage techniques from online convex optimization and interior-point methods.
- “Efficient” algorithms obtaining optimal guarantees in $\mathcal{C}^{0,1}$, $\mathcal{C}^{1,1}$ cases are still open.

Optimal Algorithm for the Contextual Bandit problem

Alekh Agarwal[†]

John Langford[†]

Daniel Hsu[‡]

Lihong Li[†]

Satyen Kale[#]

Rob Schapire[†]

[†]Microsoft Research, [‡]Columbia University,

[#]Google Research, ^{*}Princeton University

1. Introduction

Learning to interact: example #1

Loop:

1. Patient arrives with symptoms, medical history, genome ...
2. Physician prescribes treatment.
3. Patient's health responds (e.g., improves, worsens).

Goal: prescribe treatments that yield good health outcomes.

Learning to interact: example #2

Loop:

1. User visits website with profile, browsing history ...
2. Website operator chooses content/ads to display.
3. User reacts to content/ads (e.g., click, "like").

Goal: choose content/ads that yield desired user behavior.

Contextual bandit setting (i.i.d. version)

For $t = 1, 2, \dots, T$:

0. Nature draws (x_t, r_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^{\mathcal{A}}$.
1. Observe context x_t .
2. Choose action $a_t \in \mathcal{A}$.
3. Collect reward $r_t(a_t)$.

Task: Design an algorithm for choosing a_t 's that yield high reward.

Contextual bandit setting (i.i.d. version)

For $t = 1, 2, \dots, T$:

0. Nature draws (x_t, r_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^{\mathcal{A}}$.
1. Observe context x_t .
2. Choose action $a_t \in \mathcal{A}$.
3. Collect reward $r_t(a_t)$.

Task: Design an algorithm for choosing a_t 's that yield high reward.

Contextual setting: use features x_t to choose good actions a_t .

Contextual bandit setting (i.i.d. version)

For $t = 1, 2, \dots, T$:

0. Nature draws (x_t, \mathbf{r}_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^{\mathcal{A}}$.
1. Observe context x_t .
2. Choose action $a_t \in \mathcal{A}$.
3. Collect reward $r_t(a_t)$.

Task: Design an algorithm for choosing a_t 's that yield high reward.

Contextual setting: use features x_t to choose good actions a_t .

Bandit setting: $r_t(a)$ for $a \neq a_t$ is not observed.

⇒ Exploration vs. exploitation dilemma

(cf. non-bandit setting: whole reward vector $\mathbf{r}_t \in [0, 1]^{\mathcal{A}}$ observed.)

Learning objective in the contextual bandit setting

No single action is good in all situations: **must exploit context.**

Learning objective in the contextual bandit setting

No single action is good in all situations: **must exploit context.**

Policy class Π : set of functions (“policies”) $\pi : \mathcal{X} \rightarrow \mathcal{A}$
(e.g., advice of experts, linear classifiers, neural networks).

Learning objective in the contextual bandit setting

No single action is good in all situations: **must exploit context.**

Policy class Π : set of functions (“policies”) $\pi : \mathcal{X} \rightarrow \mathcal{A}$
(e.g., advice of experts, linear classifiers, neural networks).

Regret (i.e., relative performance) to a policy class Π :

$$\underbrace{\max_{\pi \in \Pi} \sum_{t=1}^T r_t(\pi(x_t))}_{\text{total reward of best policy}} - \underbrace{\sum_{t=1}^T r_t(a_t)}_{\text{total reward of learner}}$$

... a strong benchmark when Π contains a policy with high reward.

Learning objective in the contextual bandit setting

No single action is good in all situations: **must exploit context.**

Policy class Π : set of functions (“policies”) $\pi : \mathcal{X} \rightarrow \mathcal{A}$
(e.g., advice of experts, linear classifiers, neural networks).

Regret (i.e., relative performance) to a policy class Π :

$$\underbrace{\max_{\pi \in \Pi} \sum_{t=1}^T r_t(\pi(x_t))}_{\text{total reward of best policy}} - \underbrace{\sum_{t=1}^T r_t(a_t)}_{\text{total reward of learner}}$$

... a strong benchmark when Π contains a policy with high reward.

Regret is sublinear (in T) \implies (Avg.) per-round regret $\rightarrow 0$.

Challenge #1: computation

Feedback that learner observes: reward of chosen action $r_t(a_t)$
→ only directly relevant to $\pi \in \Pi$ s.t. $\pi(x_t) = a_t$.

Challenge #1: computation

Feedback that learner observes: reward of chosen action $r_t(a_t)$
→ only directly relevant to $\pi \in \Pi$ s.t. $\pi(x_t) = a_t$.

Separate explicit bookkeeping for each policy $\pi \in \Pi$ becomes
computationally intractable when Π is large (or infinite!).

Arg max oracle (AMO): supervised learning

In many cases, we know how to exploit structure of Π to design efficient algorithms or heuristics!

Arg max oracle (AMO): supervised learning

In many cases, we know how to exploit structure of Π to design efficient algorithms or heuristics!

Given **fully labeled** data $(x_1, \rho_1), \dots, (x_t, \rho_t) \in \mathcal{X} \times [0, 1]^{\mathcal{A}}$, the AMO returns

$$\arg \max_{\pi \in \Pi} \sum_{i=1}^t \rho_i(\pi(x_i)).$$

Arg max oracle (AMO): supervised learning

In many cases, we know how to exploit structure of Π to design efficient algorithms or heuristics!

Given **fully labeled** data $(x_1, \rho_1), \dots, (x_t, \rho_t) \in \mathcal{X} \times [0, 1]^{\mathcal{A}}$, the AMO returns

$$\arg \max_{\pi \in \Pi} \sum_{i=1}^t \rho_i(\pi(x_i)).$$

AMO is an abstraction for efficient search through Π .

Arg max oracle (AMO): supervised learning

In many cases, we know how to exploit structure of Π to design efficient algorithms or heuristics!

Given **fully labeled** data $(x_1, \rho_1), \dots, (x_t, \rho_t) \in \mathcal{X} \times [0, 1]^{\mathcal{A}}$, the AMO returns

$$\arg \max_{\pi \in \Pi} \sum_{i=1}^t \rho_i(\pi(x_i)).$$

AMO is an abstraction for efficient search through Π .

In practice: implement using standard heuristics—e.g., convex relaxations, backpropagation—for cost-sensitive multi-class learning.

Arg max oracle (AMO): supervised learning

In many cases, we know how to exploit structure of Π to design efficient algorithms or heuristics!

Given **fully labeled** data $(x_1, \rho_1), \dots, (x_t, \rho_t) \in \mathcal{X} \times [0, 1]^{\mathcal{A}}$, the AMO returns

$$\arg \max_{\pi \in \Pi} \sum_{i=1}^t \rho_i(\pi(x_i)).$$

AMO is an abstraction for efficient search through Π .

In practice: implement using standard heuristics—e.g., convex relaxations, backpropagation—for cost-sensitive multi-class learning.

But requires **complete reward vectors** ρ_i ; not directly usable for contextual bandits.

Challenge #2: exploration

Possible approach: AMO + **simple random exploration**

- 1: In first T_0 rounds, choose $a_t \in \mathcal{A}$ u.a.r. to get unbiased estimates $\hat{\mathbf{r}}_t$ of \mathbf{r}_t for all $t \in [T_0]$.
- 2: Get $\tilde{\pi} := \text{AMO}(\{(x_t, \hat{\mathbf{r}}_t)\}_{t \in [T_0]})$.
- 3: Use $a_t := \tilde{\pi}(x_t)$ in round $t > T_0$.

Challenge #2: exploration

Possible approach: AMO + **simple random exploration**

- 1: In first T_0 rounds, choose $a_t \in \mathcal{A}$ u.a.r. to get unbiased estimates $\hat{\mathbf{r}}_t$ of \mathbf{r}_t for all $t \in [T_0]$.
- 2: Get $\tilde{\pi} := \text{AMO}(\{(x_t, \hat{\mathbf{r}}_t)\}_{t \in [T_0]})$.
- 3: Use $a_t := \tilde{\pi}(x_t)$ in round $t > T_0$.

But $\mathbb{E}_{(x, \mathbf{r})}[r(\tilde{\pi}(x))] \approx \max_{\pi \in \Pi} \mathbb{E}_{(x, \mathbf{r})}[r(\pi(x))] - \Omega\left(\frac{1}{\sqrt{T_0}}\right)$

(Dependencies on $|\mathcal{A}|$ and $|\Pi|$ hidden.)

Challenge #2: exploration

Possible approach: AMO + **simple random exploration**

- 1: In first T_0 rounds, choose $a_t \in \mathcal{A}$ u.a.r. to get unbiased estimates $\hat{\mathbf{r}}_t$ of \mathbf{r}_t for all $t \in [T_0]$.
- 2: Get $\tilde{\pi} := \text{AMO}(\{(x_t, \hat{\mathbf{r}}_t)\}_{t \in [T_0]})$.
- 3: Use $a_t := \tilde{\pi}(x_t)$ in round $t > T_0$.

But $\mathbb{E}_{(x, \mathbf{r})}[r(\tilde{\pi}(x))] \approx \max_{\pi \in \Pi} \mathbb{E}_{(x, \mathbf{r})}[r(\pi(x))] - \Omega\left(\frac{1}{\sqrt{T_0}}\right)$
... so regret with this approach (with best T_0) could be as large as

$$\Omega\left(T_0 + \frac{1}{\sqrt{T_0}}(T - T_0)\right)$$

(Dependencies on $|\mathcal{A}|$ and $|\Pi|$ hidden.)

Challenge #2: exploration

Possible approach: AMO + **simple random exploration**

- 1: In first T_0 rounds, choose $a_t \in \mathcal{A}$ u.a.r. to get unbiased estimates $\hat{\mathbf{r}}_t$ of \mathbf{r}_t for all $t \in [T_0]$.
- 2: Get $\tilde{\pi} := \text{AMO}(\{(x_t, \hat{\mathbf{r}}_t)\}_{t \in [T_0]})$.
- 3: Use $a_t := \tilde{\pi}(x_t)$ in round $t > T_0$.

But $\mathbb{E}_{(x, \mathbf{r})}[r(\tilde{\pi}(x))] \approx \max_{\pi \in \Pi} \mathbb{E}_{(x, \mathbf{r})}[r(\pi(x))] - \Omega\left(\frac{1}{\sqrt{T_0}}\right)$
... so regret with this approach (with best T_0) could be as large as

$$\Omega\left(T_0 + \frac{1}{\sqrt{T_0}}(T - T_0)\right) \sim T^{2/3} \gg T^{1/2}.$$

(Dependencies on $|\mathcal{A}|$ and $|\Pi|$ hidden.)

Algorithms for contextual bandits

Let $K := |\mathcal{A}|$ and $N := |\Pi|$.

Our result [AHKLLS'14]: a new, fast and simple algorithm.

Optimal regret bound $\tilde{O}(\sqrt{KT \log N})$.

$\tilde{O}(\sqrt{TK})$ calls to AMO overall.

Algorithms for contextual bandits

Let $K := |\mathcal{A}|$ and $N := |\Pi|$.

Our result [AHKLLS'14]: a new, fast and simple algorithm.

Optimal regret bound $\tilde{O}(\sqrt{KT \log N})$.

$\tilde{O}(\sqrt{TK})$ calls to AMO overall.

Previous work:

[ACBFS'02] Exp4 algorithm (exponential weights).

Optimal regret bound $O(\sqrt{KT \log N})$.

Requires explicit enumeration of Π in every round.

Algorithms for contextual bandits

Let $K := |\mathcal{A}|$ and $N := |\Pi|$.

Our result [AHKLLS'14]: a new, fast and simple algorithm.

Optimal regret bound $\tilde{O}(\sqrt{KT \log N})$.

$\tilde{O}(\sqrt{TK})$ calls to AMO overall.

Previous work:

[ACBFS'02] Exp4 algorithm (exponential weights).

Optimal regret bound $O(\sqrt{KT \log N})$.

Requires explicit enumeration of Π in every round.

[LZ'07] ϵ -greedy variant (uniform exploration).

Suboptimal regret bound $\tilde{O}(T^{2/3}(K \log N)^{1/3})$.

One call to AMO overall.

Algorithms for contextual bandits

Let $K := |\mathcal{A}|$ and $N := |\Pi|$.

Our result [AHKLLS'14]: a new, fast and simple algorithm.

Optimal regret bound $\tilde{O}(\sqrt{KT \log N})$.

$\tilde{O}(\sqrt{TK})$ calls to AMO overall.

Previous work:

[ACBFS'02] Exp4 algorithm (exponential weights).

Optimal regret bound $O(\sqrt{KT \log N})$.

Requires explicit enumeration of Π in every round.

[LZ'07] ϵ -greedy variant (uniform exploration).

Suboptimal regret bound $\tilde{O}(T^{2/3}(K \log N)^{1/3})$.

One call to AMO overall.

[DHKKLRZ'11] “efficient” algorithm (careful exploration).

Optimal regret bound $\tilde{O}(\sqrt{KT \log N})$.

$O(T^6 K^4)$ calls to AMO overall.

Rest of the talk

Components of the new algorithm: Importance-weighted
Low-Variance Epoch-Timed Oracleized CONtextual BANDITS

1. “Classical” tricks: randomization, inverse probability weighting.
2. Efficient algorithm for balancing exploration/exploitation.
3. Additional tricks: warm-start and epoch structure.

Note: we assume (x_t, \mathbf{r}_t) i.i.d. from \mathcal{D}
(whereas Exp4 also works in adversarial setting).

Outline

1. Introduction
2. Classical tricks
3. Construction of good policy distributions
4. Additional tricks: warm-start and epoch structure

2. Classical tricks

What would've happened if I had done X?

For $t = 1, 2, \dots, T$:

0. Nature draws (x_t, r_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^{\mathcal{A}}$.
1. Observe context x_t .
2. Choose action $a_t \in \mathcal{A}$.
3. Collect reward $r_t(a_t)$.

What would've happened if I had done X?

For $t = 1, 2, \dots, T$:

0. Nature draws (x_t, r_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^{\mathcal{A}}$.
1. Observe context x_t .
2. Choose action $a_t \in \mathcal{A}$.
3. Collect reward $r_t(a_t)$.

Q: How do I learn about $r_t(a)$ for actions a I don't actually take?

What would've happened if I had done X?

For $t = 1, 2, \dots, T$:

0. Nature draws (x_t, r_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^{\mathcal{A}}$.
1. Observe context x_t .
2. Choose action $a_t \in \mathcal{A}$.
3. Collect reward $r_t(a_t)$.

Q: How do I learn about $r_t(a)$ for actions a I don't actually take?

A: Randomize. Draw $a_t \sim p_t$ for some pre-specified prob. dist. p_t .

Inverse probability weighting

Importance-weighted estimate of reward from round t :

$$\forall a \in \mathcal{A} . \quad \hat{r}_t(a) := \frac{r_t(a_t) \cdot \mathbb{1}\{a = a_t\}}{p_t(a_t)}$$

Inverse probability weighting

Importance-weighted estimate of reward from round t :

$$\forall a \in \mathcal{A} . \quad \hat{r}_t(a) := \frac{r_t(a_t) \cdot \mathbb{1}\{a = a_t\}}{p_t(a_t)} = \begin{cases} \frac{r_t(a_t)}{p_t(a_t)} & \text{if } a = a_t \\ 0 & \text{otherwise} \end{cases}$$

Inverse probability weighting

Importance-weighted estimate of reward from round t :

$$\forall a \in \mathcal{A} . \quad \hat{r}_t(a) := \frac{r_t(a_t) \cdot \mathbb{1}\{a = a_t\}}{p_t(a_t)} = \begin{cases} \frac{r_t(a_t)}{p_t(a_t)} & \text{if } a = a_t \\ 0 & \text{otherwise} \end{cases}$$

Unbiasedness:

$$\mathbb{E}_{a_t \sim p_t} [\hat{r}_t(a)] = \sum_{a' \in \mathcal{A}} p_t(a') \cdot \frac{r_t(a') \cdot \mathbb{1}\{a = a'\}}{p_t(a')} = r_t(a).$$

Inverse probability weighting

Importance-weighted estimate of reward from round t :

$$\forall a \in \mathcal{A}. \quad \hat{r}_t(a) := \frac{r_t(a_t) \cdot \mathbb{1}\{a = a_t\}}{p_t(a_t)} = \begin{cases} \frac{r_t(a_t)}{p_t(a_t)} & \text{if } a = a_t \\ 0 & \text{otherwise} \end{cases}$$

Unbiasedness:

$$\mathbb{E}_{a_t \sim p_t} [\hat{r}_t(a)] = \sum_{a' \in \mathcal{A}} p_t(a') \cdot \frac{r_t(a') \cdot \mathbb{1}\{a = a'\}}{p_t(a')} = r_t(a).$$

Range and variance: upper-bounded by $1/p_t(a)$.

Inverse probability weighting

Importance-weighted estimate of reward from round t :

$$\forall a \in \mathcal{A}. \quad \hat{r}_t(a) := \frac{r_t(a_t) \cdot \mathbb{1}\{a = a_t\}}{p_t(a_t)} = \begin{cases} \frac{r_t(a_t)}{p_t(a_t)} & \text{if } a = a_t \\ 0 & \text{otherwise} \end{cases}$$

Unbiasedness:

$$\mathbb{E}_{a_t \sim p_t} [\hat{r}_t(a)] = \sum_{a' \in \mathcal{A}} p_t(a') \cdot \frac{r_t(a') \cdot \mathbb{1}\{a = a'\}}{p_t(a')} = r_t(a).$$

Range and variance: upper-bounded by $1/p_t(a)$.

Expected reward of policy: $\text{Rew}(\pi) = \mathbb{E}_{(x, r)}[r(\pi(x))]$

Unbiased estimator of total reward: $\widehat{\text{Rew}}_t(\pi) := \sum_{i=1}^t \hat{r}_i(\pi(x_i))$.

Inverse probability weighting

Importance-weighted estimate of reward from round t :

$$\forall a \in \mathcal{A}. \quad \hat{r}_t(a) := \frac{r_t(a_t) \cdot \mathbb{1}\{a = a_t\}}{p_t(a_t)} = \begin{cases} \frac{r_t(a_t)}{p_t(a_t)} & \text{if } a = a_t \\ 0 & \text{otherwise} \end{cases}$$

Unbiasedness:

$$\mathbb{E}_{a_t \sim p_t} [\hat{r}_t(a)] = \sum_{a' \in \mathcal{A}} p_t(a') \cdot \frac{r_t(a') \cdot \mathbb{1}\{a = a'\}}{p_t(a')} = r_t(a).$$

Range and variance: upper-bounded by $1/p_t(a)$.

Expected reward of policy: $\text{Rew}(\pi) = \mathbb{E}_{(x, r)}[r(\pi(x))]$

Unbiased estimator of total reward: $\widehat{\text{Rew}}_t(\pi) := \sum_{i=1}^t \hat{r}_i(\pi(x_i))$.

How should we choose the p_t ?

Hedging over policies

Get action distributions via policy distributions.

$$\underbrace{(\mathbf{W}, \mathbf{x})}_{\text{(policy distribution, context)}} \mapsto \underbrace{\mathbf{p}}_{\text{action distribution}}$$

Hedging over policies

Get action distributions via policy distributions.

$$\underbrace{(\mathbf{W}, x)}_{\text{(policy distribution, context)}} \mapsto \underbrace{\mathbf{p}}_{\text{action distribution}}$$

Policy distribution: $\mathbf{W} = (W(\pi) : \pi \in \Pi)$
probability dist. over policies π in the policy class Π

Hedging over policies

Get action distributions via policy distributions.

$$\underbrace{(\mathbf{W}, x)}_{\text{(policy distribution, context)}} \mapsto \underbrace{\mathbf{p}}_{\text{action distribution}}$$

- 1: Pick initial distribution \mathbf{W}_1 over policies Π .
- 2: **for round** $t = 1, 2, \dots$ **do**
- 3: Nature draws (x_t, r_t) from dist. \mathcal{D} over $\mathcal{X} \times [0, 1]^A$.
- 4: Observe context x_t .
- 5: Compute distribution \mathbf{p}_t over \mathcal{A} (using \mathbf{W}_t and x_t).
- 6: Pick action $a_t \sim \mathbf{p}_t$.
- 7: Collect reward $r_t(a_t)$.
- 8: Compute new distribution \mathbf{W}_{t+1} over policies Π .
- 9: **end for**

Projections of policy distributions

Given policy distribution \mathbf{W} and context x ,

$$\forall a \in \mathcal{A} . \quad W(a|x) := \sum_{\pi \in \Pi} W(\pi) \cdot \mathbb{1}\{\pi(x) = a\}$$

(so $\mathbf{W} \mapsto \mathbf{W}(\cdot|x)$ is a linear map).

Projections of policy distributions

Given policy distribution \mathbf{W} and context x ,

$$\forall a \in \mathcal{A} . \quad W(a|x) := \sum_{\pi \in \Pi} W(\pi) \cdot \mathbb{1}\{\pi(x) = a\}$$

(so $\mathbf{W} \mapsto \mathbf{W}(\cdot|x)$ is a linear map).

We actually use

$$\mathbf{p}_t := \mathbf{W}_t^{\mu_t}(\cdot|x_t) := (1 - K\mu_t)\mathbf{W}_t(\cdot|x_t) + \mu_t$$

so every action has probability at least μ_t (*to be determined*).

Basic algorithm structure

```
1: Pick initial distribution  $\mathbf{W}_1$  over policies  $\Pi$ .  
2: for round  $t = 1, 2, \dots$  do  
3:   Nature draws  $(x_t, r_t)$  from dist.  $\mathcal{D}$  over  $\mathcal{X} \times [0, 1]^{\mathcal{A}}$ .  
4:   Observe context  $x_t$ .  
5:   Compute action distribution  $\mathbf{p}_t := \mathbf{W}_t^{\mu_t}(\cdot | x_t)$ .  
6:   Pick action  $a_t \sim \mathbf{p}_t$ .  
7:   Collect reward  $r_t(a_t)$ .  
8:   Compute new distribution  $\mathbf{W}_{t+1}$  over policies  $\Pi$ .  
9: end for
```

Q: How do we choose \mathbf{W}_t for good exploration/exploitation?

Basic algorithm structure

```
1: Pick initial distribution  $\mathbf{W}_1$  over policies  $\Pi$ .  
2: for round  $t = 1, 2, \dots$  do  
3:   Nature draws  $(x_t, r_t)$  from dist.  $\mathcal{D}$  over  $\mathcal{X} \times [0, 1]^{\mathcal{A}}$ .  
4:   Observe context  $x_t$ .  
5:   Compute action distribution  $\mathbf{p}_t := \mathbf{W}_t^{\mu_t}(\cdot | x_t)$ .  
6:   Pick action  $a_t \sim \mathbf{p}_t$ .  
7:   Collect reward  $r_t(a_t)$ .  
8:   Compute new distribution  $\mathbf{W}_{t+1}$  over policies  $\Pi$ .  
9: end for
```

Q: How do we choose \mathbf{W}_t for good exploration/exploitation?

Caveat: \mathbf{W}_t must be efficiently computable + representable!

3. Construction of good policy distributions

Our approach

- ▶ Define convex feasibility problem (over distributions \mathcal{W} on Π) such that solutions yield optimal regret bounds.

Our approach

- ▶ Define convex feasibility problem (over distributions \mathbf{W} on Π) such that solutions yield optimal regret bounds.
- ▶ Design algorithm that finds a *sparse* solution \mathbf{W} .

Our approach

- ▶ Define convex feasibility problem (over distributions \mathbf{W} on Π) such that solutions yield optimal regret bounds.
- ▶ Design algorithm that finds a *sparse* solution \mathbf{W} .

Algorithm only accesses Π via calls to AMO

$$\implies \text{nnz}(\mathbf{W}) = \# \text{ calls to AMO}$$

An optimal but inefficient algorithm

Policy_Elimination

Let $\Pi_1 = \Pi$.

An optimal but inefficient algorithm

Policy_Elimination

Let $\Pi_1 = \Pi$. For each $t = 1, 2, \dots$:

1. Choose distribution W_t over Π_t such that

$$\forall \pi \in \Pi_t : \quad \mathbb{E}_x \left[\frac{1}{W_t(\pi(x)|x)} \right] \leq K$$

An optimal but inefficient algorithm

Policy_Elimination

Let $\Pi_1 = \Pi$. For each $t = 1, 2, \dots$:

1. Choose distribution W_t over Π_t such that

$$\forall \pi \in \Pi_t : \quad \mathbb{E}_x \left[\frac{1}{W_t(\pi(x)|x)} \right] \leq K$$

2. Let $\overline{\text{Rew}}_t(\pi) = \frac{1}{t} \widehat{\text{Rew}}_t(\pi)$, i.e. the average of all the estimators for $\text{Rew}(\pi)$ so far. Let

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta \left(\frac{1}{\sqrt{t}} \right) \right\}$$

Analysis Sketch: Distribution Selection Step

Choose distribution W_t over Π_t such that

$$\forall \pi \in \Pi_t : \mathbb{E}_x \left[\frac{1}{W_t(\pi(x)|x)} \right] \leq K$$

Analysis Sketch: Distribution Selection Step

Choose distribution W_t over Π_t such that

$$\forall \pi \in \Pi_t : \mathbb{E}_x \left[\frac{1}{W_t(\pi(x)|x)} \right] \leq K$$

- ▶ Ensures that $\forall \pi \in \Pi_t :$

$$\mathbf{Var}_x[\hat{r}_t(\pi(x_t))] \leq O(1).$$

Analysis Sketch: Distribution Selection Step

Choose distribution W_t over Π_t such that

$$\forall \pi \in \Pi_t : \mathbb{E}_x \left[\frac{1}{W_t(\pi(x)|x)} \right] \leq K$$

- ▶ Ensures that $\forall \pi \in \Pi_t :$

$$\mathbf{Var}_x[\hat{r}_t(\pi(x_t))] \leq O(1).$$

- ▶ Hence, averaging over t iterations, we have $\forall \pi \in \Pi_t :$

$$\mathbf{Var}_x[\overline{\text{Rew}}_t(\pi)] \leq O\left(\frac{1}{t}\right).$$

Analysis Sketch: Distribution Selection Step

Choose distribution W_t over Π_t such that

$$\forall \pi \in \Pi_t : \mathbb{E}_x \left[\frac{1}{W_t(\pi(x)|x)} \right] \leq K$$

- ▶ Ensures that $\forall \pi \in \Pi_t :$

$$\mathbf{Var}_x[\hat{r}_t(\pi(x_t))] \leq O(1).$$

- ▶ Hence, averaging over t iterations, we have $\forall \pi \in \Pi_t :$

$$\mathbf{Var}_x[\overline{\text{Rew}}_t(\pi)] \leq O\left(\frac{1}{t}\right).$$

- ▶ Martingale concentration bounds imply that w.h.p. $\forall \pi \in \Pi_t :$

$$|\overline{\text{Rew}}_t(\pi) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

Analysis Sketch: Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

Analysis Sketch: Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

- W.h.p. $\forall \pi \in \Pi_t$:

$$|\overline{\text{Rew}}_t(\pi) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

Analysis Sketch: Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

- W.h.p. $\forall \pi \in \Pi_t$:

$$|\overline{\text{Rew}}_t(\pi) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

- W.h.p. $\forall \pi \in \Pi_{t+1}$:

$$|\text{Rew}(\pi^*) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

Analysis Sketch: Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

- W.h.p. $\forall \pi \in \Pi_t$:

$$|\overline{\text{Rew}}_t(\pi) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

- W.h.p. $\forall \pi \in \Pi_{t+1}$:

$$|\text{Rew}(\pi^*) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

- Thus, expected regret in time $t + 1$ is $O(\frac{1}{\sqrt{t}})$.

Analysis Sketch: Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

- W.h.p. $\forall \pi \in \Pi_t$:

$$|\overline{\text{Rew}}_t(\pi) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

- W.h.p. $\forall \pi \in \Pi_{t+1}$:

$$|\text{Rew}(\pi^*) - \text{Rew}(\pi)| \leq O\left(\frac{1}{\sqrt{t}}\right).$$

- Thus, expected regret in time $t + 1$ is $O(\frac{1}{\sqrt{t}})$.

- Thus, total regret is $\sum_{t=1}^T O(\frac{1}{\sqrt{t}}) = O(\sqrt{T})$.

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Why should such a distribution exist?

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Why should such a distribution exist?

Answer: Minimax magic.

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Why should such a distribution exist?

Answer: Minimax magic.

$$\min_W \max_{\pi} \mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] = \min_W \max_U \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right]$$

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Why should such a distribution exist?

Answer: Minimax magic.

$$\begin{aligned} \min_W \max_{\pi} \mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] &= \min_W \max_U \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right] \\ &= \max_U \min_W \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right] \end{aligned}$$

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Why should such a distribution exist?

Answer: Minimax magic.

$$\begin{aligned} \min_W \max_{\pi} \mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] &= \min_W \max_U \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right] \\ &= \max_U \min_W \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right] \\ &\leq \max_U \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{U(\pi(x)|x)} \right] \end{aligned}$$

Existence of Distribution

Key step: Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Why should such a distribution exist?

Answer: Minimax magic.

$$\begin{aligned} \min_W \max_{\pi} \mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] &= \min_W \max_U \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right] \\ &= \max_U \min_W \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{W(\pi(x)|x)} \right] \\ &\leq \max_U \mathbb{E}_{x, \pi \sim U} \left[\frac{1}{U(\pi(x)|x)} \right] \\ &= \max_U \mathbb{E}_x \left[\sum_{a \in [K]} \frac{U(a|x)}{U(a|x)} \right] \leq K. \end{aligned}$$

Problems with the algorithm

Distribution Selection Step

Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

Problems with the algorithm

Distribution Selection Step

Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

- ▶ Computing P is a convex optimization problem and takes $\text{poly}(N)$ time.

Problems with the algorithm

Distribution Selection Step

Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

- ▶ Computing P is a convex optimization problem and takes $\text{poly}(N)$ time.
- ▶ Computing P requires knowledge of actual data distribution.

Problems with the algorithm

Distribution Selection Step

Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

- ▶ Computing P is a convex optimization problem and takes $\text{poly}(N)$ time.
- ▶ Computing P requires knowledge of actual data distribution.

Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

Problems with the algorithm

Distribution Selection Step

Choose W s.t. $\forall \pi \in \Pi_t$, we have $\mathbb{E}_x \left[\frac{1}{W(\pi(x)|x)} \right] \leq K$.

- ▶ Computing P is a convex optimization problem and takes $\text{poly}(N)$ time.
- ▶ Computing P requires knowledge of actual data distribution.

Policy Elimination Step

$$\Pi_{t+1} = \left\{ \pi \in \Pi_t : \overline{\text{Rew}}_t(\pi) \geq \max_{\pi' \in \Pi_t} \overline{\text{Rew}}_t(\pi') - \Theta\left(\frac{1}{\sqrt{t}}\right) \right\}$$

- ▶ Policy Elimination Step takes $\Omega(N)$ time.

Properties of a good policy distribution

Low Regret and Low Variance constraints on W :

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq \sqrt{Kt \log N}, \quad (\text{LR})$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{\sqrt{Kt \log N}} \right) \quad \forall \pi \in \Pi \quad (\text{LV})$$

$$\widehat{\text{Reg}}_t(\pi) := \max_{\pi' \in \Pi} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi), \quad \mu_t := \sqrt{\frac{\log N}{Kt}}, \quad H_t := (x_1, \dots, x_t)$$

Properties of a good policy distribution

Low Regret and Low Variance constraints on W :

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq \sqrt{Kt \log N}, \quad (\text{LR})$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{\sqrt{Kt \log N}} \right) \quad \forall \pi \in \Pi \quad (\text{LV})$$

$$\widehat{\text{Reg}}_t(\pi) := \max_{\pi' \in \Pi} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi), \quad \mu_t := \sqrt{\frac{\log N}{Kt}}, \quad H_t := (x_1, \dots, x_t)$$

Intuition: Allow higher variance for policies π with larger regret, as they should have low weight anyway.

Properties of a good policy distribution

Low Regret and Low Variance constraints on W :

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq Kt \cdot \mu_t, \quad (\text{LR})$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{Kt \cdot \mu_t} \right) \quad \forall \pi \in \Pi \quad (\text{LV})$$

$$\widehat{\text{Reg}}_t(\pi) := \max_{\pi' \in \Pi} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi), \quad \mu_t := \sqrt{\frac{\log N}{Kt}}, \quad H_t := (x_1, \dots, x_t)$$

Intuition: Allow higher variance for policies π with larger regret, as they should have low weight anyway.

Properties of a good policy distribution

Low Regret and Low Variance constraints on W :

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq Kt \cdot \mu_t, \quad (\text{LR})$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{Kt \cdot \mu_t} \right) \quad \forall \pi \in \Pi \quad (\text{LV})$$

$$\widehat{\text{Reg}}_t(\pi) := \max_{\pi' \in \Pi} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi), \quad \mu_t := \sqrt{\frac{\log N}{Kt}}, \quad H_t := (x_1, \dots, x_t)$$

$$(\text{LV}) \implies \text{Reg}(\pi) \leq O\left(\widehat{\text{Reg}}_t(\pi) + Kt \cdot \mu_t\right) \quad \forall \pi \in \Pi;$$

$$(\text{LR}, \text{LV}) \implies \sum_{\pi \in \Pi} W_t(\pi) \cdot \text{Reg}(\pi) \leq O(Kt \cdot \mu_t).$$

Properties of a good policy distribution

Low Regret and Low Variance constraints on \mathbf{W} :

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq Kt \cdot \mu_t, \quad (\text{LR})$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{Kt \cdot \mu_t} \right) \quad \forall \pi \in \Pi \quad (\text{LV})$$

$$\widehat{\text{Reg}}_t(\pi) := \max_{\pi' \in \Pi} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi), \quad \mu_t := \sqrt{\frac{\log N}{Kt}}, \quad H_t := (x_1, \dots, x_t)$$

Theorem: If we pick \mathbf{W}_t satisfying (LR,LV) in every round t ,
then regret over all T rounds is $O\left(\sqrt{KT \log N}\right)$.

Properties of a good policy distribution

Low Regret and Low Variance constraints on \mathbf{W} :

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq Kt \cdot \mu_t, \quad (\text{LR})$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{Kt \cdot \mu_t} \right) \quad \forall \pi \in \Pi \quad (\text{LV})$$

$$\widehat{\text{Reg}}_t(\pi) := \max_{\pi' \in \Pi} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi), \quad \mu_t := \sqrt{\frac{\log N}{Kt}}, \quad H_t := (x_1, \dots, x_t)$$

Theorem: If we pick \mathbf{W}_t satisfying (LR,LV) in every round t ,
then regret over all T rounds is $O\left(\sqrt{KT \log N}\right)$.

Critical question: Is it even feasible to satisfy (LR,LV)?

Minmax proof of feasibility (simplified)

$$\sum_{\pi \in \Pi} W(\pi) \cdot \widehat{\text{Reg}}_t(\pi) \leq \sqrt{Kt \log N},$$

$$\widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W(\pi(x)|x)} \right] \leq K \left(1 + \frac{\widehat{\text{Reg}}_t(\pi)}{\sqrt{Kt \log N}} \right) \quad \forall \pi \in \Pi$$

Minmax proof of feasibility (simplified)

$$\sum_{\pi \in \Pi} b(\pi) W(\pi) - 1 \leq 0,$$

$$\frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W(\pi(x)|x)} \right] - (1 + b(\pi)) \leq 0 \quad \forall \pi \in \Pi$$

$$b(\pi) := \widehat{\text{Reg}}_t(\pi) / \sqrt{Kt \log N}$$

Minmax proof of feasibility (simplified)

$$\begin{aligned} & \min_{\mathbf{W} \in \Delta^N} \max_{(\mathbf{U}_o, \mathbf{U}) \in \Delta^{N+1}} \mathbf{U}_o \left(\sum_{\pi \in \Pi} b(\pi) \mathbf{W}(\pi) - 1 \right) \\ & + \sum_{\pi \in \Pi} \mathbf{U}(\pi) \left(\frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{\mathbf{W}(\pi(x)|x)} \right] - (1 + b(\pi)) \right) \leq 0 \end{aligned}$$

$$b(\pi) := \widehat{\text{Reg}}_t(\pi) / \sqrt{Kt \log N}$$

Minmax proof of feasibility (simplified)

$$\begin{aligned} & \max_{(U_o, U) \in \Delta^{N+1}} \min_{W \in \Delta^N} U_o \left(\sum_{\pi \in \Pi} b(\pi) W(\pi) - 1 \right) \\ & + \sum_{\pi \in \Pi} U(\pi) \left(\frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W(\pi(x)|x)} \right] - (1 + b(\pi)) \right) \leq 0 \end{aligned}$$

$$b(\pi) := \widehat{\text{Reg}}_t(\pi) / \sqrt{Kt \log N}$$

Minmax proof of feasibility (simplified)

$$\begin{aligned} & \max_{(U_o, U) \in \Delta^{N+1}} \min_{W \in \Delta^N} U_o \left(\sum_{\pi \in \Pi} b(\pi) W(\pi) - 1 \right) \\ & + \sum_{\pi \in \Pi} U(\pi) \left(\frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{W(\pi(x)|x)} \right] - (1 + b(\pi)) \right) \leq 0 \end{aligned}$$

$$b(\pi) := \widehat{\text{Reg}}_t(\pi) / \sqrt{Kt \log N}$$

Choose $W := U + U_o \mathbf{1}^{\hat{\pi}}$ for $\hat{\pi} := \arg \min_{\pi \in \Pi} b(\pi)$
to verify that value of game ≤ 0 .

Minmax proof of feasibility (simplified)

Choose $\mathbf{W} := \mathbf{U} + \mathbf{U}_o \mathbf{1}^{\hat{\pi}}$ for $\hat{\pi} := \arg \min_{\pi \in \Pi} b(\pi)$ (so $b(\hat{\pi}) = 0$)
to verify that value of game ≤ 0 .

$$\begin{aligned} & \max_{(\mathbf{U}_o, \mathbf{U}) \in \Delta^{N+1}} \min_{\mathbf{W} \in \Delta^N} \mathbf{U}_o \left(\sum_{\pi \in \Pi} b(\pi) \mathbf{W}(\pi) - 1 \right) \\ & + \sum_{\pi \in \Pi} \mathbf{U}(\pi) \left(\frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\frac{1}{\mathbf{W}(\pi(x)|x)} \right] - (1 + b(\pi)) \right) \end{aligned}$$

Minmax proof of feasibility (simplified)

Choose $\mathbf{W} := \mathbf{U} + \mathbf{U}_o \mathbf{1}^{\hat{\pi}}$ for $\hat{\pi} := \arg \min_{\pi \in \Pi} b(\pi)$ (so $b(\hat{\pi}) = 0$)
to verify that value of game ≤ 0 .

$$\begin{aligned} & \max_{(\mathbf{U}_o, \mathbf{U}) \in \Delta^{N+1}} \mathbf{U}_o \left(\sum_{\pi \in \Pi} b(\pi) \mathbf{U}(\pi) - 1 \right) \\ & + \frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\sum_{a \in \mathcal{A}} \frac{\mathbf{U}(a|x)}{\mathbf{W}(a|x)} \right] - \sum_{\pi \in \Pi} \mathbf{U}(\pi) (1 + b(\pi)) \end{aligned}$$

Minmax proof of feasibility (simplified)

Choose $\mathbf{W} := \mathbf{U} + \mathbf{U}_o \mathbf{1}^{\hat{\pi}}$ for $\hat{\pi} := \arg \min_{\pi \in \Pi} b(\pi)$ (so $b(\hat{\pi}) = 0$)
to verify that value of game ≤ 0 .

$$\begin{aligned} & \max_{(\mathbf{U}_o, \mathbf{U}) \in \Delta^{N+1}} (\mathbf{U}_o - 1) \sum_{\pi \in \Pi} b(\pi) \mathbf{U}(\pi) \\ & \quad + \frac{1}{K} \widehat{\mathbb{E}}_{x \in H_t} \left[\sum_{a \in \mathcal{A}} \frac{\mathbf{U}(a|x)}{\mathbf{W}(a|x)} \right] - 1 \leq 0 \end{aligned}$$

Feasibility and sparsity

Feasibility of LR/LV constraints is implied by minimax argument.

Feasibility and sparsity

Feasibility of LR/LV constraints is implied by minimax argument.

“Monster” solution [DHKKLRZ’11]: Can solve (variant) of feasibility problem using Ellipsoid algorithm
(where separation oracle = AMO + Perceptron + another Ellipsoid).

Feasibility and sparsity

Feasibility of LR/LV constraints is implied by minimax argument.

“Monster” solution [DHKKLRZ’11]: Can solve (variant) of feasibility problem using Ellipsoid algorithm
(where separation oracle = AMO + Perceptron + another Ellipsoid).

Existence of sparse(r) solution: given any (dense) solution, *probabilistic method* shows that there is an $\tilde{O}(\sqrt{Kt})$ -sparse approximation with comparable LR and LV constraint bounds.

Feasibility and sparsity

Feasibility of LR/LV constraints is implied by minimax argument.

“Monster” solution [DHKKLRZ’11]: Can solve (variant) of feasibility problem using Ellipsoid algorithm
(where separation oracle = AMO + Perceptron + another Ellipsoid).

Existence of sparse(r) solution: given any (dense) solution, *probabilistic method* shows that there is an $\tilde{O}(\sqrt{Kt})$ -sparse approximation with comparable LR and LV constraint bounds.

Efficient construction via “boosting”-type algorithm?

Coordinate descent algorithm

```
input Initial weights  $W$ .  
1: loop  
2:   If (LR) is violated, then replace  $W$  by  $cW$ .  
3:   if there is a policy  $\pi \in \Pi$  causing (LV) to be violated  
    then  
4:     set  $W(\pi) := W(\pi) + \alpha$ .  
5:   else  
6:     Halt and return  $W$ .  
7:   end if  
8: end loop
```

(Both $0 < c < 1$ and $\alpha > 0$ have closed form expressions.)

(Technical detail: actually optimize over subdistributions that may sum to < 1 .)

Implementation via AMO

Checking violation of (LV) constraint: for all $\pi \in \Pi$,

$$\widehat{\mathbb{E}}_x \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\max_{\pi'} \widehat{\text{Rew}}_t(\pi') - \widehat{\text{Rew}}_t(\pi)}{Kt \cdot \mu_t} \right)$$

Implementation via AMO

Checking violation of (LV) constraint: for all $\pi \in \Pi$,

$$\frac{\widehat{\text{Rew}}_t(\pi)}{t \cdot \mu_t} + \widehat{\mathbb{E}}_x \left[\frac{1}{W^{\mu_t}(\pi(x)|x)} \right] \leq K \left(1 + \frac{\max_{\pi'} \widehat{\text{Rew}}_t(\pi')}{Kt \cdot \mu_t} \right)$$

Implementation via AMO

Checking violation of (LV) constraint: for all $\pi \in \Pi$,

$$\widehat{\text{Rew}}_t(\pi) + t \cdot \widehat{\mathbb{E}}_x \left[\frac{\mu_t}{W^{\mu_t}(\pi(x)|x)} \right] \leq Kt \cdot \mu_t + \max_{\pi'} \widehat{\text{Rew}}_t(\pi')$$

Implementation via AMO

Checking violation of (LV) constraint: for all $\pi \in \Pi$,

$$\widehat{\text{Rew}}_t(\pi) + t \cdot \widehat{\mathbb{E}}_x \left[\frac{\mu_t}{W^{\mu_t}(\pi(x)|x)} \right] \leq Kt \cdot \mu_t + \max_{\pi'} \widehat{\text{Rew}}_t(\pi')$$

1. Obtain $\hat{\pi} := \text{AMO}((x_1, \hat{\mathbf{r}}_1), \dots, (x_t, \hat{\mathbf{r}}_t))$.
2. Create fictitious rewards for each $i = 1, 2, \dots, t$:

$$\tilde{r}_i(a) := \frac{\mu}{W^{\mu_t}(a|x_i)} + \hat{r}_i(a) \quad \forall a \in \mathcal{A}.$$

Obtain $\tilde{\pi} := \text{AMO}((x_1, \tilde{\mathbf{r}}_1), \dots, (x_t, \tilde{\mathbf{r}}_t))$.

3. $\widetilde{\text{Rew}}_t(\tilde{\pi}) > Kt \cdot \mu_t + \widehat{\text{Rew}}_t(\hat{\pi})$ iff (LV) is violated by $\tilde{\pi}$.

Iteration bound for coordinate descent

Using unnormalized relative entropy-based potential function

$$\Phi(W) := t\mu_t \left(\frac{\widehat{\mathbb{E}}_{x \in H_t} [\text{RE}(\text{unif} \| W^{\mu_t}(\cdot|x))]}{1 - K\mu_t} + \frac{\sum_{\pi \in \Pi} W(\pi) \widehat{\text{Reg}}_t(\pi)}{Kt \cdot \mu_t} \right),$$

can show coordinate descent returns a feasible solution after

$$\tilde{O}\left(\frac{1}{\mu_t}\right) = \tilde{O}\left(\sqrt{\frac{Kt}{\log N}}\right) \text{ steps.}$$

(Every step decreases potential by about $t \cdot \mu_t^2 = \frac{\log N}{K}$.)

Recap

Recap

Low Regret / Low Variance constraints:

implies $\tilde{O}(\sqrt{KT \log N})$ regret bound.

Recap

Low Regret / Low Variance constraints:

implies $\tilde{O}(\sqrt{KT \log N})$ regret bound.

Coordinate descent to solve LR/LV constraints:

repeatedly find a violated constraint and adjust W to satisfy it.

Recap

Low Regret / Low Variance constraints:

implies $\tilde{O}(\sqrt{KT \log N})$ regret bound.

Coordinate descent to solve LR/LV constraints:

repeatedly find a violated constraint and adjust W to satisfy it.

Coordinate descent analysis:

In round t ,

$$\text{nnz}(W_t) = O(\# \text{ calls to } \arg \max_{\pi \in \Pi} \text{oracle}) = \tilde{O}\left(\sqrt{\frac{Kt}{\log N}}\right)$$

(same as guarantee via probabilistic method).

4. Additional tricks: warm-start and epoch structure

Total complexity over all rounds

In round t , coordinate descent for computing \mathbf{W}_t requires

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{Kt}{\log N}}\right) \text{ AMO calls.}$$

Total complexity over all rounds

In round t , coordinate descent for computing \mathbf{W}_t requires

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{Kt}{\log N}}\right) \text{ AMO calls.}$$

To compute \mathbf{W}_t in all rounds $t = 1, 2, \dots, T$, need

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{K}{\log N}} T^{1.5}\right) \text{ AMO calls over } T \text{ rounds.}$$

Warm start

To compute \mathbf{W}_{t+1} using coordinate descent, initialize with \mathbf{W}_t .

Warm start

To compute \mathbf{W}_{t+1} using coordinate descent, initialize with \mathbf{W}_t .

1. Total epoch-to-epoch increase in potential is $\tilde{O}(\sqrt{T/K})$ over all T rounds (w.h.p.—exploiting i.i.d. assumption).

Warm start

To compute \mathbf{W}_{t+1} using coordinate descent, initialize with \mathbf{W}_t .

1. Total epoch-to-epoch increase in potential is $\tilde{O}(\sqrt{T/K})$ over all T rounds (w.h.p.—exploiting i.i.d. assumption).
2. Each coordinate descent step decreases potential by $\Omega\left(\frac{\log N}{K}\right)$.

Warm start

To compute \mathbf{W}_{t+1} using coordinate descent, initialize with \mathbf{W}_t .

1. Total epoch-to-epoch increase in potential is $\tilde{O}(\sqrt{T/K})$ over all T rounds (w.h.p.—exploiting i.i.d. assumption).
2. Each coordinate descent step decreases potential by $\Omega\left(\frac{\log N}{K}\right)$.
3. Over all T rounds,

$$\text{total \# calls to AMO} \leq \tilde{O}\left(\sqrt{\frac{KT}{\log N}}\right)$$

Warm start

To compute \mathbf{W}_{t+1} using coordinate descent, initialize with \mathbf{W}_t .

1. Total epoch-to-epoch increase in potential is $\tilde{O}(\sqrt{T/K})$ over all T rounds (w.h.p.—exploiting i.i.d. assumption).
2. Each coordinate descent step decreases potential by $\Omega\left(\frac{\log N}{K}\right)$.
3. Over all T rounds,

$$\text{total \# calls to AMO} \leq \tilde{O}\left(\sqrt{\frac{KT}{\log N}}\right)$$

But still need an AMO call to even check if \mathbf{W}_t is feasible!

Epoch trick

Regret analysis: \mathbf{W}_t has low instantaneous per-round regret (roughly $K\mu_t$)—this also crucially relies on i.i.d. assumption.

Epoch trick

Regret analysis: W_t has low instantaneous per-round regret (roughly $K\mu_t$)—this also crucially relies on i.i.d. assumption.

⇒ same W_t can be used for $O(t)$ more rounds!

Epoch trick

Regret analysis: \mathbf{W}_t has low instantaneous per-round regret (roughly $K\mu_t$)—this also crucially relies on i.i.d. assumption.

⇒ same \mathbf{W}_t can be used for $O(t)$ more rounds!

Epoch trick: split T rounds into epochs, only compute \mathbf{W}_t at start of each epoch.

Epoch trick

Regret analysis: \mathbf{W}_t has low instantaneous per-round regret (roughly $K\mu_t$)—this also crucially relies on i.i.d. assumption.

⇒ same \mathbf{W}_t can be used for $O(t)$ more rounds!

Epoch trick: split T rounds into epochs, only compute \mathbf{W}_t at start of each epoch.

Doubling: only update on rounds $2^1, 2^2, 2^3, 2^4, \dots$

Epoch trick

Regret analysis: \mathbf{W}_t has low instantaneous per-round regret (roughly $K\mu_t$)—this also crucially relies on i.i.d. assumption.

⇒ same \mathbf{W}_t can be used for $O(t)$ more rounds!

Epoch trick: split T rounds into epochs, only compute \mathbf{W}_t at start of each epoch.

Doubling: only update on rounds $2^1, 2^2, 2^3, 2^4, \dots$

log T epochs, so $\tilde{O}(\sqrt{KT/\log N})$ AMO calls overall.

Blackwell Approachability

MEHRYAR MOHRI MOHRI@
GOOGLE RESEARCH & COURANT INSTITUTE

Motivation

- Extension of online learning to vectorial payoffs.
- Extension of zero-sum games theory to vectorial payoffs.
- Repeated games.

von Neumann's Theorem

(von Neumann, 1928)

- **Theorem** (von Neumann's minimax theorem): for any two-player zero-sum game with finite action sets,

$$\max_{p \in \Delta_1(\mathcal{A}_1)} \min_{q \in \Delta_1(\mathcal{A}_2)} E_{\substack{a_1 \sim p \\ a_2 \sim q}} [u(a_1, a_2)] = \min_{q \in \Delta_1(\mathcal{A}_2)} \max_{p \in \Delta_1(\mathcal{A}_1)} E_{\substack{a_1 \sim p \\ a_2 \sim q}} [u(a_1, a_2)].$$

- single strategy good against any strategy by other player.
- order of play does not matter.
- mixed Nash equilibria, same payoff (value of the game).

Sion's Minimax Theorem

(Sion, 1958)

- **Theorem** (simplified version): let \mathcal{X} and \mathcal{Y} be convex and compact sets and $f: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ a function that is convex with respect to its first argument and concave with respect to its second argument. Then,

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).$$

Biaffine functions

- **Definition:** let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be convex and compact sets, then biaffine is $u: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is **biaffine** if

$$u(\alpha \mathbf{x} + (1 - \alpha) \mathbf{x}', \mathbf{y}) = \alpha u(\mathbf{x}, \mathbf{y}) + (1 - \alpha) u(\mathbf{x}', \mathbf{y})$$

$$u(\mathbf{x}, \alpha \mathbf{y} + (1 - \alpha) \mathbf{y}') = \alpha u(\mathbf{x}, \mathbf{y}) + (1 - \alpha) u(\mathbf{x}, \mathbf{y}')$$

- for all $\alpha \in [0, 1]$, $\mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathbf{y}, \mathbf{y}' \in \mathcal{Y}$.

Equivalent Statement

- Equivalent formulation of Sion's theorem for biaffine function: for all $c \in \mathbb{R}$,

$$\forall \mathbf{y} \in \mathcal{Y} \exists \mathbf{x} \in \mathcal{X}: u(\mathbf{x}, \mathbf{y}) \in [c, +\infty) \implies \exists \mathbf{x} \in \mathcal{X} \forall \mathbf{y} \in \mathcal{Y}: u(\mathbf{x}, \mathbf{y}) \in [c, +\infty);$$

or $\min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} u(\mathbf{x}, \mathbf{y}) \geq c \implies \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} u(\mathbf{x}, \mathbf{y}) \geq c.$

Vectorial Payoffs

- Vector valued games: biaffine function $u: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$.
- **Question:** can the result be extended? Given convex set S ,
 $\forall \mathbf{y} \in \mathcal{Y}, \exists \mathbf{x} \in \mathcal{X}: u(\mathbf{x}, \mathbf{y}) \in S \implies \exists \mathbf{x} \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y}: u(\mathbf{x}, \mathbf{y}) \in S$.
- Counter-example:
 - clearly, $\forall \mathbf{y} \in \mathcal{Y}, \exists \mathbf{x} = \mathbf{y}, u(\mathbf{x}, \mathbf{y}) \in S$.
 - but, $\nexists \mathbf{x} \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y}: u(\mathbf{x}, \mathbf{y}) \in S$.

Approachability

(Blackwell, 1956)

- **Assumptions:** $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ convex and compact sets, biaffine vectorial payoff $u: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$, $S \subset \mathbb{R}^d$ a closed convex set.
- **Definition:** S is approachable if there exists an algorithm \mathcal{A} such that for any sequence $y_1, \dots, y_T \in \mathcal{Y}$ and $x_1, \dots, x_T \in \mathcal{X}$ with $x_t = \mathcal{A}(y_1, \dots, y_{t-1})$,

$$\lim_{T \rightarrow +\infty} d\left(\frac{1}{T} \sum_{t=1}^T u(x_t, y_t), S\right) = 0.$$

- for any $z \in \mathbb{R}^d$, $d(z, S) = \inf_{s \in S} \|z - s\|_2$.
- repeated game, adaptive player strategy.

Blackwell's Theorem

- **Theorem:** let $S \subset \mathbb{R}^d$ be a closed convex set with $S \subseteq B(0, R)$. Assume that: $\forall y \in \mathcal{Y}, \exists x \in \mathcal{X} : u(x, y) \in S$. Then, S is approachable and there exists \mathcal{A} such that

$$d\left(\frac{1}{T} \sum_{t=1}^T u(x_t, y_t), S\right) \leq \frac{2R}{\sqrt{T}}.$$

- **Proof:** let $H = \{z \in \mathbb{R}^d : w \cdot z \geq c\}$ be a halfspace that contains S . By assumption: $\min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} w \cdot u(x, y) \geq c$. By Sion's theorem, this implies:

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} w \cdot u(x, y) \geq c.$$

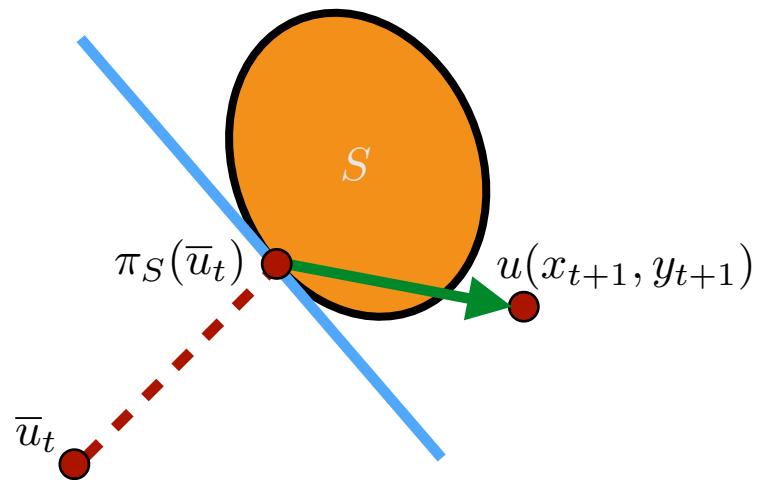
- thus,

$$\exists x^* \in \mathcal{X} : \forall y \in \mathcal{Y}, u(x^*, y) \in H.$$

Proof

- Choice of H : $H_t = \left\{ z \in \mathbb{R}^d : [z - \pi_S(\bar{u}_t)] \cdot [\pi_S(\bar{u}_t) - \bar{u}_t] \geq 0 \right\}$, where $\bar{u}_t = \frac{1}{T} \sum_{i=1}^t u(x_i, y_i)$. Define x_{t+1} to be x^* for H_t .
 - since $\bar{u}_{t+1} = \frac{t}{t+1} \bar{u}_t + \frac{1}{t+1} u(x_{t+1}, y_{t+1})$,

$$\begin{aligned}
 & d(\bar{u}_{t+1}, S)^2 \\
 & \leq d(\bar{u}_{t+1}, \pi_S(\bar{u}_t))^2 \\
 & = \|\bar{u}_{t+1} - \pi_S(\bar{u}_t)\|_2^2 \\
 & = \left\| \frac{t[\bar{u}_t - \pi_S(\bar{u}_t)]}{t+1} + \frac{[u(x_{t+1}, y_{t+1}) - \pi_S(\bar{u}_t)]}{t+1} \right\|_2^2 \\
 & \leq \frac{t^2 d(\bar{u}_t, S)^2}{(t+1)^2} + \frac{\|u(x_{t+1}, y_{t+1}) - \pi_S(\bar{u}_t)\|^2}{(t+1)^2}.
 \end{aligned}$$



Proof

■ This gives:

$$(t+1)^2 d(\bar{u}_{t+1}, S)^2 \leq t^2 d(\bar{u}_t, S)^2 + \|u(x_{t+1}, y_{t+1}) - \pi_S(\bar{u}_t)\|^2.$$

- assuming that $\|u(x_{t+1}, y_{t+1})\| \leq R$, this implies

$$(t+1)^2 d(\bar{u}_{t+1}, S)^2 \leq t^2 d(\bar{u}_t, S)^2 + 4R^2;$$

- Thus, $T^2 d(\bar{u}_T, S)^2 \leq 4TR^2$ and

$$d(\bar{u}_T, S) \leq \frac{2R}{\sqrt{T}}.$$

External Regret

- K actions, $\mathcal{X} = \Delta_K$, losses bounded by one.
- Regret per round of algorithm \mathcal{A} :

$$\frac{\text{Reg}_T(\mathcal{A})}{T} = \max_{a \in [K]} \frac{1}{T} \sum_{t=1}^T [\mathbf{p}_t \cdot \ell_t - \ell_t(a)].$$

- Vectorial payoff:

$$u(\mathbf{p}_t, \ell_t) = \begin{bmatrix} \mathbf{p}_t \cdot \ell_t - \ell_t(a_1) \\ \vdots \\ \mathbf{p}_t \cdot \ell_t - \ell_t(a_K) \end{bmatrix}.$$

External regret can be understood as a special case of Blackwell's approachability

- Approachable set: $S = [-1, 0]^K$.
- Assumption holds: for any $\ell \in [0, 1]^K$,

$$u(\mathbf{p}_t, \ell_t) \in S \text{ for } \mathbf{p}_t = \mathbf{e}_k, \text{ with } k \in \operatorname{argmin}_{k \in [K]} \ell(a_k).$$

- Regret bound: $O\left(\sqrt{\frac{K}{T}}\right)$ (suboptimal).

Approachability Using OCO

(Abernethy et al., 2011; Dann et al., 2023)

■ Fenchel duality:

$$\begin{aligned} d(x, S) &= \min_{s \in S} \|x - s\|_2 \\ &= \min_{s \in \mathbb{R}^d} \|x - s\|_2 + I_S(s) && (\text{def. of } I_S) \\ &= \max_{\theta \in \mathbb{R}^d} -\left\{ I_{B_2(0,1)}(\theta) + \theta \cdot x \right\} - \sup_{s \in S} \{-\theta \cdot s\} && (\text{Fenchel duality theorem}) \\ &= \max_{\theta \in B_2(0,1)} -\theta \cdot x - \sup_{s \in S} \{-\theta \cdot s\} && (\text{def. of } I_{B_2(0,1)}) \\ &= \max_{\theta \in B_2(0,1)} \left\{ \theta \cdot x - \underbrace{\sup_{s \in S} \theta \cdot s}_{I_S^*(\theta)} \right\}. \end{aligned}$$

OCO-Based Algorithm

OCO2APP(\mathcal{A}, S)

- 1 $\theta_1 \leftarrow \text{RANDOM}(\mathbb{B}(0, 1))$
- 2 **for** $t \leftarrow 1$ **to** T **do**
- 3 $p_t \leftarrow p \in \Delta_K : \forall y \in \mathcal{Y}, \theta_t \cdot u(p, y) \leq \max_{s \in S} \theta_t \cdot s \triangleright \text{exists by sep. assumption}$
- 4 $y_t \leftarrow \text{RECEIVE}(\mathcal{Y})$
- 5 $f_t \leftarrow \theta \mapsto \max_{s \in S} \theta \cdot s - \theta \cdot u(p_t, y_t)$
- 6 $\theta_{t+1} \leftarrow \mathcal{A}(f_1^t, \theta_1^t)$

■ Guarantee:

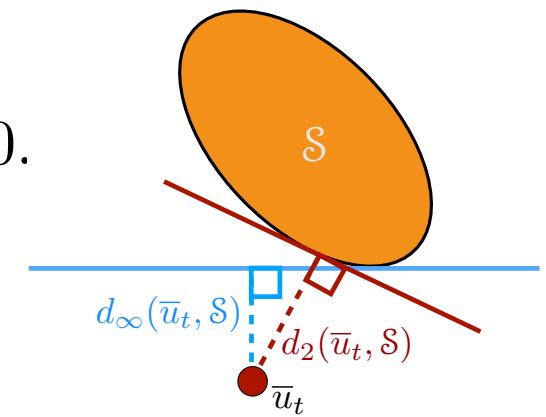
$$\begin{aligned} d(\bar{u}_T, S) &= \max_{\theta \in \mathbb{B}_2(0,1)} \left\{ \theta \cdot \bar{u}_T - \max_{s \in S} \theta \cdot s \right\} \\ &= \max_{\theta \in \mathbb{B}_2(0,1)} \left\{ -\frac{1}{T} \sum_{t=1}^T f_t(\theta) \right\} = -\min_{\theta \in \mathbb{B}_2(0,1)} \left\{ \frac{1}{T} \sum_{t=1}^T f_t(\theta) \right\} \\ &= \frac{\text{Reg}_T(\mathcal{A})}{T} - \frac{1}{T} \sum_{t=1}^T \underbrace{f_t(\theta_t)}_{\geq 0} \leq \frac{\text{Reg}_T(\mathcal{A})}{T}. \end{aligned}$$

ℓ_∞ -Approachability

(e.g., Dann et al., 2023)

- **Assumptions:** $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ convex and compact sets, biaffine vectorial payoff $u: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$, $S \subset \mathbb{R}^d$ a closed convex set.
- **Definition:** S is ℓ_∞ -approachable if there exists an algorithm \mathcal{A} such that for any sequence $y_1, \dots, y_T \in \mathcal{Y}$ and $x_1, \dots, x_T \in \mathcal{X}$ with $x_t = \mathcal{A}(y_1, \dots, y_{t-1})$,

$$\lim_{t \rightarrow +\infty} d_\infty \left(\frac{1}{T} \sum_{t=1}^T u(x_t, y_t), S \right) = 0.$$



- for any $z \in \mathbb{R}^d$, $d(z, S) = \inf_{s \in S} \|z - s\|_\infty$.
- repeated game, adaptive player strategy.

ℓ_∞ -Approachability Using OCO

(Abernethy et al., 2011; Dann et al., 2023)

■ Fenchel duality:

$$\begin{aligned} d_\infty(x, S) &= \min_{s \in S} \|x - s\|_\infty \\ &= \min_{s \in \mathbb{R}^d} \|x - s\|_\infty + I_S(s) && \text{(def. of } I_S\text{)} \\ &= \max_{\theta \in \mathbb{R}^d} -\left\{ I_{B_1(0,1)}(\theta) + \theta \cdot x \right\} - \sup_{s \in S} \{-\theta \cdot s\} && \text{(Fenchel duality theorem)} \\ &= \max_{\theta \in B_1(0,1)} -\theta \cdot x - \sup_{s \in S} \{-\theta \cdot s\} && \text{(def. of } I_{B_1(0,1)}\text{)} \\ &= \max_{\theta \in B_1(0,1)} \left\{ \theta \cdot x - \sup_{s \in S} \theta \cdot s \right\}. \end{aligned}$$

ℓ_∞ -OCO-Based Algorithm

OCO2APP(\mathcal{A}, S)

- 1 $\theta_1 \leftarrow \text{RANDOM}(\mathbb{B}(0, 1))$
- 2 **for** $t \leftarrow 1$ **to** T **do**
- 3 $p_t \leftarrow p \in \Delta_K : \forall y \in \mathcal{Y}, \theta_t \cdot u(p, y) \leq \max_{s \in S} \theta_t \cdot s \triangleright \text{exists by sep. assumption}$
- 4 $y_t \leftarrow \text{RECEIVE}(\mathcal{Y})$
- 5 $f_t \leftarrow \theta \mapsto \max_{s \in S} \theta \cdot s - \theta \cdot u(p_t, y_t)$
- 6 $\theta_{t+1} \leftarrow \mathcal{A}(f_1^t, \theta_1^t)$

■ Guarantee:

$$\begin{aligned} d(\bar{u}_T, S) &= \max_{\theta \in \mathbb{B}_1(0,1)} \left\{ \theta \cdot \bar{u}_T - \max_{s \in S} \theta \cdot s \right\} \\ &= \max_{\theta \in \mathbb{B}_1(0,1)} \left\{ -\frac{1}{T} \sum_{t=1}^T f_t(\theta) \right\} = -\min_{\theta \in \mathbb{B}_1(0,1)} \left\{ \frac{1}{T} \sum_{t=1}^T f_t(\theta) \right\} \\ &= \frac{\text{Reg}_T(\mathcal{A})}{T} - \frac{1}{T} \sum_{t=1}^T \underbrace{f_t(\theta_t)}_{\geq 0} \leq \frac{\text{Reg}_T(\mathcal{A})}{T}. \end{aligned}$$

External Regret

- K actions, $\mathcal{X} = \Delta_K$, losses bounded by one.
- Vectorial payoff and regret per round of algorithm \mathcal{A} :

$$u(\mathbf{p}_t, \ell_t) = \begin{bmatrix} \mathbf{p}_t \cdot \ell_t - \ell_t(a_1) \\ \vdots \\ \mathbf{p}_t \cdot \ell_t - \ell_t(a_K) \end{bmatrix} \quad \left| \frac{\text{Reg}_T(\mathcal{A})}{T} \right| = \left\| \frac{1}{T} \sum_{t=1}^T u(\mathbf{p}_t, y_t) \right\|_\infty.$$

- Approachable set: $S = [-1, 0]^K$.
- Choice of \mathbf{p}_t : $\mathbf{p}_t = \theta_t \in \Delta_K$.
- Loss function: $f_t(\theta) = -\theta \cdot u(\mathbf{p}_t, \ell_t) = -\theta \cdot u(\theta_t, \ell_t)$.
- Algorithm \mathcal{A} : use EG algorithm.
- Regret bound: $2\sqrt{T \log K}$.

References

- ([Blackwell, 1956](#)): paper introducing approachability with constructive proof.
- ([Abernethy et al., 2011](#)): reduction of approachability to online linear optimization.
- ([Foster and Vohra, 1999](#)): internal regret and relationship with approachability; see also references therein.
- ([Hart and Mas-Colell, 2000](#)): internal regret as approachability, with explicit strategies.
- ([Hart and Mas-Colell, 2001](#)): class of approachability algorithms, related to projection in some norm.

References

- [\(Shimkin, 2016\)](#): approachability for an arbitrary norm, general duality result using Sion's minimax theorem.
- [\(Kwon, 2021\)](#): duality theorem similar to (Shimkin, 2016), FTRL algorithm for general norm.
- [\(Perchet, 2015\)](#): ℓ_∞ -approachability, exponential weight algorithm + several other interesting approachability publications from this author.
- [\(Dann, Mansour, MM, Schneider, and Sivan, 2023\)](#): emphasizes importance of ℓ_∞ -approachability, general conversion to lower-dim space of vectorial payoffs, general Fenchel duality result and pseudonorm approachability.

Advanced Machine Learning

Learning Kernels

MEHRYAR MOHRI

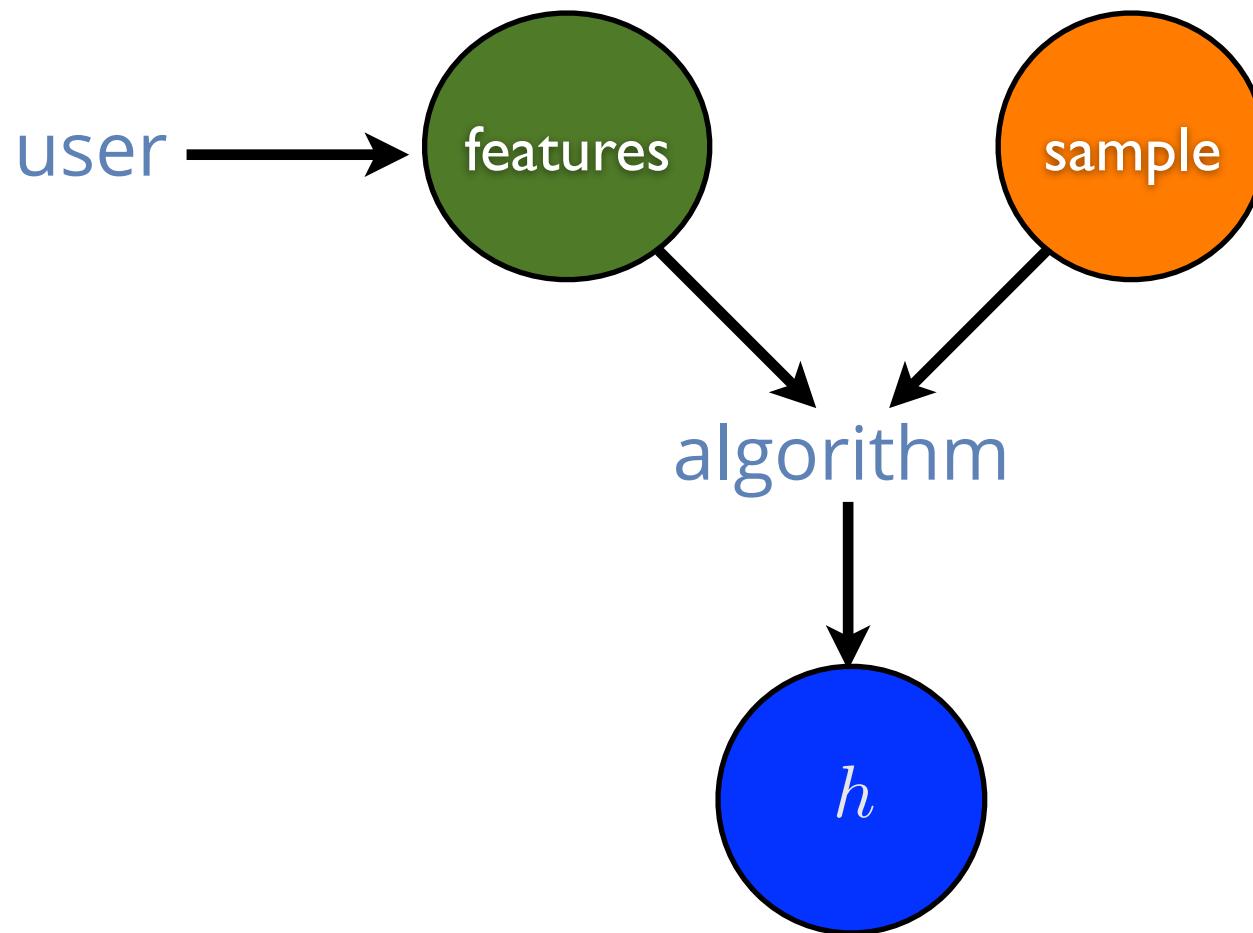
MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

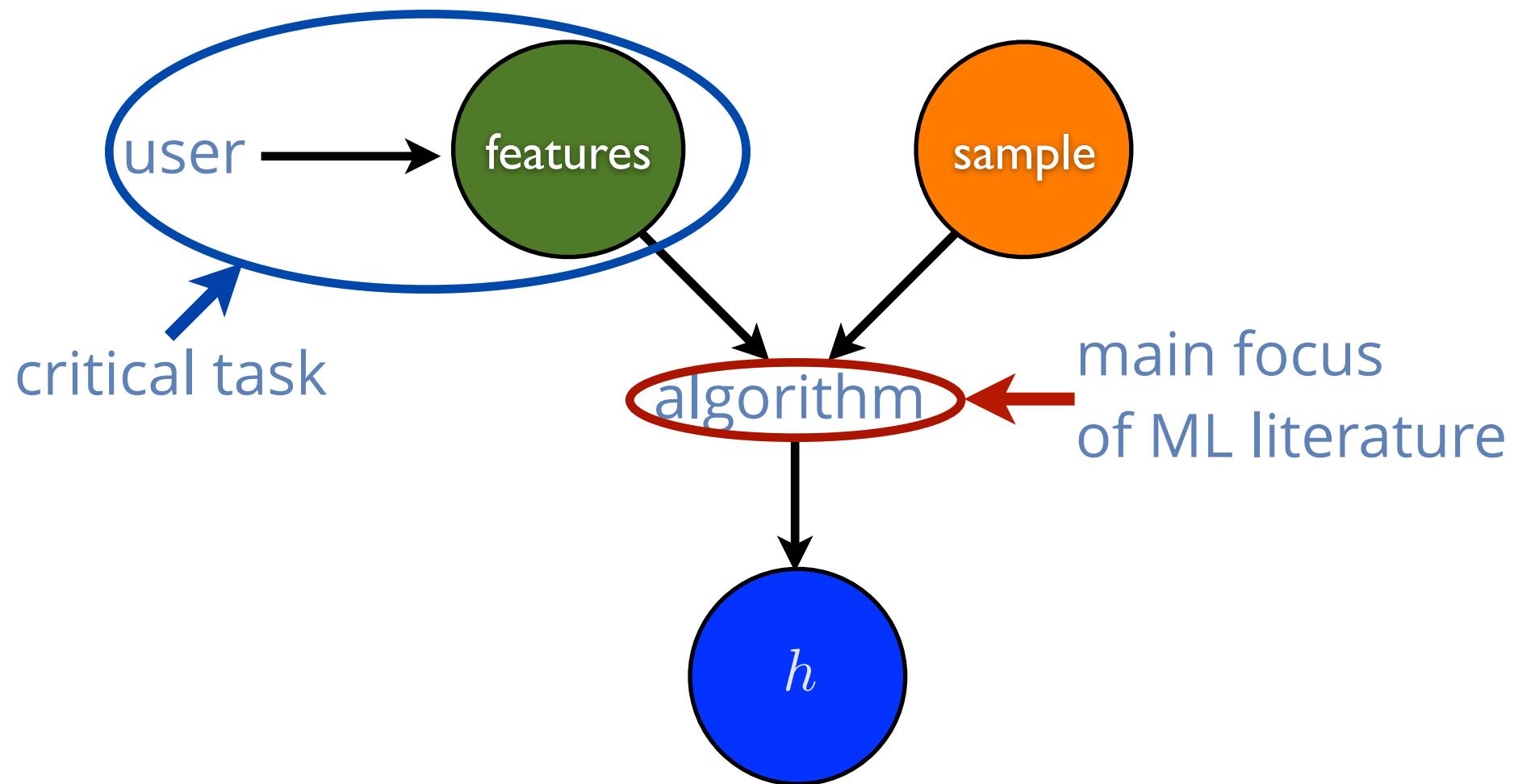
Outline

- Kernel methods.
- Learning kernels
 - scenario.
 - learning bounds.
 - algorithms.

Machine Learning Components



Machine Learning Components



Kernel Methods

- Features $\Phi: X \rightarrow \mathbb{H}$ implicitly defined via the choice of a PDS kernel K

$$\forall x, y \in X, \quad \Phi(x) \cdot \Phi(y) = K(x, y).$$

- K interpreted as a similarity measure.
- Flexibility: PDS kernel can be chosen arbitrarily.
- Help extend a variety of algorithms to non-linear predictors, e.g., SVMs, KRR, SVR, KPCA.
- PDS condition directly related to convexity of optimization problem.

Example - Polynomial Kernels

- **Definition:**

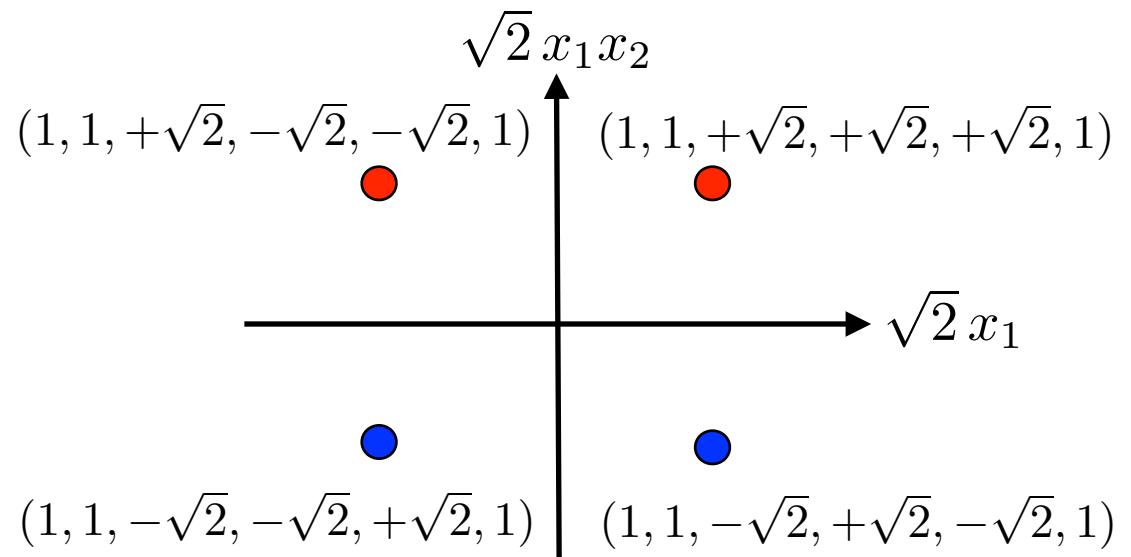
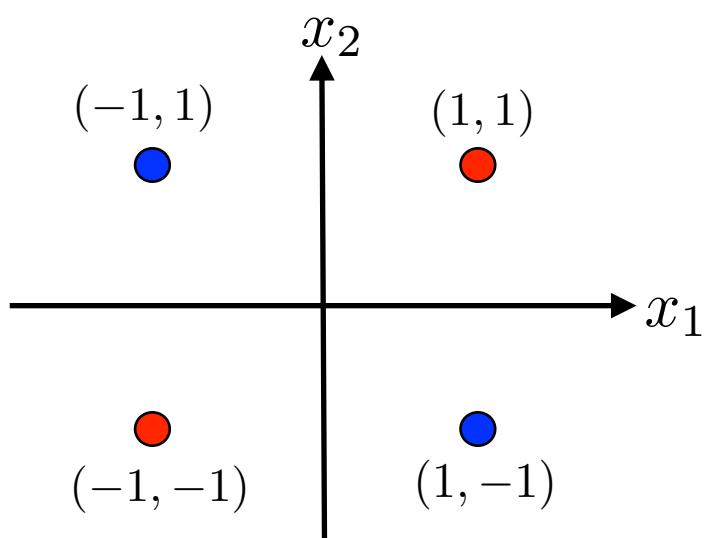
$$\forall x, y \in \mathbb{R}^N, K(x, y) = (x \cdot y + c)^d, \quad c > 0.$$

- **Example:** for $N=2$ and $d=2$,

$$K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2$$
$$= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2c} x_1 \\ \sqrt{2c} x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2} y_1 y_2 \\ \sqrt{2c} y_1 \\ \sqrt{2c} y_2 \\ c \end{bmatrix}.$$

XOR Problem

- Use second-degree polynomial kernel with $c = 1$:



Linearly non-separable

Linearly separable by $x_1x_2 = 0$.

Other Standard PDS Kernels

■ Gaussian kernels:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \quad \sigma \neq 0.$$

- Normalized kernel of $(\mathbf{x}, \mathbf{x}') \mapsto \exp\left(\frac{\mathbf{x} \cdot \mathbf{x}'}{\sigma^2}\right)$.

■ Sigmoid Kernels:

$$K(x, y) = \tanh(a(x \cdot y) + b), \quad a, b \geq 0.$$

SVM

(Cortes and Vapnik, 1995; Boser, Guyon, and Vapnik, 1992)

■ Primal:

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \left(1 - y_i (\mathbf{w} \cdot \Phi_K(x_i) + b) \right)_+$$

■ Dual:

$$\max_{\alpha} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to: $0 \leq \alpha_i \leq C \wedge \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$.

Kernel Ridge Regression

(Hoerl and Kennard, 1970; Sanders et al., 1998)

■ Primal:

$$\min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^m (\mathbf{w} \cdot \Phi_K(x_i) + b - y_i)^2.$$

■ Dual:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} -\boldsymbol{\alpha}^\top (\mathbf{K} + \lambda \mathbf{I}) \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{y}.$$

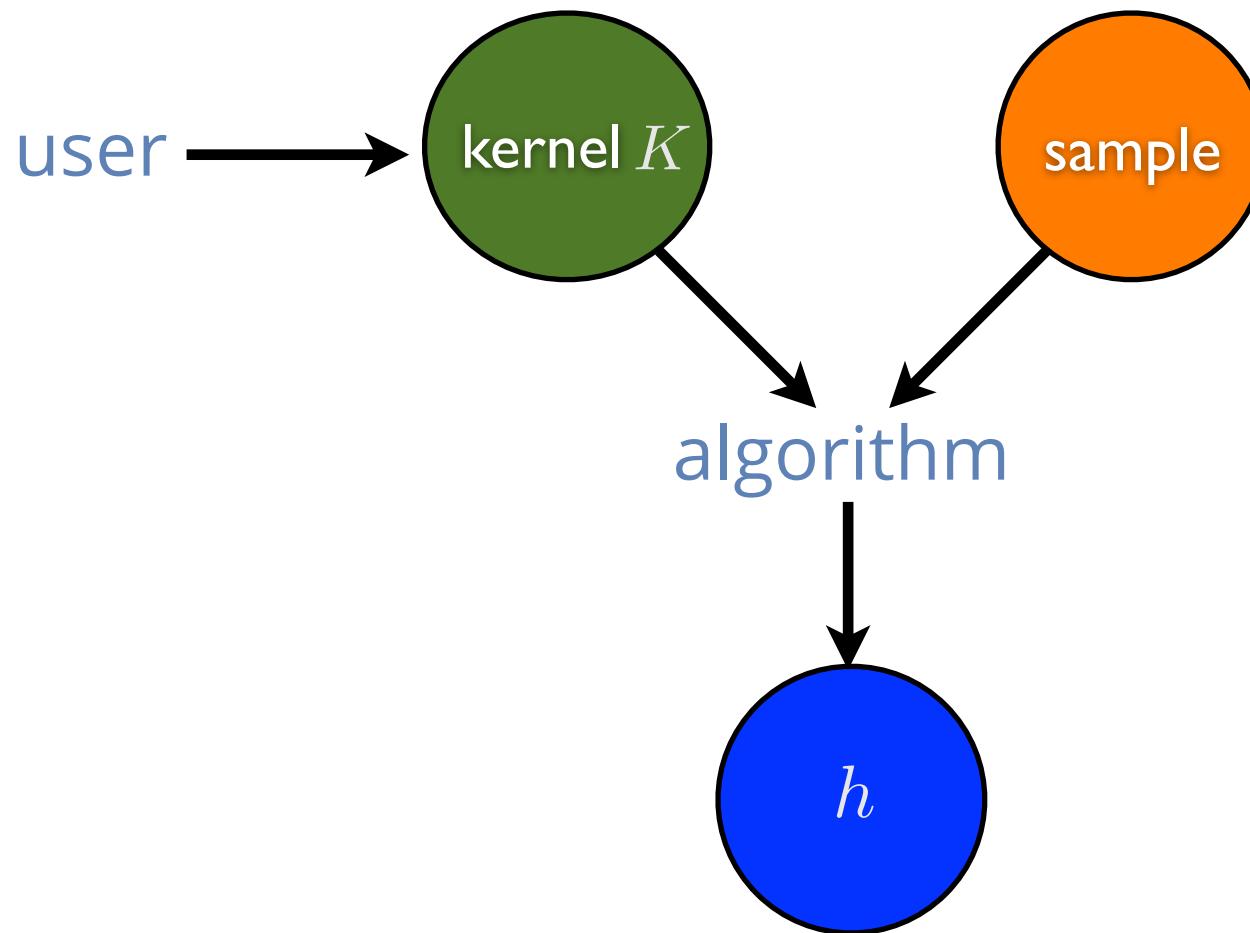
Questions

- How should the user choose the kernel?
 - problem similar to that of selecting features for other learning algorithms.
 - poor choice → learning made very difficult.
 - good choice → even poor learners could succeed.
- The requirement from the user is thus critical.
 - can this requirement be lessened?
 - is a more automatic selection of features possible?

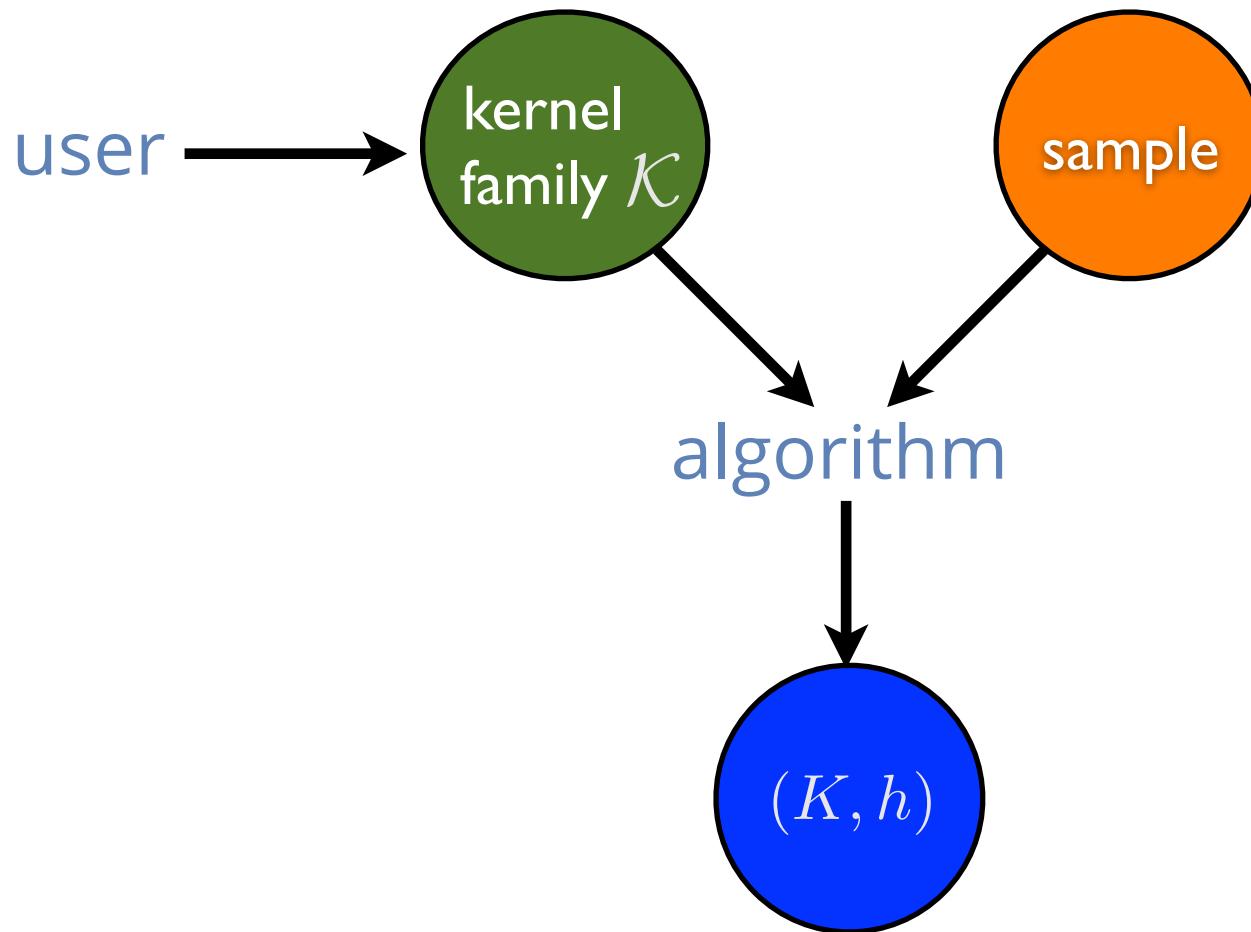
Outline

- Kernel methods.
- Learning kernels
 - scenario.
 - learning bounds.
 - algorithms.

Standard Learning with Kernels



Learning Kernel Framework



Kernel Families

- Most frequently used kernel families, $q \geq 1$,

$$\mathcal{K}_q = \left\{ K_{\boldsymbol{\mu}} : K_{\boldsymbol{\mu}} = \sum_{k=1}^p \mu_k K_k, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \in \Delta_q \right\}$$

with $\Delta_q = \left\{ \boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \|\boldsymbol{\mu}\|_q = 1 \right\}$.

- Hypothesis sets:

$$H_q = \left\{ h \in \mathbb{H}_K : K \in \mathcal{K}_q, \|h\|_{\mathbb{H}_K} \leq 1 \right\}.$$

Relation between Norms

- **Lemma:** for $p, q \in (0, +\infty]$, the following holds:

$$\forall \mathbf{x} \in \mathbb{R}^N, p \leq q \Rightarrow \|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p \leq N^{\frac{1}{p} - \frac{1}{q}} \|\mathbf{x}\|_q.$$

- **Proof:** for the left inequalities, observe that for $\mathbf{x} \neq 0$,

$$\left[\frac{\|\mathbf{x}\|_p}{\|\mathbf{x}\|_q} \right]^p = \sum_{i=1}^N \underbrace{\left[\frac{|x_i|}{\|\mathbf{x}\|_q} \right]^p}_{\leq 1} \geq \sum_{i=1}^N \left[\frac{|x_i|}{\|\mathbf{x}\|_q} \right]^q = 1.$$

- Right inequalities follow immediately Hölder's inequality:

$$\|\mathbf{x}\|_p = \left[\sum_{i=1}^N |x_i|^p \right]^{\frac{1}{p}} \leq \left[\left(\sum_{i=1}^N (|x_i|^p)^{\frac{q}{p}} \right)^{\frac{p}{q}} \left(\sum_{i=1}^N (1)^{\frac{q}{q-p}} \right)^{1-\frac{p}{q}} \right]^{\frac{1}{p}} = \|\mathbf{x}\|_q N^{\frac{1}{p} - \frac{1}{q}}.$$

Single Kernel Guarantee

(Koltchinskii and Panchenko, 2002)

- **Theorem:** fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H_1$,

$$R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \frac{\sqrt{\text{Tr}[\mathbf{K}]}}{m} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Multiple Kernel Guarantee

(Cortes, MM, and Rostamizadeh, 2010)

- **Theorem:** fix $\rho > 0$. Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H_q$ and any integer $1 \leq s \leq r$:

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2}{\rho} \frac{\sqrt{s \|\mathbf{u}\|_s}}{m} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

with $\mathbf{u} = (\text{Tr}[\mathbf{K}_1], \dots, \text{Tr}[\mathbf{K}_p])^\top$.

Proof

■ Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$.

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H_q) &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in H_q} \sum_{i=1}^m \sigma_i h(x_i) \right] \\
&= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q, \boldsymbol{\alpha}^\top \mathbf{K}_\mu \boldsymbol{\alpha} \leq 1} \sum_{i,j=1}^m \sigma_i \alpha_j K_\mu(x_i, x_j) \right] \\
&= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q, \boldsymbol{\alpha}^\top \mathbf{K}_\mu \boldsymbol{\alpha} \leq 1} \boldsymbol{\sigma}^\top \mathbf{K}_\mu \boldsymbol{\alpha} \right] = \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q, \|\boldsymbol{\alpha}\|_{\mathbf{K}_\mu^{1/2}} \leq 1} \langle \boldsymbol{\sigma}, \boldsymbol{\alpha} \rangle_{\mathbf{K}_\mu^{1/2}} \right] \\
&= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q} \sqrt{\boldsymbol{\sigma}^\top \mathbf{K}_\mu \boldsymbol{\sigma}} \right] \quad (\text{Cauchy-Schwarz}) \\
&= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \Delta_q} \sqrt{\boldsymbol{\mu} \cdot \mathbf{u}_\sigma} \right] \quad [\mathbf{u}_\sigma = (\boldsymbol{\sigma}^\top \mathbf{K}_1 \boldsymbol{\sigma}, \dots, \boldsymbol{\sigma}^\top \mathbf{K}_p \boldsymbol{\sigma})^\top] \\
&= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sqrt{\|\mathbf{u}_\sigma\|_r} \right]. \quad (\text{definition of dual norm})
\end{aligned}$$

Lemma

(Cortes, MM, and Rostamizadeh, 2010)

- **Lemma:** Let \mathbf{K} be a kernel matrix for a finite sample. Then, for any integer r ,

$$\underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[(\boldsymbol{\sigma}^\top \mathbf{K} \boldsymbol{\sigma})^r \right] \leq \left(r \operatorname{Tr}[\mathbf{K}] \right)^r.$$

- **Proof:** combinatorial argument.

Proof

- For any $1 \leq s \leq r$,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(H_q) &= \frac{1}{m} \underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_r} \right] \\ &\leq \frac{1}{m} \underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[\sqrt{\|\mathbf{u}_{\boldsymbol{\sigma}}\|_s} \right] \\ &= \frac{1}{m} \underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[\left[\sum_{k=1}^p (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^s \right]^{\frac{1}{2s}} \right] \\ &\leq \frac{1}{m} \left[\underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[\sum_{k=1}^p (\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^s \right] \right]^{\frac{1}{2s}} \text{ (Jensen's inequality)} \\ &= \frac{1}{m} \left[\sum_{k=1}^p \underset{\boldsymbol{\sigma}}{\mathbb{E}} \left[(\boldsymbol{\sigma}^\top \mathbf{K}_k \boldsymbol{\sigma})^s \right] \right]^{\frac{1}{2s}} \\ &\leq \frac{1}{m} \left[\sum_{k=1}^p \left(s \operatorname{Tr}[\mathbf{K}_k] \right)^s \right]^{\frac{1}{2s}} = \frac{\sqrt{s \|\mathbf{u}\|_s}}{m}. \quad (\text{lemma})\end{aligned}$$

L₁ Learning Bound

(Cortes, MM, and Rostamizadeh, 2010)

- **Corollary:** fix $\rho > 0$. For any $\delta > 0$, with probability $1 - \delta$, the following holds for all $h \in H_1$:

$$R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \sqrt{\frac{e \lceil \log p \rceil \max_{k=1}^p \text{Tr}[\mathbf{K}_k]}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

- weak dependency on p .
- bound valid for $p \gg m$.
- $\text{Tr}[\mathbf{K}_k] \leq m \max_x K_k(x, x)$.

Proof

- For $q = 1$, the bound holds for any integer $s \geq 1$

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2}{\rho} \frac{\sqrt{s\|\mathbf{u}\|_s}}{m} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

$$\text{with } s\|\mathbf{u}\|_s = s \left[\sum_{k=1}^p \text{Tr}[\mathbf{K}_k]^s \right]^{\frac{1}{s}} \leq sp^{\frac{1}{s}} \max_{k=1}^p \text{Tr}[\mathbf{K}_k].$$

- The function $s \mapsto sp^{\frac{1}{s}}$ reaches its minimum at $\log p$.

Lower Bound

■ Tight bound:

- dependency $\sqrt{\log p}$ cannot be improved.
- argument based on VC dimension or example.

■ Observations: case $\mathcal{X}=\{-1, +1\}^p$.

- canonical projection kernels $K_k(\mathbf{x}, \mathbf{x}') = x_k x'_k$.
- H_1 contains $J_p = \{\mathbf{x} \mapsto s x_k : k \in [1, p], s \in \{-1, +1\}\}$.
- $\text{VCdim}(J_p) = \Omega(\log p)$.
- for $\rho=1$ and $h \in J_p$, $\widehat{R}_\rho(h) = \widehat{R}(h)$.
- VC lower bound: $\Omega(\sqrt{\text{VCdim}(J^p)/m})$.

Pseudo-Dimension Bound

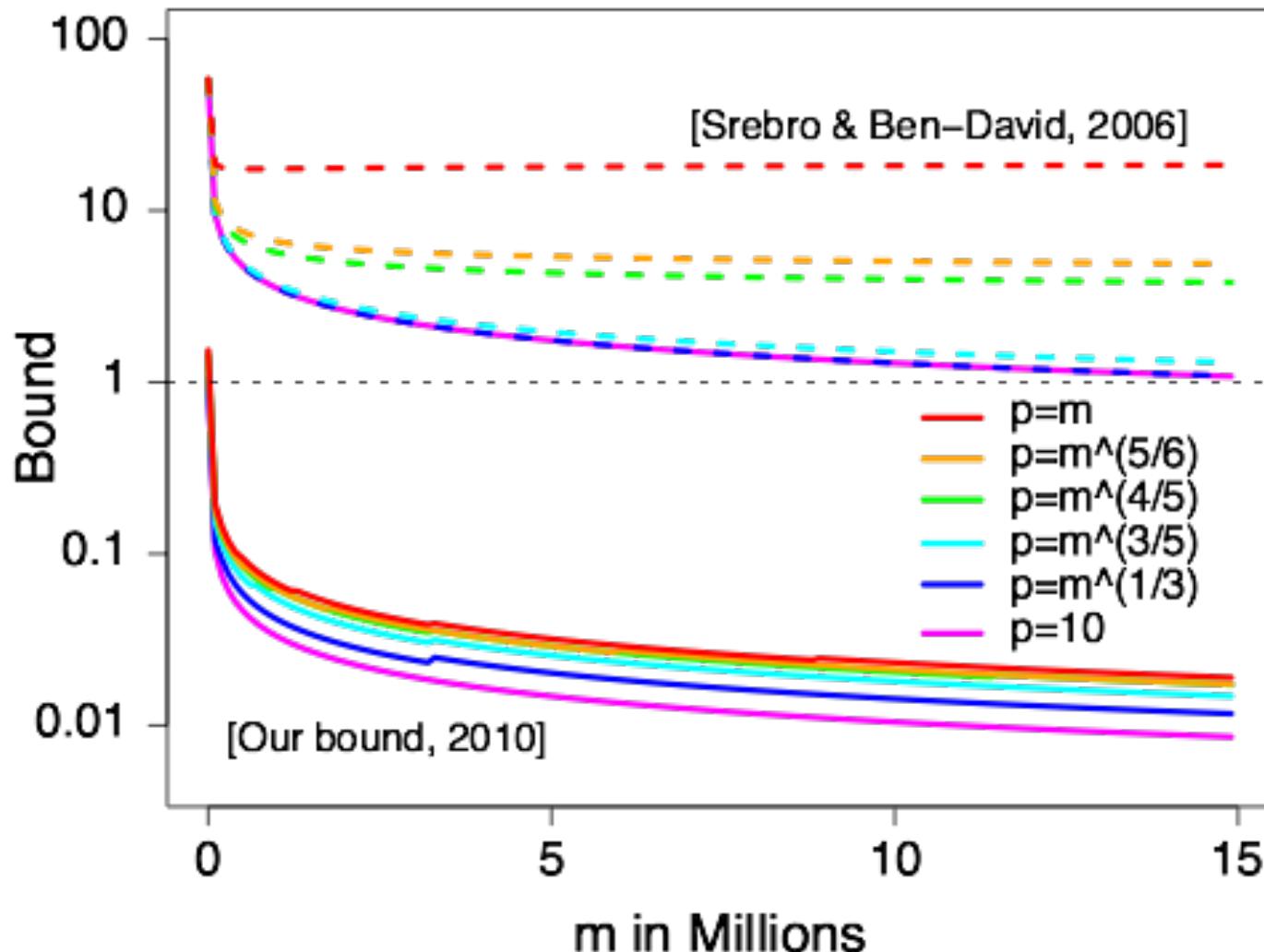
(Srebro and Ben-David, 2006)

- Assume that for all $k \in [1, p]$, $K_k(x, x) \leq R^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H_1$,

$$R(h) \leq \widehat{R}_\rho(h) + \sqrt{8 \frac{2 + p \log \frac{128em^3R^2}{\rho^2p} + 256 \frac{R^2}{\rho^2} \log \frac{\rho em}{8R} \log \frac{128mR^2}{\rho^2} + \log(1/\delta)}{m}}.$$

- bound additive in p (modulo log terms).
- not informative for $p > m$.
- based on pseudo-dimension of kernel family.
- similar guarantees for other families.

Comparison



$$\rho/R = .2$$

L_q Learning Bound

(Cortes, MM, and Rostamizadeh, 2010)

- **Corollary:** fix $\rho > 0$. Let $q, r \geq 1$ with $\frac{1}{q} + \frac{1}{r} = 1$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H_q$:

$$R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \sqrt{\frac{rp^{\frac{1}{r}} \max_{k=1}^p \text{Tr}[\mathbf{K}_k]}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

- mild dependency on p .
- $\text{Tr}[\mathbf{K}_k] \leq m \max_x K_k(x, x)$.

Lower Bound

■ Tight bound:

- dependency $p^{\frac{1}{2r}}$ cannot be improved.
- in particular $p^{\frac{1}{4}}$ tight for L_2 regularization.

■ Observations: equal kernels.

- $\sum_{k=1}^p \mu_k K_k = \left(\sum_{k=1}^p \mu_k \right) K_1$.
- thus, $\|h\|_{\mathbb{H}_{K_1}}^2 = \left(\sum_{k=1}^p \mu_k \right) \|h\|_{\mathbb{H}_K}^2$ for $\sum_{k=1}^p \mu_k \neq 0$.
- $\sum_{k=1}^p \mu_k \leq p^{\frac{1}{r}} \|\boldsymbol{\mu}\|_q = p^{\frac{1}{r}}$ (Hölder's inequality).
- H_q coincides with $\{h \in \mathbb{H}_{K_1} : \|h\|_{\mathbb{H}_{K_1}} \leq p^{\frac{1}{2r}}\}$.

Outline

- Kernel methods.
- Learning kernels
 - scenario.
 - learning bounds.
 - algorithms.

General LK Formulation - SVMs

■ Notation:

- \mathcal{K} set of PDS kernel functions.
- $\bar{\mathcal{K}}$ kernel matrices associated to \mathcal{K} , assumed convex.
- $\mathbf{Y} \in \mathbb{R}^{m \times m}$ diagonal matrix with $\mathbf{Y}_{ii} = \mathbf{y}_i$.

■ Optimization problem:

$$\min_{\mathbf{K} \in \bar{\mathcal{K}}} \max_{\boldsymbol{\alpha}} 2 \boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K} \mathbf{Y} \boldsymbol{\alpha}$$

subject to: $\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0$.

- convex problem: function linear in \mathbf{K} , convexity of pointwise maximum.

Parameterized LK Formulation

■ Notation:

- $(K_\mu)_{\mu \in \Delta}$ parameterized set of PDS kernel functions.
- Δ convex set, $\mu \mapsto \mathbf{K}_\mu$ concave function.
- $\mathbf{Y} \in \mathbb{R}^{m \times m}$ diagonal matrix with $\mathbf{Y}_{ii} = \mathbf{y}_i$.

■ Optimization problem:

$$\min_{\mu \in \Delta} \max_{\alpha} 2\alpha^\top \mathbf{1} - \alpha^\top \mathbf{Y}^\top \mathbf{K}_\mu \mathbf{Y} \alpha$$

subject to: $\mathbf{0} \leq \alpha \leq \mathbf{C} \wedge \alpha^\top \mathbf{y} = 0$.

- convex problem: function convex in μ , convexity of pointwise maximum.

Non-Negative Combinations

- $K_\mu = \sum_{k=1}^p \mu_k K_k, \mu \in \Delta_1.$
- By von Neumann's generalized minimax theorem (convexity wrt μ , concavity wrt α , Δ_1 convex and compact, \mathcal{A} convex and compact):

$$\begin{aligned}& \min_{\mu \in \Delta_1} \max_{\alpha \in \mathcal{A}} 2\alpha^\top \mathbf{1} - \alpha^\top \mathbf{Y}^\top \mathbf{K}_\mu \mathbf{Y} \alpha \\&= \max_{\alpha \in \mathcal{A}} \min_{\mu \in \Delta_1} 2\alpha^\top \mathbf{1} - \alpha^\top \mathbf{Y}^\top \mathbf{K}_\mu \mathbf{Y} \alpha \\&= \max_{\alpha \in \mathcal{A}} 2\alpha^\top \mathbf{1} - \max_{\mu \in \Delta_1} \alpha^\top \mathbf{Y}^\top \mathbf{K}_\mu \mathbf{Y} \alpha \\&= \max_{\alpha \in \mathcal{A}} 2\alpha^\top \mathbf{1} - \max_{k \in [1, p]} \alpha^\top \mathbf{Y}^\top \mathbf{K}_k \mathbf{Y} \alpha.\end{aligned}$$

Non-Negative Combinations

(Lanckriet et al., 2004)

- Optimization problem: in view of the previous analysis, the problem can be rewritten as the following QCQP.

$$\max_{\alpha, t} 2\alpha^\top \mathbf{1} - t$$

subject to: $\forall k \in [1, p], t \geq \alpha^\top \mathbf{Y}^\top \mathbf{K}_k \mathbf{Y} \alpha;$

$$\mathbf{0} \leq \alpha \leq \mathbf{C} \wedge \alpha^\top \mathbf{y} = 0.$$

- complexity (interior-point methods): $O(pm^3)$.

Equivalent Primal Formulation

- Optimization problem:

$$\min_{w, \mu \in \Delta_q} \frac{1}{2} \sum_{k=1}^p \frac{\|\mathbf{w}_k\|_2^2}{\mu_k} + C \sum_{i=1}^m \max \left\{ 0, 1 - y_i \left(\sum_{k=1}^p \mathbf{w}_k \cdot \Phi_k(x_i) \right) \right\}.$$

Lots of Optimization Solutions

- QCQP (Lanckriet et al., 2004).
- Wrapper methods — interleaving call to SVM solver and update of μ :
 - SILP (Sonnenburg et al., 2006).
 - Reduced gradient (SimpleML) (Rakotomamonjy et al., 2008).
 - Newton's method (Kloft et al., 2009).
 - Mirror descent (Nath et al., 2009).
- On-line method (Orabona & Jie, 2011).
- Many other methods proposed.

Does It Work?

■ Experiments:

- this algorithm and its different optimization solutions often do not significantly outperform the simple uniform combination kernel in practice!
- observations corroborated by NIPS workshops.

■ Alternative algorithms: significant improvement (see empirical results of (Gönen and Alpaydin, 2011)).

- **centered alignment-based LK algorithms** (Cortes, MM, and Rostamizadeh, 2010 and 2012).
- **non-linear combination of kernels** (Cortes, MM, and Rostamizadeh, 2009).

LK Formulation - KRR

(Cortes, MM, and Rostamizadeh, 2009)

Kernel family:

- non-negative combinations.
- L_q regularization.

Optimization problem:

$$\min_{\mu} \max_{\alpha} -\lambda \alpha^\top \alpha - \sum_{k=1}^p \mu_k \alpha^\top \mathbf{K}_k \alpha + 2\alpha^\top \mathbf{y}$$

subject to: $\mu \geq 0 \wedge \|\mu - \mu_0\|_q \leq \Lambda$.

- convex optimization: linearity in μ and convexity of pointwise maximum.

Projected Gradient

- Solving maximization problem in α , closed-form solution $\alpha = (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{y}$, reduces problem to

$$\min_{\mu} \mathbf{y}^\top (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{y}$$

subject to: $\mu \geq 0 \wedge \|\mu - \mu_0\|_2 \leq \Lambda$.

- Convex optimization problem, one solution using projection-based gradient descent:

$$\begin{aligned}\frac{\partial F}{\partial \mu_k} &= \text{Tr} \left[\frac{\partial \mathbf{y}^\top (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{y}}{\partial (\mathbf{K}_\mu + \lambda \mathbf{I})} \frac{\partial (\mathbf{K}_\mu + \lambda \mathbf{I})}{\partial \mu_k} \right] \\ &= - \text{Tr} \left[(\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{y} \mathbf{y}^\top (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \frac{\partial (\mathbf{K}_\mu + \lambda \mathbf{I})}{\partial \mu_k} \right] \\ &= - \text{Tr} \left[(\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{y} \mathbf{y}^\top (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{K}_k \right] \\ &= - \mathbf{y}^\top (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{K}_k (\mathbf{K}_\mu + \lambda \mathbf{I})^{-1} \mathbf{y} = -\alpha^\top \mathbf{K}_k \alpha.\end{aligned}$$

□

Proj. Grad. KRR - L₂ Reg.

PROJECTIONBASEDGRADIENTDESCENT($((\mathbf{K}_k)_{k \in [1, p]}, \boldsymbol{\mu}_0)$)

```
1    $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_0$ 
2    $\boldsymbol{\mu}' \leftarrow \infty$ 
3   while  $\|\boldsymbol{\mu}' - \boldsymbol{\mu}\| > \epsilon$  do
4        $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}'$ 
5        $\boldsymbol{\alpha} \leftarrow (\mathbf{K}_{\boldsymbol{\mu}} + \lambda \mathbf{I})^{-1} \mathbf{y}$ 
6        $\boldsymbol{\mu}' \leftarrow \boldsymbol{\mu} + \eta (\boldsymbol{\alpha}^\top \mathbf{K}_1 \boldsymbol{\alpha}, \dots, \boldsymbol{\alpha}^\top \mathbf{K}_p \boldsymbol{\alpha})^\top$ 
7       for  $k \leftarrow 1$  to  $p$  do
8            $\boldsymbol{\mu}'_k \leftarrow \max(0, \boldsymbol{\mu}'_k)$ 
9            $\boldsymbol{\mu}' \leftarrow \boldsymbol{\mu}_0 + \Lambda \frac{\boldsymbol{\mu}' - \boldsymbol{\mu}_0}{\|\boldsymbol{\mu}' - \boldsymbol{\mu}_0\|}$ 
10  return  $\boldsymbol{\mu}'$ 
```

Interpolated Step KRR - L₂ Reg.

INTERPOLATEDITERATIVEALGORITHM($(\mathbf{K}_k)_{k \in [1, p]}, \boldsymbol{\mu}_0$)

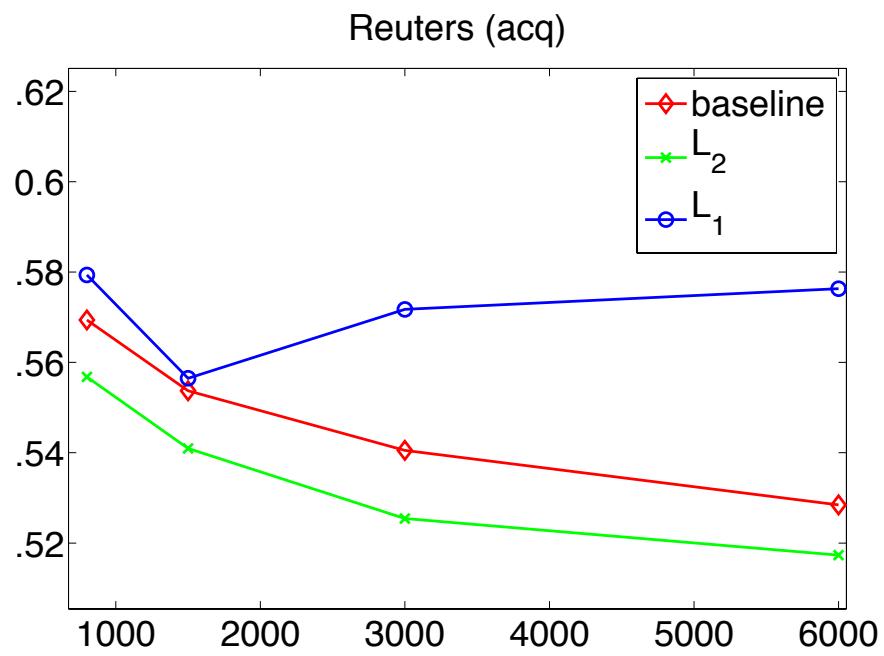
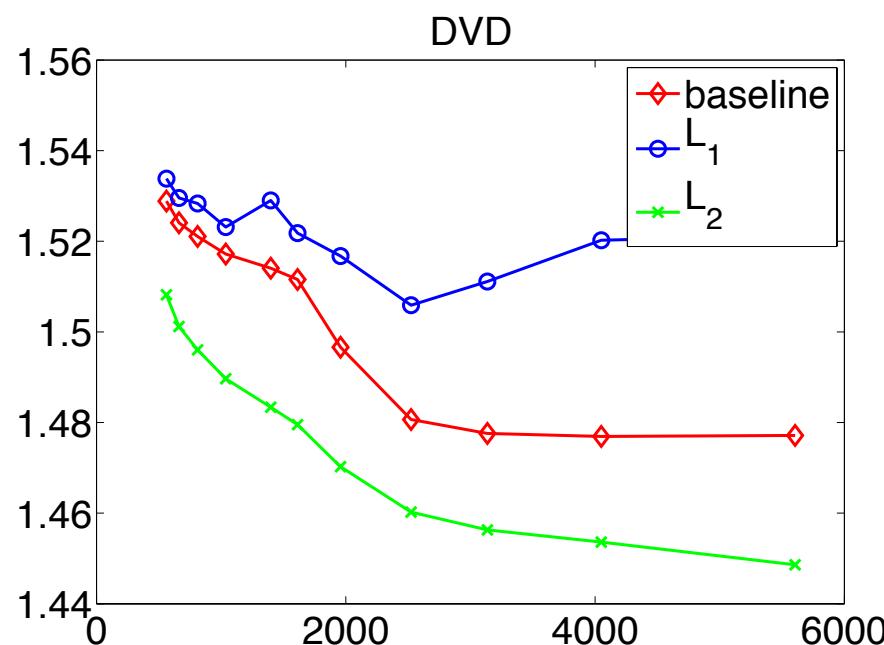
```
1    $\alpha \leftarrow \infty$ 
2    $\alpha' \leftarrow (\mathbf{K}_{\boldsymbol{\mu}_0} + \lambda \mathbf{I})^{-1} \mathbf{y}$ 
3   while  $\|\alpha' - \alpha\| > \epsilon$  do
4        $\alpha \leftarrow \alpha'$ 
5        $\mathbf{v} \leftarrow (\alpha^\top \mathbf{K}_1 \alpha, \dots, \alpha^\top \mathbf{K}_p \alpha)^\top$ 
6        $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_0 + \Lambda \frac{\mathbf{v}}{\|\mathbf{v}\|}$ 
7        $\alpha' \leftarrow \eta \alpha + (1 - \eta)(\mathbf{K}_{\boldsymbol{\mu}} + \lambda \mathbf{I})^{-1} \mathbf{y}$ 
8   return  $\alpha'$ 
```

Simple and very efficient: few iterations (less than 15).

L_2 -Regularized Combinations

(Cortes, MM, and Rostamizadeh, 2009)

- Dense combinations are beneficial when using many kernels.
- Combining kernels based on single features, can be viewed as principled feature weighting.



Conclusion

- Solid theoretical guarantees suggesting the use of a large number of base kernels.
- Broad literature on optimization techniques but often no significant improvement over uniform combination.
- Recent algorithms with significant improvements, in particular non-linear combinations.
- Still many theoretical and algorithmic questions left to explore.

References

- Bousquet, Olivier and Herrmann, Daniel J. L. On the complexity of learning the kernel matrix. In NIPS, 2002.
- Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local Rademacher complexity. In Proceedings of NIPS, 2013.
- Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In NIPS, 2009.
- Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Generalization Bounds for Learning Kernels. In ICML, 2010.
- Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Two-stage learning kernel methods. In ICML, 2010.
- Corinna Cortes, Mehryar Mohri, Afshin Rostamizadeh. Algorithms for Learning Kernels Based on Centered Alignment. JMLR 13: 795-828, 2012.

References

- Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Tutorial: Learning Kernels. ICML 2011, Bellevue, Washington, July 2011.
- Zakria Hussain, John Shawe-Taylor. Improved Loss Bounds For Multiple Kernel Learning. In AISTATS, 2011 [see arxiv for corrected version].
- Sham M. Kakade, Shai Shalev-Shwartz, Ambuj Tewari: Regularization Techniques for Learning with Matrices. JMLR 13: 1865-1890, 2012.
- Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics, 30, 2002.
- Koltchinskii, Vladimir and Yuan, Ming. Sparse recovery in large ensembles of kernel machines on-line learning and bandits. In COLT, 2008.
- Lanckriet, Gert, Cristianini, Nello, Bartlett, Peter, Ghaoui, Laurent El, and Jordan, Michael. Learning the kernel matrix with semidefinite programming. JMLR, 5, 2004.

References

- Mehmet Gönen, Ethem Alpaydin: Multiple Kernel Learning Algorithms. JMLR 12: 2211-2268 (2011).
- Srebro, Nathan and Ben-David, Shai. Learning bounds for support vector machines with learned kernels. In COLT, 2006.
- Ying, Yiming and Campbell, Colin. Generalization bounds for learning the kernel problem. In COLT, 2009.

Advanced Machine Learning

Deep Boosting

MEHRYAR MOHRI

MOHRI@

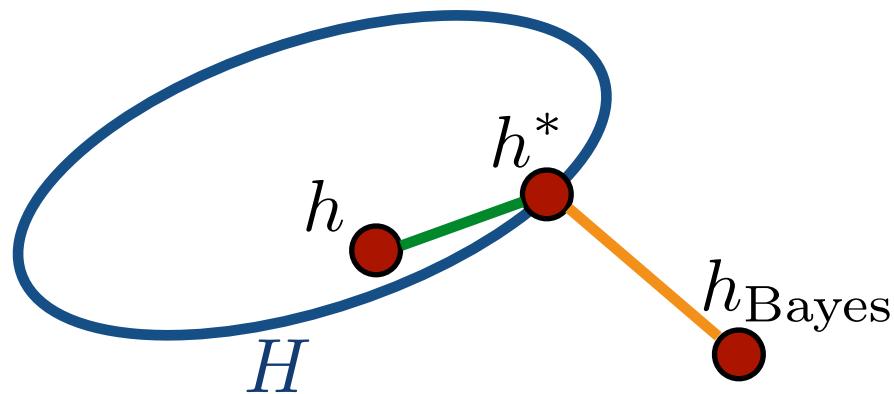
COURANT INSTITUTE & GOOGLE RESEARCH

Outline

- Model selection.
- Deep boosting.
 - theory.
 - algorithm.
 - experiments.

Model Selection

- **Problem:** how to select hypothesis set H ?
 - H too complex, no gen. bound, overfitting.
 - H too simple, gen. bound, but underfitting.
- balance between estimation and approx. errors.



Estimation and Approximation

- General equality: for any $h \in H$,

$$R(h) - R^* = \underbrace{[R(h) - R(h^*)]}_{\text{Bayes error}} + \underbrace{[R(h^*) - R^*]}_{\substack{\text{estimation} \\ \text{best in class}}}.$$

approximation

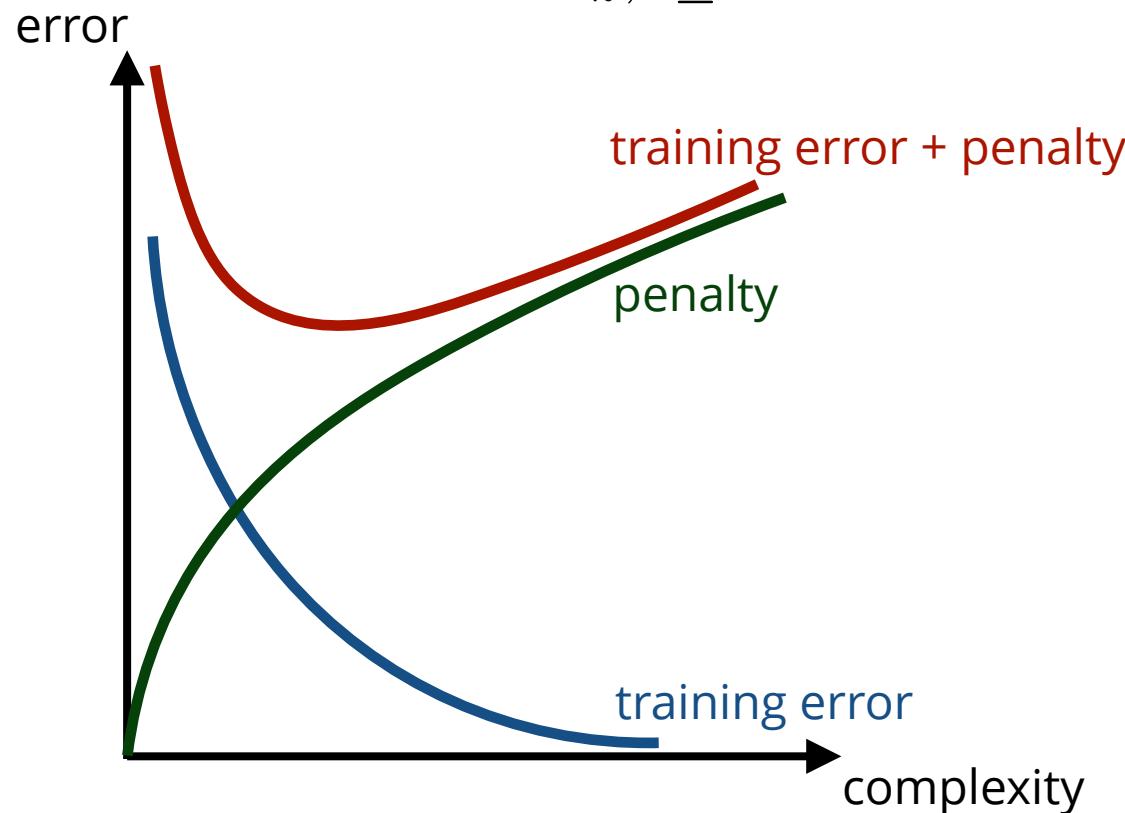
- Approximation: not a random variable, only depends on H .
- Estimation: only term we can hope to bound; for ERM, bounded by two times gen. bound:

$$\begin{aligned} R(h_{\text{ERM}}) - R(h^*) &= R(h_{\text{ERM}}) - \hat{R}(h_{\text{ERM}}) + \hat{R}(h_{\text{ERM}}) - R(h^*) \\ &\leq R(h_{\text{ERM}}) - \hat{R}(h_{\text{ERM}}) + \hat{R}(h^*) - R(h^*) \\ &\leq 2 \sup_{h \in H} |R(h) - \hat{R}(h)|. \end{aligned}$$

Structural Risk Minimization

(Vapnik and Chervonenkis, 1974; Vapnik, 1995)

- SRM: $H = \bigcup_{k=1}^{\infty} H_k$ with $H_1 \subset H_2 \subset \cdots \subset H_k \subset \cdots$
 - solution: $f^* = \operatorname{argmin}_{h \in H_k, k \geq 1} \hat{R}_S(h) + \text{pen}(k, m).$



SRM Guarantee

■ Definitions:

- $H_{k(h)}$ simplest hypothesis set containing h .
- f^* the hypothesis returned by SRM:

$$f^* = \operatorname{argmin}_{h \in H_k, k \geq 1} \widehat{R}_S(h) + R_m(H_k) + \sqrt{\frac{\log k}{m}} = F_k(h).$$

■ Theorem: for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(f^*) \leq R(h^*) + 2\mathfrak{R}_m(H_{k(h^*)}) + \sqrt{\frac{\log k(h^*)}{m}} + \sqrt{\frac{2 \log \frac{3}{\delta}}{m}}.$$

Proof

■ General bound for all $h \in H$:

$$\begin{aligned} & \Pr \left[\sup_{h \in H} R(h) - F_{k(h)}(h) > \epsilon \right] \\ &= \Pr \left[\sup_{k \geq 1} \sup_{h \in H_k} R(h) - F_k(h) > \epsilon \right] \\ &\leq \sum_{k=1}^{\infty} \Pr \left[\sup_{h \in H_k} R(h) - F_k(h) > \epsilon \right] \\ &= \sum_{k=1}^{\infty} \Pr \left[\sup_{h \in H_k} R(h) - \widehat{R}_S(h) - \mathfrak{R}_m(H_k) > \epsilon + \sqrt{\frac{\log k}{m}} \right] \\ &\leq \sum_{k=1}^{\infty} \exp \left(-2m \left[\epsilon + \sqrt{\frac{\log k}{m}} \right]^2 \right) \\ &\leq \sum_{k=1}^{\infty} e^{-2m\epsilon^2} e^{-2\log k} \\ &= e^{-2m\epsilon^2} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} e^{-2m\epsilon^2} \leq 2e^{-2m\epsilon^2}. \end{aligned}$$

Proof

■ Using the union bound and the bound just derived gives:

$$\begin{aligned} & \Pr \left[R(f^*) - R(h^*) - 2\mathfrak{R}_m(H_{k(h^*)}) - \sqrt{\frac{\log k(h^*)}{m}} > \epsilon \right] \\ & \leq \Pr \left[R(f^*) - F_{k(f^*)}(f^*) > \frac{\epsilon}{2} \right] + \Pr \left[F_{k(f^*)}(f^*) - R(h^*) - 2\mathfrak{R}_m(H_{k(h^*)}) - \sqrt{\frac{\log k(h^*)}{m}} > \frac{\epsilon}{2} \right] \\ & \leq 2e^{-\frac{m\epsilon^2}{2}} + \Pr \left[F_{k(h^*)}(h^*) - R(h^*) - 2\mathfrak{R}_m(H_{k(h^*)}) - \sqrt{\frac{\log k(h^*)}{m}} > \frac{\epsilon}{2} \right] \\ & = 2e^{-\frac{m\epsilon^2}{2}} + \Pr \left[\widehat{R}_S(h^*) - R(h^*) - \mathfrak{R}_m(H_{k(h^*)}) > \frac{\epsilon}{2} \right] \\ & = 2e^{-\frac{m\epsilon^2}{2}} + e^{-\frac{m\epsilon^2}{2}} = 3e^{-\frac{m\epsilon^2}{2}}. \end{aligned}$$

Remarks

■ SRM bound:

- similar to learning bound when $k(h^*)$ is known!
- can be extended if approximation error assumed to be small or zero.
- if H contains the Bayes classifier, only finitely many hypothesis sets need to be considered in practice.
- restriction: H decomposed as countable union of families with converging Rademacher complexity.

■ Issues: (1) SRM typically computationally intractable; (2) how should we choose H_k s?

Voted Risk Minimization

■ Ideas:

- no selection of specific H_k .
- instead, use all H_k s: $h = \sum_{k=1}^p \alpha_k h_k$, $h_k \in H_k$, $\alpha \in \Delta$.
- hypothesis-dependent penalty:

$$\sum_{k=1}^p \alpha_k \mathfrak{R}_m(H_k).$$

→ Deep ensembles.

Outline

- Model selection.

- Deep boosting.

- theory.
- algorithm.
- experiments.

Ensemble Methods in ML

- Combining several base classifiers to create a more accurate one.
 - Bagging (Breiman 1996).
 - AdaBoost (Freund and Schapire 1997).
 - Stacking (Smyth and Wolpert 1999).
 - Bayesian averaging (MacKay 1996).
 - Other averaging schemes e.g., (Freund et al. 2004).
- Often very effective in practice.
- Benefit of favorable learning guarantees.

Convex Combinations

- Base classifier set H .
 - boosting stumps.
 - decision trees with limited depth or number of leaves.
- Ensemble combinations: convex hull of base classifier set.

$$\text{conv}(H) = \left\{ \sum_{t=1}^T \alpha_t h_t : \alpha_t \geq 0; \sum_{t=1}^T \alpha_t \leq 1; \forall t, h_t \in H \right\}.$$

Ensembles - Margin Bound

(Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002)

- **Theorem:** let H be a family of real-valued functions. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f = \sum_{t=1}^T \alpha_t h_t \in \text{conv}(H)$:

$$R(f) \leq \hat{R}_{S,\rho}(f) + \frac{2}{\rho} \mathfrak{R}_m(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

- where $\hat{R}_{S,\rho}(f) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{y_i f(x_i) \leq \rho}$.

Questions

- Can we use a much richer or deeper base classifier set?
 - richer families needed for difficult tasks in speech and image processing.
 - but generalization bound indicates risk of overfitting.

AdaBoost

(Freund and Schapire, 1997)

- **Description:** coordinate descent applied to

$$F(\boldsymbol{\alpha}) = \sum_{i=1}^m e^{-y_i f(x_i)} = \sum_{i=1}^m \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i)\right).$$

- **Guarantees:** ensemble margin bound.
 - but AdaBoost does not maximize the margin!
 - some margin maximizing algorithms such as arc-gv are outperformed by AdaBoost! (Reyzin and Schapire, 2006)

Suspicions

- Complexity of hypotheses used:
 - arc-gv tends to use deeper decision trees to achieve a larger margin.

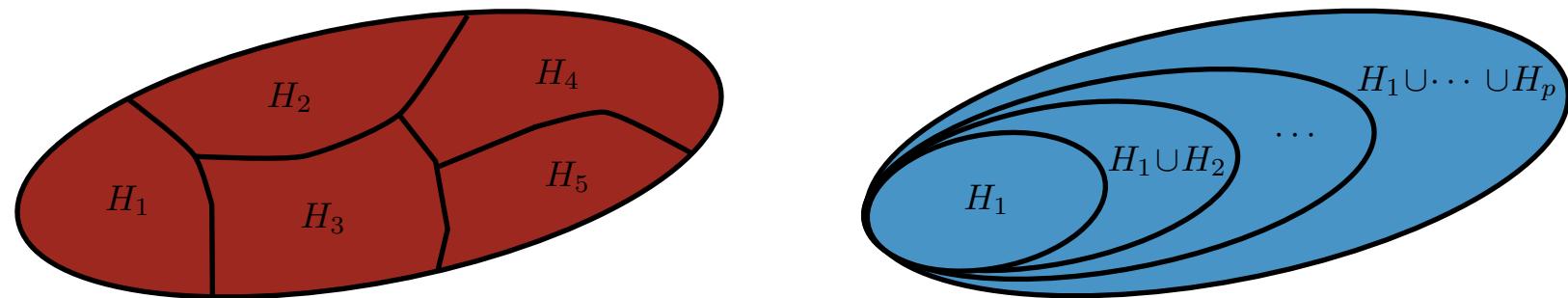
 - Notion of margin:
 - minimal margin perhaps not the appropriate notion.
 - margin distribution is key.
- can we shed more light on these questions?

Question

- **Main question:** how can we design ensemble algorithms that can succeed even with very deep decision trees or other complex sets?
 - theory.
 - algorithms.
 - experimental results.

Base Classifier Set H

- Decomposition in terms of sub-families or their union.

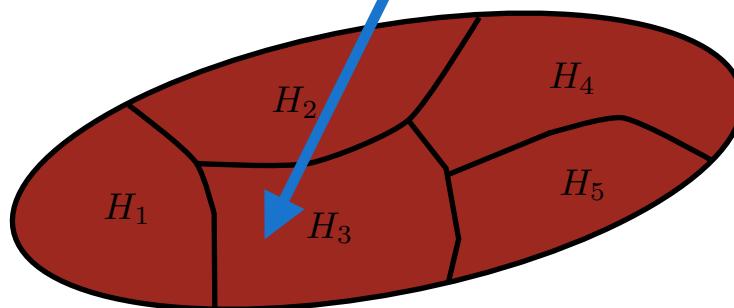


Ensemble Family

- Non-negative linear ensembles $\mathcal{F} = \text{conv}(\cup_{k=1}^p H_k)$:

$$f = \sum_{t=1}^T \alpha_t h_t$$

with $\alpha_t \geq 0$, $\sum_{t=1}^T \alpha_t \leq 1$, $h_t \in H_{k_t}$.



Ideas

- Use hypotheses drawn from H_k s with larger k s but allocate more weight to hypotheses drawn from smaller k s.
 - how can we determine quantitatively the amounts of mixture weights apportioned to different families?
 - can we provide learning guarantees guiding these choices?

Learning Guarantee

(Cortes, MM, and Syed, 2014)

- **Theorem:** Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f = \sum_{t=1}^T \alpha_t h_t \in \mathcal{F}$:

$$R(f) \leq \widehat{R}_{S,\rho}(f) + \frac{4}{\rho} \sum_{t=1}^T \alpha_t \mathfrak{R}_m(H_{k_t}) + \tilde{O}\left(\sqrt{\frac{\log p}{\rho^2 m}}\right).$$

Consequences

- Complexity term with explicit dependency on mixture weights.
 - quantitative guide for controlling weights assigned to more complex sub-families.
 - bound can be used to inspire, or directly define an ensemble algorithm.

Set-Up

- H_1, \dots, H_p : disjoint sub-families of functions taking values in $[-1, +1]$.
- Further assumption (not necessary): symmetric sub-families, i.e. $h \in H_k \Leftrightarrow -h \in H_k$.
- Notation:
 - $r_j = \mathfrak{R}_m(H_{k_j})$ with $h_j \in H_{k_j}$.

Derivation

- Learning bound suggests seeking $\alpha \geq 0$ with $\sum_{t=1}^T \alpha_t \leq 1$ to minimize

$$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{y_i \sum_{t=1}^T \alpha_t h_t(x_i) \leq \rho} + \frac{4}{\rho} \sum_{t=1}^T \alpha_t r_t.$$

Convex Surrogates

- Let $u \mapsto \Phi(-u)$ be a decreasing convex function upper bounding $u \mapsto 1_{u \leq 0}$, with Φ differentiable.
- Two principal choices:
 - Exponential loss: $\Phi(-u) = \exp(-u)$.
 - Logistic loss: $\Phi(-u) = \log_2(1 + \exp(-u))$.

Optimization Problem

(Cortes, MM, and Syed, 2014)

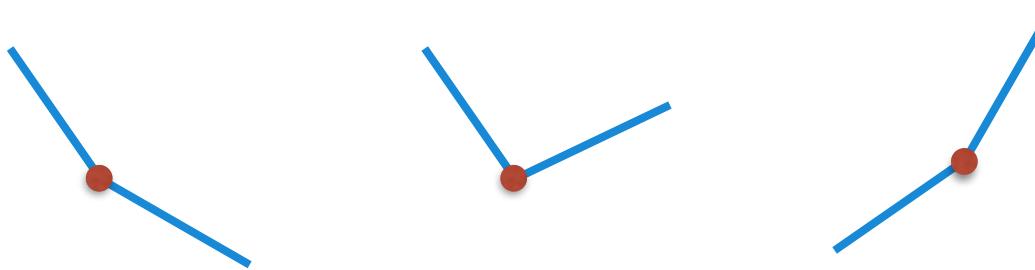
- Moving the constraint to the objective and using the fact that the sub-families are symmetric leads to:

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{m} \sum_{i=1}^m \Phi \left(1 - y_i \sum_{j=1}^N \alpha_j h_j(x_i) \right) + \sum_{t=1}^N (\lambda r_j + \beta) |\alpha_j|,$$

where $\lambda, \beta \geq 0$, and for each hypothesis, keep either h or $-h$.

DeepBoost Algorithm

- Coordinate descent applied to convex objective.
 - non-differentiable function.
 - definition of maximum coordinate descent.



Direction & Step

- Maximum direction: definition based on the error

$$\epsilon_{t,j} = \frac{1}{2} \left[1 - \underset{i \sim \mathcal{D}_t}{\mathbb{E}} [y_i h_j(x_i)] \right],$$

where D_t is the distribution over sample at iteration t .

- Step:
 - closed-form expressions for exponential and logistic losses.
 - general case: line search.

Pseudocode

```

DEEPBOOST( $S = ((x_1, y_1), \dots, (x_m, y_m))$ )
1   for  $i \leftarrow 1$  to  $m$  do
2        $D_1(i) \leftarrow \frac{1}{m}$ 
3   for  $t \leftarrow 1$  to  $T$  do
4       for  $j \leftarrow 1$  to  $N$  do
5           if ( $\alpha_{t-1,j} \neq 0$ ) then
6                $d_j \leftarrow (\epsilon_{t,j} - \frac{1}{2}) + \text{sgn}(\alpha_{t-1,j}) \frac{\Lambda_j m}{2S_t}$   $\Lambda_j = \lambda r_j + \beta.$ 
7               elseif ( $|\epsilon_{t,j} - \frac{1}{2}| \leq \frac{\Lambda_j m}{2S_t}$ ) then
8                    $d_j \leftarrow 0$ 
9               else  $d_j \leftarrow (\epsilon_{t,j} - \frac{1}{2}) - \text{sgn}(\epsilon_{t,j} - \frac{1}{2}) \frac{\Lambda_j m}{2S_t}$ 
10               $k \leftarrow \underset{j \in [1, N]}{\text{argmax}} |d_j|$ 
11               $\epsilon_t \leftarrow \epsilon_{t,k}$ 
12              if ( $|(1 - \epsilon_t)e^{\alpha_{t-1,k}} - \epsilon_t e^{-\alpha_{t-1,k}}| \leq \frac{\Lambda_k m}{S_t}$ ) then
13                   $\eta_t \leftarrow -\alpha_{t-1,k}$ 
14              elseif ( $(1 - \epsilon_t)e^{\alpha_{t-1,k}} - \epsilon_t e^{-\alpha_{t-1,k}} > \frac{\Lambda_k m}{S_t}$ ) then
15                   $\eta_t \leftarrow \log \left[ -\frac{\Lambda_k m}{2\epsilon_t S_t} + \sqrt{\left[ \frac{\Lambda_k m}{2\epsilon_t S_t} \right]^2 + \frac{1-\epsilon_t}{\epsilon_t}} \right]$ 
16              else  $\eta_t \leftarrow \log \left[ +\frac{\Lambda_k m}{2\epsilon_t S_t} + \sqrt{\left[ \frac{\Lambda_k m}{2\epsilon_t S_t} \right]^2 + \frac{1-\epsilon_t}{\epsilon_t}} \right]$ 
17               $\alpha_t \leftarrow \alpha_{t-1} + \eta_t e_k$ 
18               $S_{t+1} \leftarrow \sum_{i=1}^m \Phi' \left( 1 - y_i \sum_{j=1}^N \alpha_{t,j} h_j(x_i) \right)$ 
19              for  $i \leftarrow 1$  to  $m$  do
20                   $D_{t+1}(i) \leftarrow \frac{\Phi' \left( 1 - y_i \sum_{j=1}^N \alpha_{t,j} h_j(x_i) \right)}{S_{t+1}}$ 
21           $f \leftarrow \sum_{j=1}^N \alpha_{t,j} h_j$ 
22      return  $f$ 

```

Connections with Previous Work

- For $\lambda = \beta = 0$, DeepBoost coincides with
 - AdaBoost (Freund and Schapire 1997), run with union of sub-families, for the exponential loss.
 - additive Logistic Regression (Friedman et al., 1998), run with union of sub-families, for the logistic loss.
- For $\lambda = 0$ and $\beta \neq 0$, DeepBoost coincides with
 - L1-regularized AdaBoost (Raetsch, Mika, and Warmuth 2001), for the exponential loss.
 - coincides with L1-regularized Logistic Regression (Duchi and Singer 2009), for the logistic loss.

Rad. Complexity Estimates

- Benefit of data-dependent analysis:
 - empirical estimates of each $\mathfrak{R}_m(H_k)$.
 - example: for kernel function K_k ,

$$\widehat{\mathfrak{R}}_S(H_k) \leq \frac{\sqrt{\text{Tr}[\mathbf{K}_k]}}{m}.$$

- alternatively, upper bounds in terms of growth functions,

$$\mathfrak{R}_m(H_k) \leq \sqrt{\frac{2 \log \Pi_{H_k}(m)}{m}}.$$

Experiments (1)

- Family of base classifiers defined by boosting stumps:

- boosting stumps H_1^{stumps} (threshold functions).

- in dimension d , $\Pi_{H_1^{\text{stumps}}}(m) \leq 2md$, thus

$$\mathfrak{R}_m(H_1^{\text{stumps}}) \leq \sqrt{\frac{2 \log(2md)}{m}}.$$

- decision trees of depth 2, H_2^{stumps} , with the same question at the internal nodes of depth 1.

- in dimension d , $\Pi_{H_2^{\text{stumps}}}(m) \leq (2m)^2 \frac{d(d-1)}{2}$, thus

$$\mathfrak{R}_m(H_2^{\text{stumps}}) \leq \sqrt{\frac{2 \log(2m^2 d(d-1))}{m}}.$$

Experiments (1)

- Base classifier set: $H_1^{\text{stumps}} \cup H_2^{\text{stumps}}$.
- Data sets:
 - same UCI Irvine data sets as (Breiman 1999) and (Reyzin and Schapire 2006).
 - OCR data sets used by (Reyzin and Schapire 2006): ocr17, ocr49.
 - MNIST data sets: ocr17-mnist, ocr49-mnist.
- Experiments with exponential loss.
- Comparison with AdaBoost and AdaBoost-L1.

Data Statistics

	breastcancer	ionosphere	german (numeric)
Examples	699	351	1000
Attributes	9	34	24

	diabetes	ocr17	ocr49
Examples	768	2000	2000
Attributes	8	196	196

	ocr17-mnist	ocr49-mnist
Examples	15170	13782
Attributes	400	400

Experiments - Stumps Exp Loss

(Cortes, MM, and Syed, 2014)

Table 1. Results for boosted decision stumps and the exponential loss function.

breastcancer	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.0429	0.0437	0.0408	0.0373
(std dev)	(0.0248)	(0.0214)	(0.0223)	(0.0225)
Avg tree size	1	2	1.436	1.215
Avg no. of trees	100	100	43.6	21.6

ocr17	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.0085	0.008	0.0075	0.0070
(std dev)	0.0072	0.0054	0.0068	(0.0048)
Avg tree size	1	2	1.086	1.369
Avg no. of trees	100	100	37.8	36.9

ionosphere	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.1014	0.075	0.0708	0.0638
(std dev)	(0.0414)	(0.0413)	(0.0331)	(0.0394)
Avg tree size	1	2	1.392	1.168
Avg no. of trees	100	100	39.35	17.45

ocr49	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.0555	0.032	0.03	0.0275
(std dev)	0.0167	0.0114	0.0122	(0.0095)
Avg tree size	1	2	1.99	1.96
Avg no. of trees	100	100	99.3	96

german	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.243	0.2705	0.2455	0.2395
(std dev)	(0.0445)	(0.0487)	(0.0438)	(0.0462)
Avg tree size	1	2	1.54	1.76
Avg no. of trees	100	100	54.1	76.5

ocr17-mnist	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.0056	0.0048	0.0046	0.0040
(std dev)	0.0017	0.0014	0.0013	(0.0014)
Avg tree size	1	2	2	1.99
Avg no. of trees	100	100	100	100

diabetes	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.253	0.260	0.254	0.253
(std dev)	(0.0330)	(0.0518)	(0.04868)	(0.0510)
Avg tree size	1	2	1.9975	1.9975
Avg no. of trees	100	100	100	100

ocr49_mnist	AdaBoost H_1^{stumps}	AdaBoost H_2^{stumps}	AdaBoost-L1	DeepBoost
Error	0.0414	0.0209	0.0200	0.0177
(std dev)	0.00539	0.00521	0.00408	(0.00438)
Avg tree size	1	2	1.9975	1.9975
Avg no. of trees	100	100	100	100

Experiments (2)

- Family of base classifiers defined by decision trees of depth k . For trees with at most n nodes:

$$\mathfrak{R}_m(T_n) \leq \sqrt{\frac{(4n + 2) \log_2(d + 2) \log(m + 1)}{m}}.$$

- Base classifier set: $\cup_{k=1}^K H_k^{\text{trees}}$.
- Same data sets as with Experiments (1).
- Both exponential and logistic loss.
- Comparison with AdaBoost and AdaBoost-L1.

Experiments - Trees Exp Loss

(Cortes, MM, and Syed, 2014)

breastcancer	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.0267	0.0264	0.0243
(std dev)	(0.00841)	(0.0098)	(0.00797)
Avg tree size	29.1	28.9	20.9
Avg no. of trees	67.1	51.7	55.9

ocr17	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.004	0.003	0.002
(std dev)	(0.00316)	(0.00100)	(0.00100)
Avg tree size	15.0	30.4	26.0
Avg no. of trees	88.3	65.3	61.8

ionosphere	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.0661	0.0657	0.0501
(std dev)	(0.0315)	(0.0257)	(0.0316)
Avg tree size	29.8	31.4	26.1
Avg no. of trees	75.0	69.4	50.0

ocr49	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.0180	0.0175	0.0175
(std dev)	(0.00555)	(0.00357)	(0.00510)
Avg tree size	30.9	62.1	30.2
Avg no. of trees	92.4	89.0	83.0

german	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.239	0.239	0.234
(std dev)	(0.0165)	(0.0201)	(0.0148)
Avg tree size	3	7	16.0
Avg no. of trees	91.3	87.5	14.1

ocr17-mnist	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.00471	0.00471	0.00409
(std dev)	(0.0022)	(0.0021)	(0.0021)
Avg tree size	15	33.4	22.1
Avg no. of trees	88.7	66.8	59.2

diabetes	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.249	0.240	0.230
(std dev)	(0.0272)	(0.0313)	(0.0399)
Avg tree size	3	3	5.37
Avg no. of trees	45.2	28.0	19.0

ocr49-mnist	AdaBoost	AdaBoost-L1	DeepBoost
Error	0.0198	0.0197	0.0182
(std dev)	(0.00500)	(0.00512)	(0.00551)
Avg tree size	29.9	66.3	30.1
Avg no. of trees	82.4	81.1	80.9

Experiments - Trees Log Loss

(Cortes, MM, and Syed, 2014)

breastcancer	LogReg	LogReg-L1	DeepBoost
Error	0.0351	0.0264	0.0264
(std dev)	(0.0101)	(0.0120)	(0.00876)
Avg tree size	15	59.9	14.0
Avg no. of trees	65.3	16.0	23.8

ocr17	LogReg	LogReg-L1	DeepBoost
Error	0.00300	0.00400	0.00250
(std dev)	(0.00100)	(0.00141)	(0.000866)
Avg tree size	15.0	7	22.1
Avg no. of trees	75.3	53.8	25.8

ionosphere	LogReg	LogReg-L1	DeepBoost
Error	0.074	0.060	0.043
(std dev)	(0.0236)	(0.0219)	(0.0188)
Avg tree size	7	30.0	18.4
Avg no. of trees	44.7	25.3	29.5

ocr49	LogReg	LogReg-L1	DeepBoost
Error	0.0205	0.0200	0.0170
(std dev)	(0.00654)	(0.00245)	(0.00361)
Avg tree size	31.0	31.0	63.2
Avg no. of trees	63.5	54.0	37.0

german	LogReg	LogReg-L1	DeepBoost
Error	0.233	0.232	0.225
(std dev)	(0.0114)	(0.0123)	(0.0103)
Avg tree size	7	7	14.4
Avg no. of trees	72.8	66.8	67.8

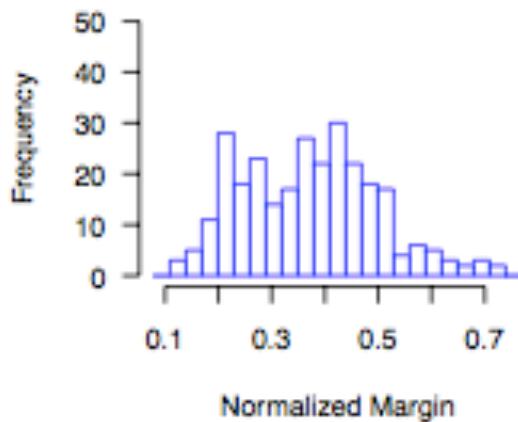
ocr17-mnist	LogReg	LogReg-L1	DeepBoost
Error	0.00422	0.00417	0.00399
(std dev)	(0.00191)	(0.00188)	(0.00211)
Avg tree size	15	15	25.9
Avg no. of trees	71.4	55.6	27.6

diabetes	LogReg	LogReg-L1	DeepBoost
Error	0.250	0.246	0.246
(std dev)	(0.0374)	(0.0356)	(0.0356)
Avg tree size	3	3	3
Avg no. of trees	46.0	45.5	45.5

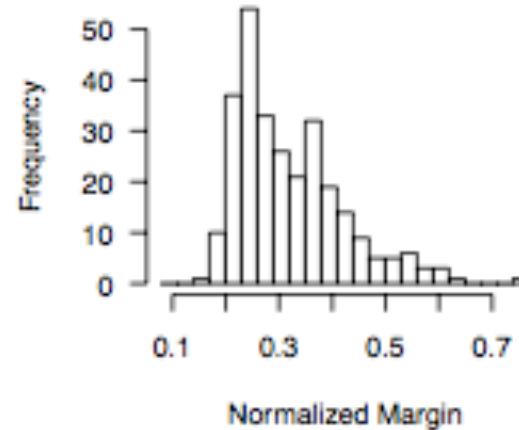
ocr49-mnist	LogReg	LogReg-L1	DeepBoost
Error	0.0211	0.0201	0.0201
(std dev)	(0.00412)	(0.00433)	(0.00411)
Avg tree size	28.7	33.5	72.8
Avg no. of trees	79.3	61.7	41.9

Margin Distribution

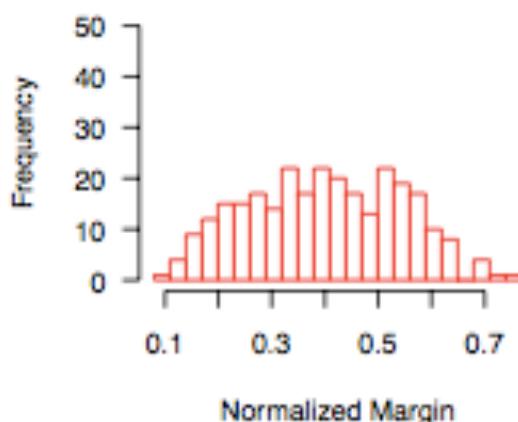
Ion: AdaBoost-L1, fold = 6



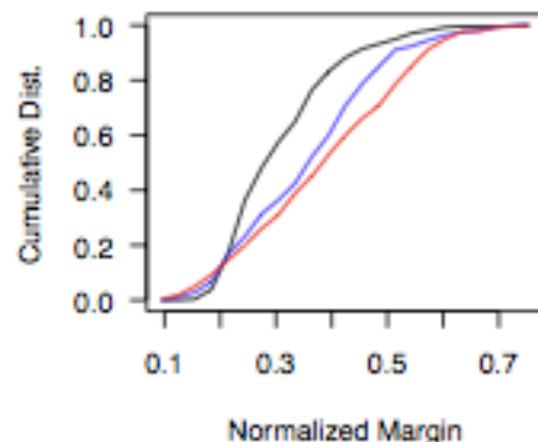
Ion: AdaBoost, fold = 6



Ion: DeepBoost, fold = 6



Cumulative Distribution of Margins



Multi-Class Learning Guarantee

(Kuznetsov, MM, and Syed, 2014)

- **Theorem:** Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f = \sum_{t=1}^T \alpha_t h_t \in \mathcal{F}$:

$$R(f) \leq \widehat{R}_{S,\rho}(f) + \frac{8c}{\rho} \sum_{t=1}^T \alpha_t \mathfrak{R}_m(\Pi_1(H_{k_t})) + O\left(\sqrt{\frac{\log p}{\rho^2 m} \log \left[\frac{\rho^2 c^2 m}{4 \log p}\right]}\right).$$

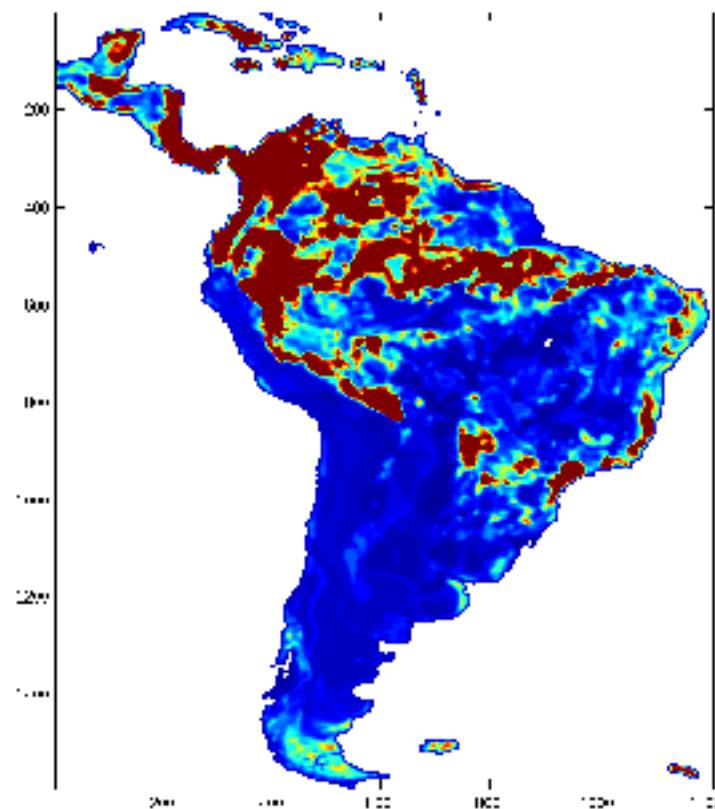
- with c number of classes.
- and $\Pi_1(H_k) = \{x \mapsto h(x, y) : y \in \mathcal{Y}, h \in H_k\}$.

Extension to Multi-Class

- Similar data-dependent learning guarantee proven for the multi-class setting.
 - bound depending on mixture weights and complexity of sub-families.
- Deep Boosting algorithm for multi-class:
 - similar extension taking into account the complexities of sub-families.
 - several variants depending on number of classes.
 - different possible loss functions for each variant.

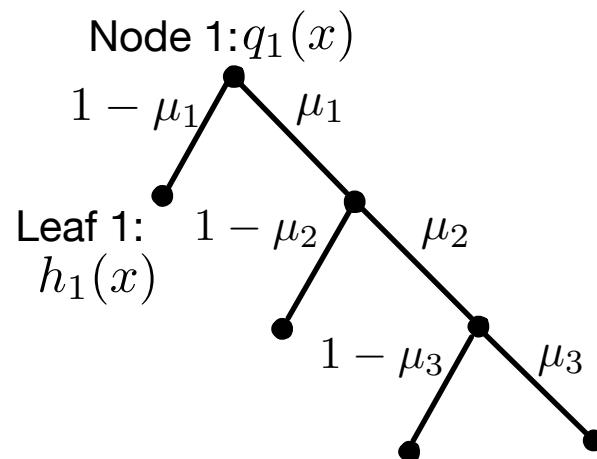
Other Related Algorithms

- Structural Maxent models (Cortes, Kuznetsov, MM, and Syed, ICML 2015): feature functions chosen from a union of very complex families.



Other Related Algorithms

- Deep Cascades (DeSalvo, MM, and Syed, ALT 2015): cascade of predictors with leaf predictors and node questions selected from very rich families.



Conclusion

- Deep Boosting: ensemble learning with increasingly complex families.
 - data-dependent theoretical analysis.
 - algorithm based on learning bound.
 - extension to multi-class.
 - ranking and other losses.
 - enhancement of many existing algorithms.
 - compares favorably to AdaBoost and additive Logistic Regression or their L1-regularized variants in experiments.

References

- Bartlett, Peter L. and Mendelson, Shahar. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3, 2002.
- Breiman, Leo. Bagging predictors. *Machine Learning*, 24 (2):123–140, 1996.
- Breiman, Leo. Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1517, 1999.
- Corinna Cortes, Mehryar Mohri, and Umar Syed. Deep boosting. In *ICML*, 2014.
- Duchi, John C. and Singer, Yoram. Boosting with structural sparsity. In *ICML*, pp. 38, 2009.
- Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Sciences*, 55(1):119–139, 1997.

References

- Koltchinskii, Vladimir and Panchenko, Dmitry. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.
- Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. In NIPS, 2014.
- Raetsch, Gunnar, Onoda, Takashi, and Mueller, Klaus- Robert. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.
- Raetsch, Gunnar, Mika, Sebastian, and Warmuth, Manfred K. On the convergence of leveraging. In NIPS, pp. 487–494, 2001.

References

- Reyzin, Lev and Schapire, Robert E. How boosting the margin can also boost classifier complexity. In ICML, pp. 753–760, 2006.
- Schapire, Robert E., Freund, Yoav, Bartlett, Peter, and Lee, Wee Sun. Boosting the margin: A new explanation for the effectiveness of voting methods. In ICML, pp. 322– 330, 1997.
- Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.
- Vladimir N. Vapnik and Alexey Ya.Chervonenkis. Theory of Pattern Recognition. Nauka, Moscow, 1974.

Advanced Machine Learning

Structured Prediction

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Structured Prediction

- Structured output:

$$\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l.$$

- Loss function: $L: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ decomposable.

- Example: Hamming loss.

$$L(y, y') = \frac{1}{l} \sum_{k=1}^l 1_{y_k \neq y'_k}$$

- Example: edit-distance loss.

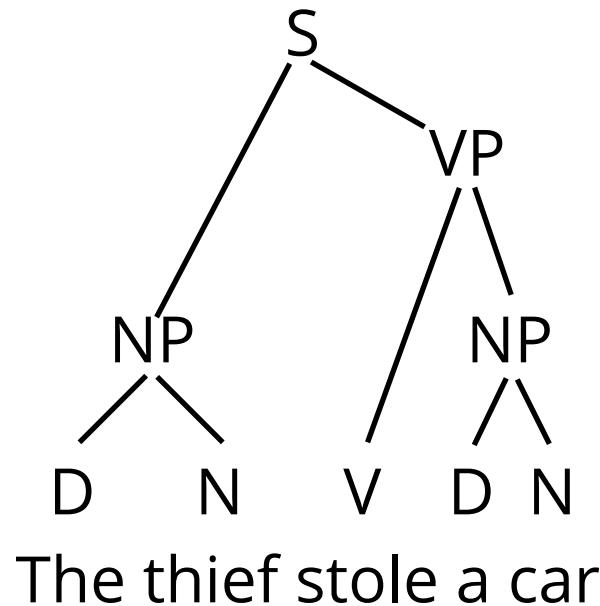
$$L(y, y') = \frac{1}{l} d_{\text{edit}}(y_1 \cdots y_l, y'_1 \cdots y'_l).$$

Examples

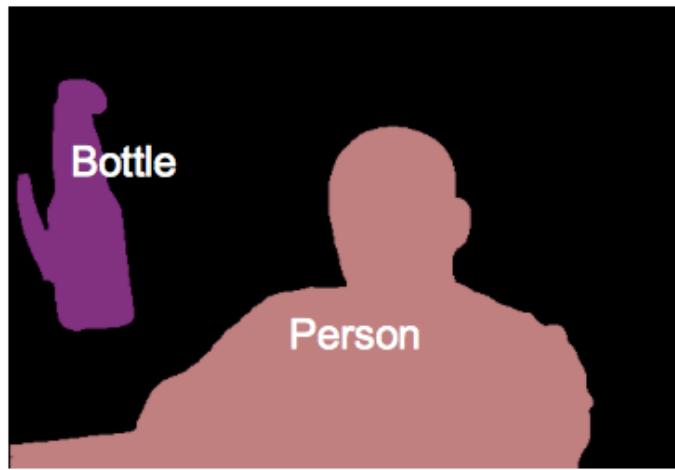
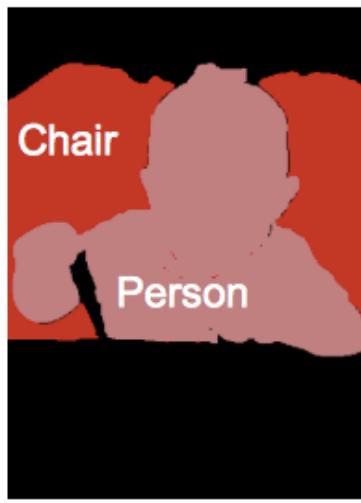
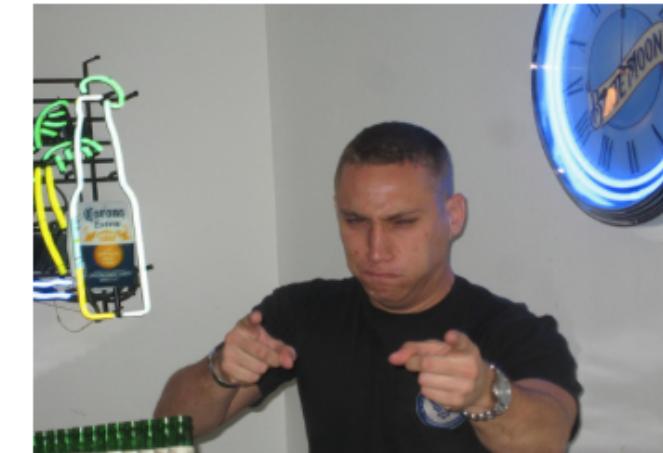
- Pronunciation modeling.
- Part-of-speech tagging.
- Named-entity recognition.
- Context-free parsing.
- Dependency parsing.
- Machine translation.
- Image segmentation.

Examples: NLP Tasks

- Pronunciation: I have formulated a
ay hh ae v f ow r m y ax l ey t ih d ax
- POS tagging: The thief stole a car
D N V D N
- Context-free parsing/Dependency parsing:



Examples: Image Segmentation



Predictors

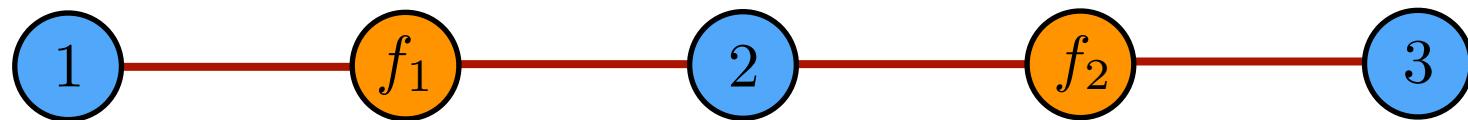
- Family of scoring functions \mathcal{H} mapping from $\mathcal{X} \times \mathcal{Y}$ to \mathbb{R} .
- For any $h \in \mathcal{H}$, prediction based on highest score:

$$\forall x \in \mathcal{X}, h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x, y).$$

- Decomposition as a sum modeled by factor graphs.

Factor Graph Examples

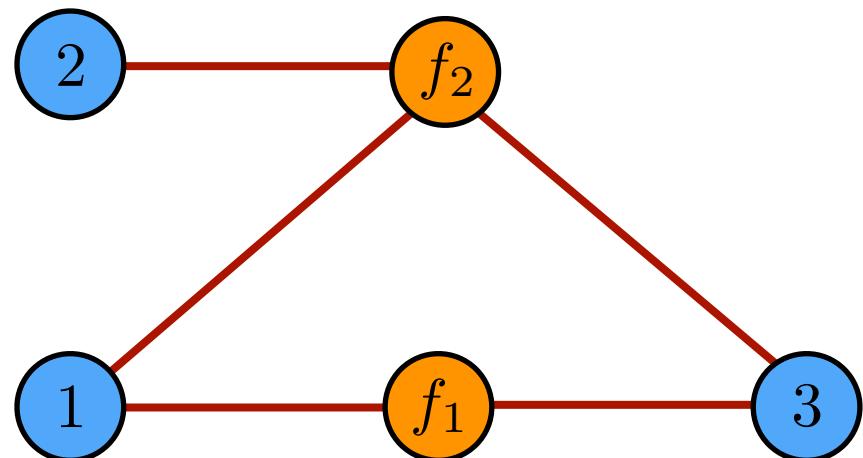
- Pairwise Markov network decomposition:



$$h(x, y) = h_{f_1}(x, y_1, y_2) + h_{f_2}(x, y_2, y_3).$$

- Other decomposition:

$$h(x, y) = h_{f_1}(x, y_1, y_3) + \\ h_{f_2}(x, y_1, y_2, y_3).$$



Factor Graphs

- $G = (V, F, E)$: factor graph.
- $\mathcal{N}(f)$: neighborhood of f .
- $\mathcal{Y}_f = \prod_{k \in \mathcal{N}(f)} \mathcal{Y}_k$: substructure set cross-product at f .
- Decomposition:

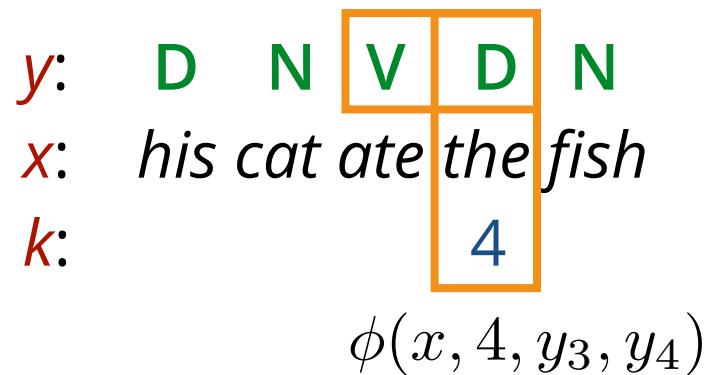
$$h(x, y) = \sum_{f \in F} h_f(x, y_f).$$

- More generally, example-dependent factor graph,

$$G_i = G(x_i, y_i) = (V_i, F_i, E_i).$$

Linear Hypotheses

- Feature decomposition → Hypothesis decomposition.
 - Example: bigram decomposition.



$$\Phi(x, y) = \sum_{s=1}^l \phi(x, s, y_{s-1}, y_s).$$

$$h(x, y) = \mathbf{w} \cdot \Phi(x, y) = \sum_{s=1}^l \underbrace{\mathbf{w} \cdot \phi(x, s, y_{s-1}, y_s)}_{h_s(x, y_{s-1}, y_s)}.$$

Structured Prediction Problem

- **Training data:** sample drawn i.i.d. from $\mathcal{X} \times \mathcal{Y}$ according to some distribution \mathcal{D} ,

$$S = ((x_1, y_1), \dots, (x_m, y_m)) \in \mathcal{X} \times \mathcal{Y}.$$

- **Problem:** find hypothesis $h: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ in \mathcal{H} with small expected loss:

$$R(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathsf{L}(h(x), y)].$$

- learning guarantees?
- role of factor graph?
- better algorithms?

Outline

- Generalization bounds.
- Algorithms.

Learning Guarantees

- Standard multi-class learning bounds:
 - number of classes is exponential!
- Structured prediction bounds:
 - covering number bounds: Hamming loss, linear hypotheses ([Taskar et al., 2003](#)).
 - PAC-Bayesian bounds (randomized algorithms) ([David McAllester, 2007](#)).
→ can we derive learning guarantees for general hypothesis sets and general loss functions?

Covering Number Bound

(Taskar et al., 2003)

- **Theorem:** fix $\rho > 0$. Then, with probability at least $1 - \rho$ over the choice of sample S of size m , the following holds for any hypothesis $h: (x, y) \rightarrow \mathbf{w} \cdot \Phi(x, y)$:

$$\mathbb{E}_{(x,y) \sim D} [L_H(h, x, y)] \leq \frac{1}{m} \sum_{i=1}^m \sup_{f \in \mathcal{F}_S^\rho(h)} L_H(f, x_i, y_i) + O\left(\sqrt{\frac{1}{m} \frac{R^2 \|\mathbf{w}\|^2}{\rho^2} (\log m + \log l + \log \max_k |\mathcal{Y}_k|)}\right),$$

where $\mathcal{F}_S^\rho(h) = \{f: X \times Y \rightarrow \mathbb{R} \mid \forall y \in Y, \forall i \in [1, m], |f(x_i, y) - h(x_i, y)| \leq \rho H(y, y_i)\}$.

Factor Graph Complexity

(Cortes, Kuznetsov, MM, Yang, 2016)

- Empirical factor graph complexity for hypothesis set \mathcal{H} and sample $S = (x_1, \dots, x_m)$:

$$\begin{aligned}\widehat{\mathfrak{R}}_S^G(\mathcal{H}) &= \frac{1}{m} \mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} \sqrt{|F_i|} \epsilon_{i,f,y} h_f(x_i, y) \right] \\ &= \mathbb{E}_\epsilon \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \underbrace{\begin{bmatrix} \vdots \\ \epsilon_{i,f,y} \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} \vdots \\ \sqrt{|F_i|} h_f(x_i, y) \\ \vdots \end{bmatrix}}_{\text{correlation with random noise}} \right].\end{aligned}$$

- Factor graph complexity:

$$\mathfrak{R}_m^G(\mathcal{H}) = \mathbb{E}_{S \sim \mathcal{D}^m} [\widehat{\mathfrak{R}}_S^G(\mathcal{H})].$$

Margin

- **Definition:** the margin of h at a labeled point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is

$$\rho_h(x, y) = \min_{y' \neq y} h(x, y) - h(x, y').$$

- error when $\rho_h(x, y) \leq 0$.
- small margin interpreted as low confidence.

Loss Function

■ Assumptions:

- bounded: $\max_{y,y'} L(y, y') \leq M$ for some $M > 0$.
- definite: $L(y, y') = 0 \Rightarrow y = y'$.

■ Consequence:

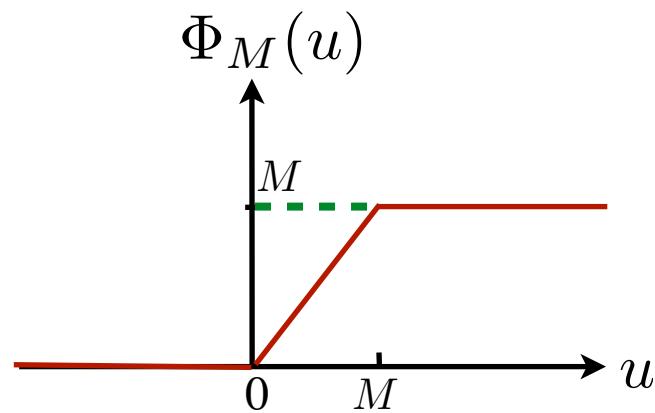
$$L(h(x), y) = L(h(x), y) \mathbf{1}_{\rho_h(x, y) \leq 0}.$$

Empirical Margin Losses

- For any $\rho > 0$,

$$\widehat{R}_{S,\rho}^{\text{add}}(h) = \mathbb{E}_{(x,y) \sim S} \left[\Phi_M \left(\max_{y' \neq y} \mathsf{L}(y', y) - \frac{h(x,y) - h(x,y')}{\rho} \right) \right]$$

$$\widehat{R}_{S,\rho}^{\text{mult}}(h) = \mathbb{E}_{(x,y) \sim S} \left[\Phi_M \left(\max_{y' \neq y} \mathsf{L}(y', y) \left(1 - \frac{h(x,y) - h(x,y')}{\rho} \right) \right) \right],$$



Generalization Bounds

(Cortes, Kuznetsov, MM, Yang, 2016)

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \widehat{R}_{S,\rho}^{\text{add}}(h) + \frac{4\sqrt{2}}{\rho} \mathfrak{R}_m^G(\mathcal{H}) + M \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

$$R(h) \leq \widehat{R}_{S,\rho}^{\text{mult}}(h) + \frac{4\sqrt{2}M}{\rho} \mathfrak{R}_m^G(\mathcal{H}) + M \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

- tightest margin bounds for structured prediction.
- data-dependent.
- improves upon bound of (Taskar et al., 2003) by log terms (in the special case they study).

Linear Hypotheses

- Hypothesis set used by most convex structured prediction algorithms (StructSVM, M3N, CRF):

$$\mathcal{H}_p = \left\{ (x, y) \mapsto \mathbf{w} \cdot \Psi(x, y) : \mathbf{w} \in \mathbb{R}^N, \|\mathbf{w}\|_p \leq \Lambda_p \right\},$$

with $p \geq 1$ and $\Psi(x, y) = \sum_{f \in F} \Psi_f(x, y_f)$.

Complexity Bounds

- Bounds on factor graph complexity of linear hypothesis sets:

$$\widehat{\mathfrak{R}}_S^G(\mathcal{H}_1) \leq \frac{\Lambda_1 r_\infty \sqrt{s \log(2N)}}{m}$$

$$\widehat{\mathfrak{R}}_S^G(\mathcal{H}_2) \leq \frac{\Lambda_2 r_2 \sqrt{\sum_{i=1}^m \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} |F_i|}}{m}$$

with $r_q = \max_{i,f,y} \|\Psi_f(x_i, y)\|_q$

$$s = \max_{j \in [1, N]} \sum_{i=1}^m \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} |F_i| \mathbf{1}_{\Psi_{f,j}(x_i, y) \neq 0}.$$

Key Term

■ Sparsity parameter:

$$s \leq \sum_{i=1}^m \sum_{f \in F_i} \sum_{y \in \mathcal{Y}_f} |F_i| \leq \sum_{i=1}^m |F_i|^2 d_i \leq m \max_i |F_i|^2 d_i,$$

where $d_i = \max_{f \in F_i} |\mathcal{Y}_f|$.

- • factor graph complexity in $O(\sqrt{\log(N) \max_i |F_i|^2 d_i / m})$ for hypothesis set \mathcal{H}_1 .
- key term: average factor graph size.

NLP Applications

■ Features:

- $\Psi_{f,j}$ is often a binary function, non-zero for a single pair $(x, y) \in \mathcal{X} \times \mathcal{Y}_f$.
- example: presence of n-gram (indexed by j) at position f of the output with input sentence x_i .
- complexity term only in $O\left(\max_i |F_i| \sqrt{\log(N)/m}\right)$.

Theory Takeaways

- Key generalization terms:
 - average size of factor graphs.
 - empirical margin loss.
- But, is learning with very complex hypothesis sets (factor graph complexity) possible?
 - richer families needed for difficult NLP tasks.
 - but generalization bound indicates risk of overfitting.

→ Voted Risk Minimization (VRM) theory

(Cortes, Kuznetsov, MM, Yang, 2016).

Outline

- Generalization bounds.
- Algorithms.

Surrogate Loss

- **Lemma:** for any $u \in \mathbb{R}_+$, let $\Phi_u: \mathbb{R} \rightarrow \mathbb{R}$ be an upper bound on $v \mapsto u1_{v \leq 0}$. Then, the following upper bound holds for any $h \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$L(h(x), y) \leq \max_{y' \neq y} \Phi_{L(y', y)}(h(x, y) - h(x, y')).$$

- **Proof:** if $h(x) \neq y$, then the following holds:

$$\begin{aligned} L(h(x), y) &= L(h(x), y)1_{\rho_h(x, y) \leq 0} \\ &\leq \Phi_{L(h(x), y)}(\rho_h(x, y)) \\ &= \Phi_{L(h(x), y)}(h(x, y) - \max_{y' \neq y} h(x, y')) \\ &= \Phi_{L(h(x), y)}(h(x, y) - h(x, h(x))) \\ &\leq \max_{y' \neq y} \Phi_{L(y', y)}(h(x, y) - h(x, y')), \end{aligned}$$

Φ -Choices

■ Different algorithms:

- StructSVM: $\Phi_u(v) = \max(0, u(1 - v))$.
- M3N: $\Phi_u(v) = \max(0, u - v)$.
- CRF: $\Phi_u(v) = \log(1 + e^{u-v})$.
- StructBoost: $\Phi_u(v) = ue^{-v}$ (Cortes, Kuznetsov, MM, Yang, 2016).

Algorithms

- StructSVM
- Maximum Margin Markov Networks (M3N)
- Conditional Random Fields (CRF)
- Regression for Learning Transducers (RLT)

Linear Prediction

- **Features:** function $\Phi: X \times Y \rightarrow \mathbb{R}^N$.
- **Hypothesis set:** functions $h: X \rightarrow Y$ of the form

$$h(x) = \operatorname{argmax}_{y \in Y} \mathbf{w} \cdot \Phi(x, y),$$

where the vector \mathbf{w} is learned from data.

- **Formulation:**
 - scoring functions.
 - multi-class classification.
 - margin: $\rho_{\mathbf{w}}(x_i, y_i) = \mathbf{w} \cdot \Phi(x_i, y_i) - \max_{y \neq y_i} \mathbf{w} \cdot \Phi(x_i, y)$.

Multi-Class SVM

■ Optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \max_{y \neq y_i} \left(0, 1 - \mathbf{w} \cdot [\Phi(\mathbf{x}_i, y_i) - \Phi(\mathbf{x}_i, y)] \right)_+.$$

■ Decision function:

$$x \mapsto \operatorname{argmax}_{y \in \mathcal{Y}} \mathbf{w} \cdot \Phi(x, y).$$

SVMStruct

(Tsochantaridis et al., 2005)

■ Optimization problem (StructSVM):

$$\min_{\mathbf{w}} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \max_{y \neq y_i} L(y_i, y) \max \left(0, 1 - \underbrace{\mathbf{w} \cdot [\Phi(x_i, y_i) - \Phi(x_i, y)]}_{=\rho(x_i, y_i, y)} \right).$$

- solution based on iteratively solving QP and adding most violating constraint.
- no specific assumption on loss.
- use of kernels.

M3N

(Taskar et al., 2003)

■ Optimization problem:

$$\min_{\mathbf{w}} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \max_{y \neq y_i} \max \left(0, L(y_i, y) - \underbrace{\mathbf{w} \cdot [\Phi(x_i, y_i) - \Phi(x_i, y)]}_{=\rho(x_i, y_i, y)} \right).$$

- \mathcal{Y} assumed to have a graph structure with a Markov property, typically a chain or a tree.
- loss assumed decomposable in the same way.
- polynomial-time algorithm using graphical model structure.
- use of kernels.

Equivalent Formulations

■ Optimization problems:

$$\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i$$

$$s.t. \quad \mathbf{w} \cdot [\Phi(x_i, y_i) - \Phi(x_i, y)] \geq 1 - \frac{\xi_i}{L(y, y_i)}, \xi_i \geq 0, \forall i \in [1, m], y \neq y_i.$$

$$\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i$$

$$s.t. \quad \mathbf{w} \cdot [\Phi(x_i, y_i) - \Phi(x_i, y)] \geq L(y, y_i) - \xi_i, \xi_i \geq 0, \forall i \in [1, m], y \neq y_i.$$

Dual Problem

- Optimization problem: $\Delta\Psi_i(y) = \Phi(x_i, y_i) - \Phi(x_i, y)$

$$\max_{\alpha \geq 0} \sum_{i, y \neq y_i} \alpha_{iy} - \frac{1}{2} \sum_{\substack{i, y \neq y_i \\ j, y' \neq y_j}} \alpha_{iy} \alpha_{jy'} \langle \Delta\Psi_i(y), \Delta\Psi_j(y') \rangle$$

$$s.t. \quad \sum_{y \neq y_i} \frac{\alpha_{iy}}{L(y_i, y)} \leq \frac{C}{m}, \forall i \in [1, m].$$

→ can use PDS kernel.

Optimization Solution

(Tsochantaridis et al., 2005)

- Cutting plane method: number of steps $\text{poly}\left(\frac{1}{\epsilon}, C, \max_{y,i} L(y, y_i)\right)$.

- start with empty constraints $S_i = \emptyset, i = 1 \dots m$.
- do until no new constraint:
 - for $i = 1 \dots m$ do
 - find most violating constraint:

$$\hat{y} = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmax}} \ L(y, y_i) \left[1 - \mathbf{w} \cdot [\Phi(x_i, y_i) - \Phi(x_i, y)] \right] = \xi_i(y)$$

- if $(\xi_i(\hat{y}) > \max_{y \in S_i} \xi_i(y) + \epsilon)$
 - $S_i \leftarrow S_i \cup \{\hat{y}\}$
 - $\alpha \leftarrow$ dual solution for $\cup_{i=1}^m S_i$

CRF = Cond. Maxent Model

(Lafferty et al., 2001)

- **Definition:** conditional probability distribution over the outputs $\mathbf{y} \in \mathcal{Y}$:

$$p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}) = \frac{\exp(\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}))}{Z_{\mathbf{w}}(\mathbf{x})},$$

with $Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} \exp(\mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y})).$

- \mathcal{Y} assumed to have a graph structure with a Markov property, typically a chain or a tree.

CRF

■ Optimization problem (CRFs):

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \log \sum_{y \in \mathcal{Y}} \exp \left(L(y_i, y) - \underbrace{\mathbf{w} \cdot [\Phi(x_i, y_i) - \Phi(x_i, y)]}_{=\rho(x_i, y_i, y)} \right).$$

$\max \text{ (M3N)} \xrightarrow{\text{soft-max (CRF)}}$

- comparison with M3N.
- smooth optimization problem, $O(C \log(1/\epsilon))$ solutions.

Features

■ Definitions:

- output alphabet Δ , $|\Delta| = r$.
- input: $\mathbf{x} = x_1 \cdots x_l$.
- output: $\mathbf{y} = y_1 \cdots y_l \in \Delta^l$.

■ Decomposition: bigram case.

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^l \phi(\mathbf{x}, k, y_{k-1}, y_k).$$

Prediction

■ Computation:

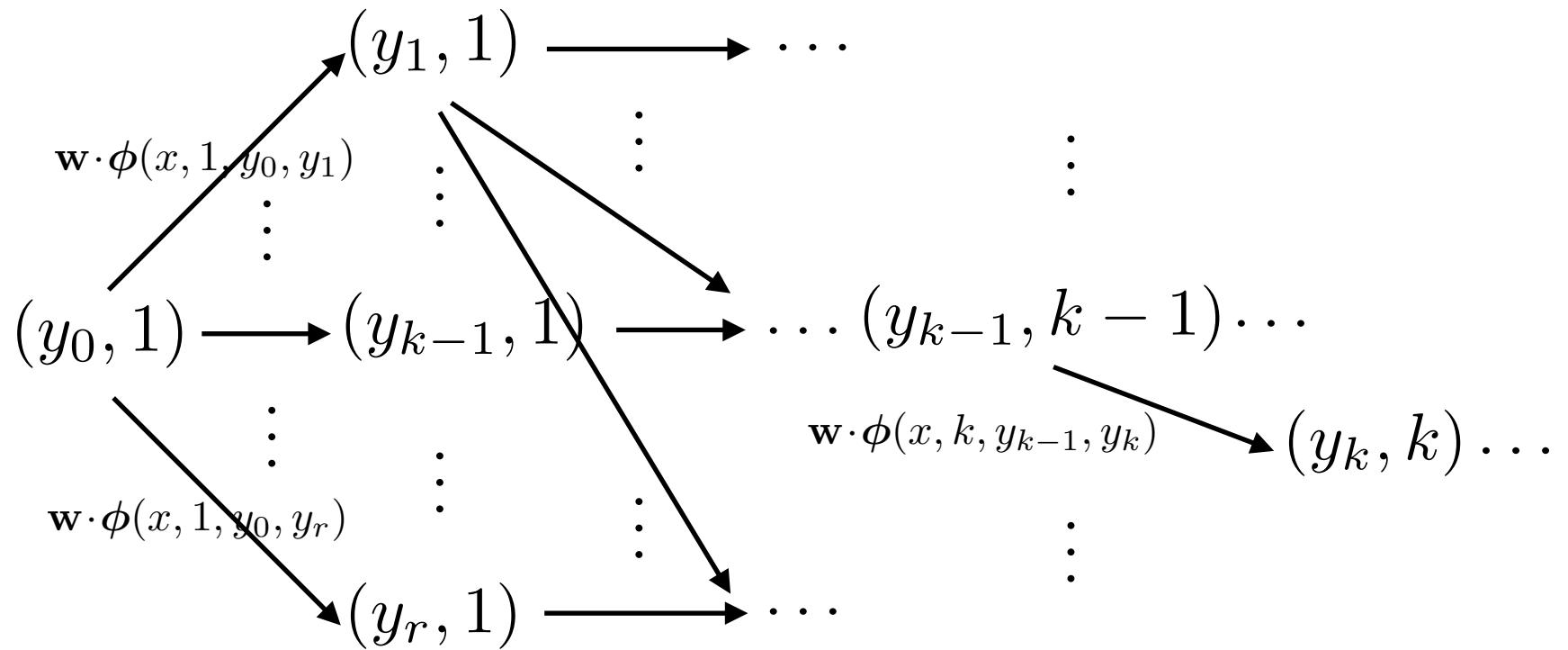
$$\underset{\mathbf{y} \in \Delta^l}{\operatorname{argmax}} \mathbf{w} \cdot \Phi(\mathbf{x}, \mathbf{y}) = \underset{\mathbf{y} \in \Delta^l}{\operatorname{argmax}} \sum_{k=1}^l \mathbf{w} \cdot \phi(\mathbf{x}, k, y_{k-1}, y_k).$$

- exponentially many possible outputs.

■ Solution:

- cast as single-source shortest-distance problem in acyclic directed graph with $(r^2 l + r)$ edges.
- linear-time algorithms: standard acyclic shortest-distance algorithm (Lawler) or the Viterbi algorithm.

Directed Graph



$$y_0 = \epsilon.$$

Estimation

- Key term in gradient computation:

$$\nabla_{\mathbf{w}} F(\mathbf{w}) = \boxed{\frac{1}{m} \sum_{i=1}^m \underset{\mathbf{y} \sim p_{\mathbf{w}}[\cdot | \mathbf{x}_i]}{\text{E}} [\Phi(\mathbf{x}_i, \mathbf{y})] - \underset{(\mathbf{x}, \mathbf{y}) \sim S}{\text{E}} [\Phi(\mathbf{x}, \mathbf{y})] + \lambda \mathbf{w}.}$$

- Computation:

$$\begin{aligned} \underset{\mathbf{y} \sim p_{\mathbf{w}}[\cdot | \mathbf{x}_i]}{\text{E}} [\Phi(\mathbf{x}_i, \mathbf{y})] &= \sum_{\mathbf{y} \in \Delta^l} p_{\mathbf{w}}[\mathbf{y} | \mathbf{w}] \Phi(\mathbf{x}_i, \mathbf{y}) \\ &= \sum_{\mathbf{y} \in \Delta^l} p_{\mathbf{w}}[\mathbf{y} | \mathbf{w}] \left[\sum_{k=1}^l \phi(\mathbf{x}_i, k, y_{k-1}, y_k) \right] \\ &= \sum_{k=1}^l \sum_{(y, y') \in \Delta^2} \boxed{\left[\begin{array}{c} \sum_{\substack{y_{k-1}=y \\ y_k=y'}} p_{\mathbf{w}}[\mathbf{y} | \mathbf{w}] \end{array} \right]} \phi(\mathbf{x}_i, k, y, y'). \end{aligned}$$

Flow Computation

■ Decomposition:

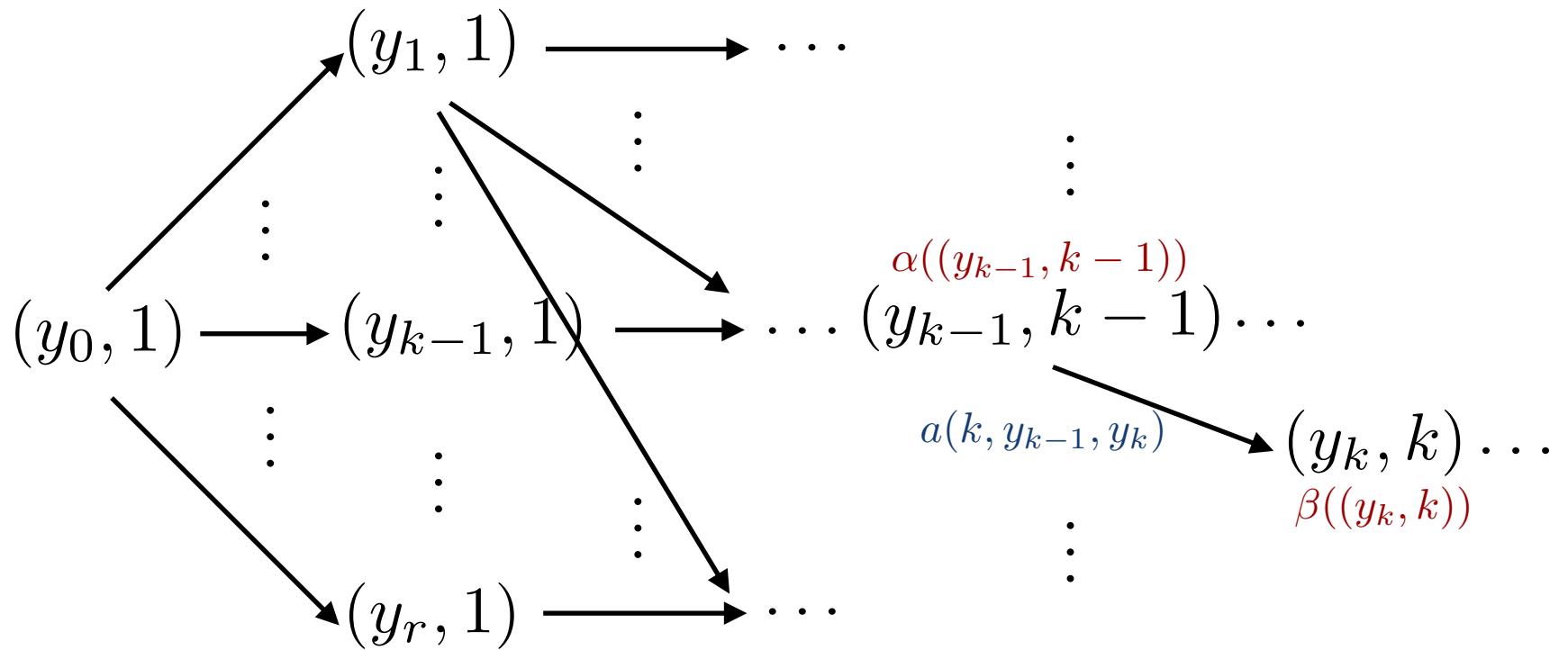
$$p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}_i) = \frac{\exp\left(\mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{y})\right)}{Z_{\mathbf{w}}(\mathbf{x}_i)}$$

$$\text{with } \exp\left(\mathbf{w} \cdot \Phi(\mathbf{x}_i, \mathbf{y})\right) = \prod_{k=1}^l \underbrace{\exp\left(\mathbf{w} \cdot \phi(\mathbf{x}_i, k, y_{k-1}, y_k)\right)}_{a(k, y_{k-1}, y_k)}.$$

■ Flow: sum of the weights of all paths going through a given transition.

- linear-time computation.
- two single-source shortest-distance algorithms.
- computational cost in $O(r^2 l)$.

Directed Graph



Computation

- Single-source shortest distance problems in $(+, \times)$:
 - $\alpha(q)$: sum of the weights of all paths from initial to q .
 - $\beta(q)$: sum of the weights of all paths from final to q .
 - linear-time algorithms for acyclic graphs.
- Partition function $Z_{\mathbf{w}}(\mathbf{x}_i)$: sum of the weights of all accepting paths, $\beta((y_0, 0))$.
- Formula:

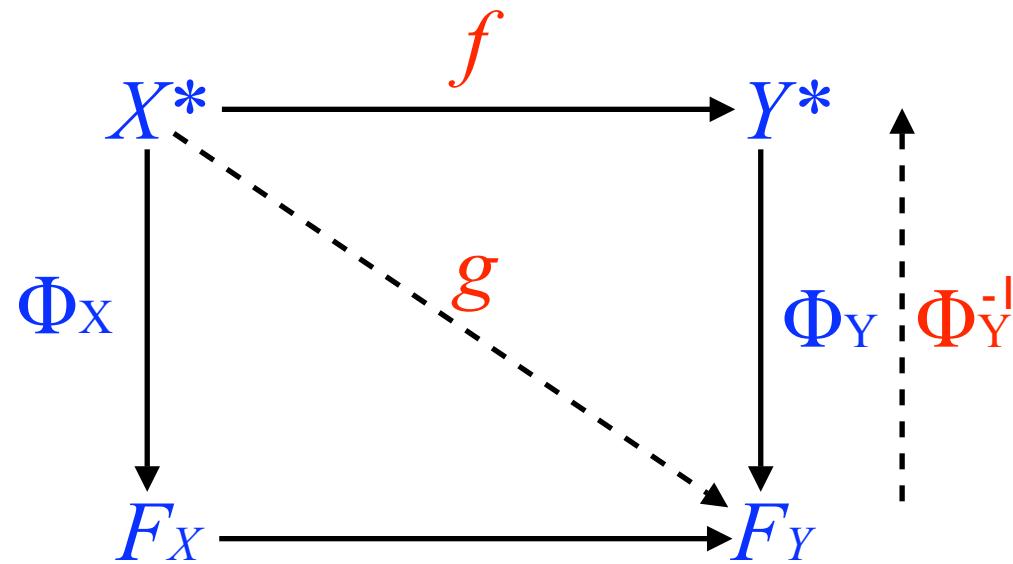
$$\sum_{\substack{y_{k-1}=y \\ y_k=y'}} p_{\mathbf{w}}[\mathbf{y} | \mathbf{w}] = \frac{\alpha((y, k-1)) \cdot a(k, y, y') \cdot \beta((y', k))}{\beta((y_0, 0))}.$$

RLT

(Cortes, MM, Weston, 2005)

■ **Definition:** formulated as a regression problem.

- learning transduction (regression).
- prediction: finding pre-image.



RLT

■ Optimization problem:

$$\operatorname{argmin}_{\mathbf{W} \in \mathbb{R}^{N_2 \times N_1}} F(\mathbf{W}) = \gamma \|\mathbf{W}\|_F^2 + \sum_{i=1}^m \|\mathbf{W} \mathbf{M}_{x_i} - \mathbf{M}_{y_i}\|^2.$$

- generalized ridge regression problem.
- closed-form solution, single matrix inversion.
- can be generalized to encoding constraints.
- use of kernels.

Solution

■ Primal:

$$\mathbf{W} = \mathbf{M}_Y \mathbf{M}_X^\top (\mathbf{M}_X \mathbf{M}_X^\top + \gamma \mathbf{I})^{-1}.$$

■ Dual:

$$\mathbf{W} = \mathbf{M}_Y (\mathbf{K}_X + \gamma \mathbf{I})^{-1} \mathbf{M}_X^\top.$$

■ Regression solution:

$$g(x) = \mathbf{W} \mathbf{M}_x.$$

Prediction

■ Prediction using kernels:

$$\begin{aligned} f(x) &= \underset{y \in Y^*}{\operatorname{argmin}} \| \mathbf{W} \mathbf{M}_x - \mathbf{M}_y \|^2 \\ &= \underset{y \in Y^*}{\operatorname{argmin}} (\mathbf{M}_y^\top \mathbf{M}_y - 2 \mathbf{M}_y^\top \mathbf{W} \mathbf{M}_x) \\ &= \underset{y \in Y^*}{\operatorname{argmin}} (\mathbf{M}_y^\top \mathbf{M}_y - 2 \mathbf{M}_y^\top \mathbf{M}_Y (\mathbf{K}_X + \gamma \mathbf{I})^{-1} \mathbf{M}_X^\top \mathbf{M}_x) \\ &= \underset{y \in Y^*}{\operatorname{argmin}} (K_Y(y, y) - 2(\mathbf{K}_Y^y)^\top (\mathbf{K}_X + \gamma \mathbf{I})^{-1} \mathbf{K}_X^x), \end{aligned}$$

with $\mathbf{K}_Y^y = \begin{bmatrix} K_Y(y, y_1) \\ \vdots \\ K_Y(y, y_m) \end{bmatrix}$ and $\mathbf{K}_X^x = \begin{bmatrix} K_X(x, x_1) \\ \vdots \\ K_X(x, x_m) \end{bmatrix}$.

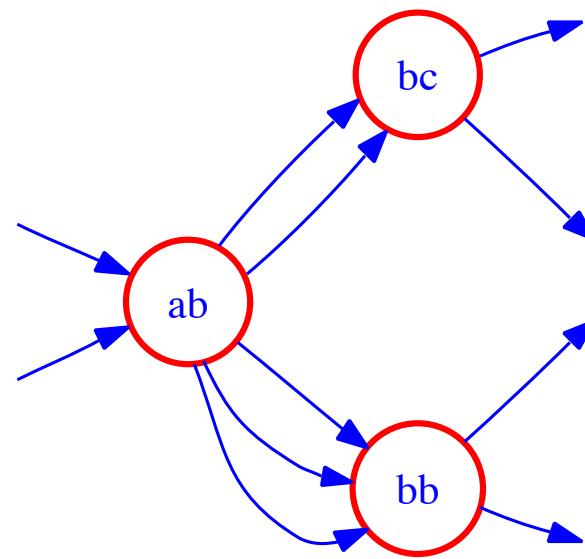
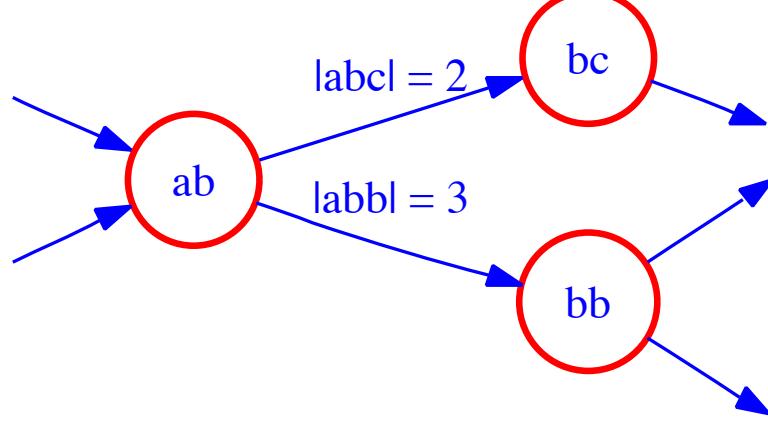
Example: N-gram kernel

- **Definition:** for any two strings y_1 and y_2 ,

$$k_n(y_1, y_2) = \sum_{|u|=n} |y_1|_u |y_2|_u.$$

Pre-Image Problem

- **Example:** pre-image for n-gram features.
 - find sequence x with matching n-gram counts.
 - use de Bruijn graph, Euler circuit.



Existence

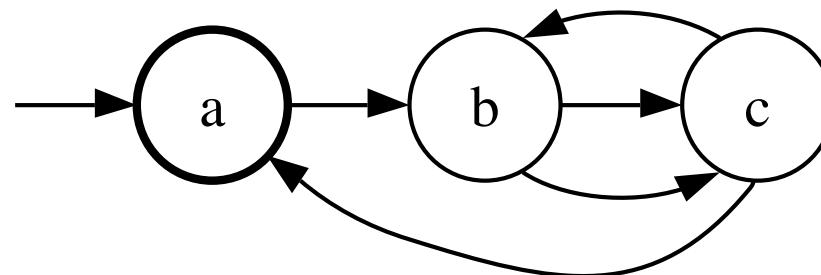
- **Theorem:** the vector of n-gram counts \mathbf{z} admits a pre-image iff for any vertex q the directed graph $G_{\mathbf{z}}$
 $\text{in-degree}(q) = \text{out-degree}(q)$.
- **Proof:** direct consequence of theorem of Euler (1736).

Pre-Image Problem

- **Example:** bigram count vector predicted

$$\mathbf{z} = (0, 1, 0, 0, 0, 2, 1, 1, 0)^\top.$$

- de Bruijn graph $G_{\mathbf{z}}$:



- Euler circuit: $x = bcbca$.

Algorithm

(Cortes, MM, Weston, 2005)

■ Algorithm:

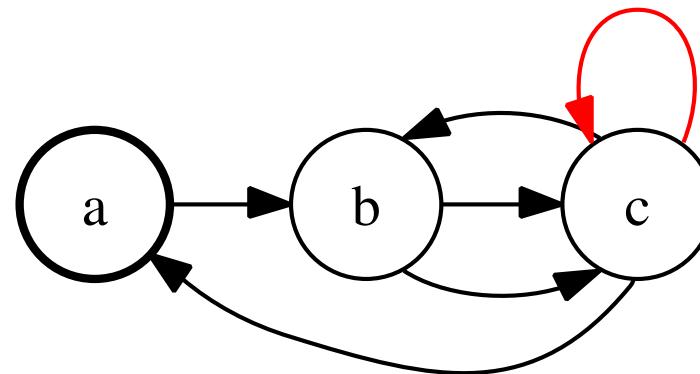
EULER(q)

```
1 path  $\leftarrow \epsilon$ 
2 for each unmarked edge  $e$  leaving  $q$  do
3     MARK( $e$ )
4     path  $\leftarrow e$  EULER( $dest(e)$ ) path
5 return path
```

- proof of correctness non-trivial.
- linear-time algorithm.

Uniqueness

- In general not unique.
- Set of strings with unique pre-image regular (Kontorovich, 2004).



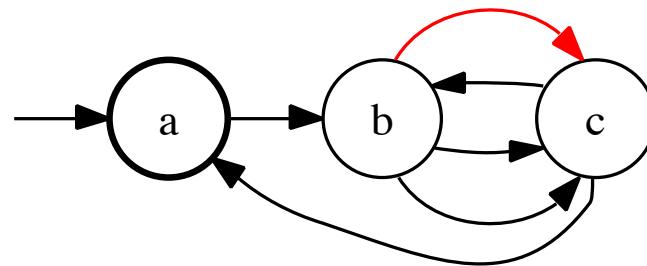
$$x = bcbcca/bccbca.$$

Generalized Euler Circuit

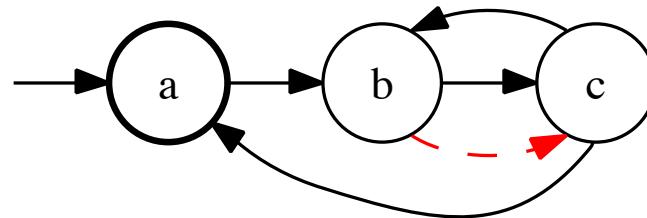
■ Extensions:

- round components of vector.
- cost of one extra or missing count for an n-gram: one local insertion or deletion.
- potentially more pre-image candidates: potentially use n-gram model to select most likely candidate.
- regression errors and potential absence of pre-image: restart Euler at every vertex for which not all edges are marked.

Illustration



$$x = bccbca/bcbcca.$$



$$x = bcba.$$

RLT

■ Benefits:

- regression formulation structured prediction problems.
- simple algorithm.
- can be generalized to regression with constraints (Cortes, MM, Weston, 2007).

■ Drawbacks:

- input-output features not natural (but constraints).
- pre-image problem for arbitrary PDS kernels?

Conclusion

- Structured prediction theory:
 - tightest margin guarantees for structured prediction.
 - general loss functions, data-dependent.
 - key notion of factor graph complexity.
 - additionally, tightest margin bounds for standard classification.

References

- Corinna Cortes, Vitaly Kuznetsov, and Mehryar Mohri. Ensemble Methods for Structured Prediction. In ICML, 2014.
- Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured Prediction Theory Based on Factor Graph Complexity. In NIPS, 2016.
- Corinna Cortes, Mehryar Mohri, and Jason Weston. A General Regression Framework for Learning String-to-String Mappings. In Predicting Structured Data. The MIT Press, 2007.
- John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In ICML, 2001.
- David McAllester. Generalization Bounds and Consistency. In Predicting Structured Data. The MIT Press, 2007.

References

- Mehryar Mohri. Semiring Frameworks and Algorithms for Shortest-Distance Problems. *Journal of Automata, Languages and Combinatorics* 7(3), 2002.
- Ben Taskar and Carlos Guestrin and Daphne Koller. Max-Margin Markov Networks. In *NIPS*, 2003.
- Ioannis Tschantzidis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large Margin Methods for Structured and Interdependent Output Variables, *JMLR*, 6, 2005.

Advanced Machine Learning

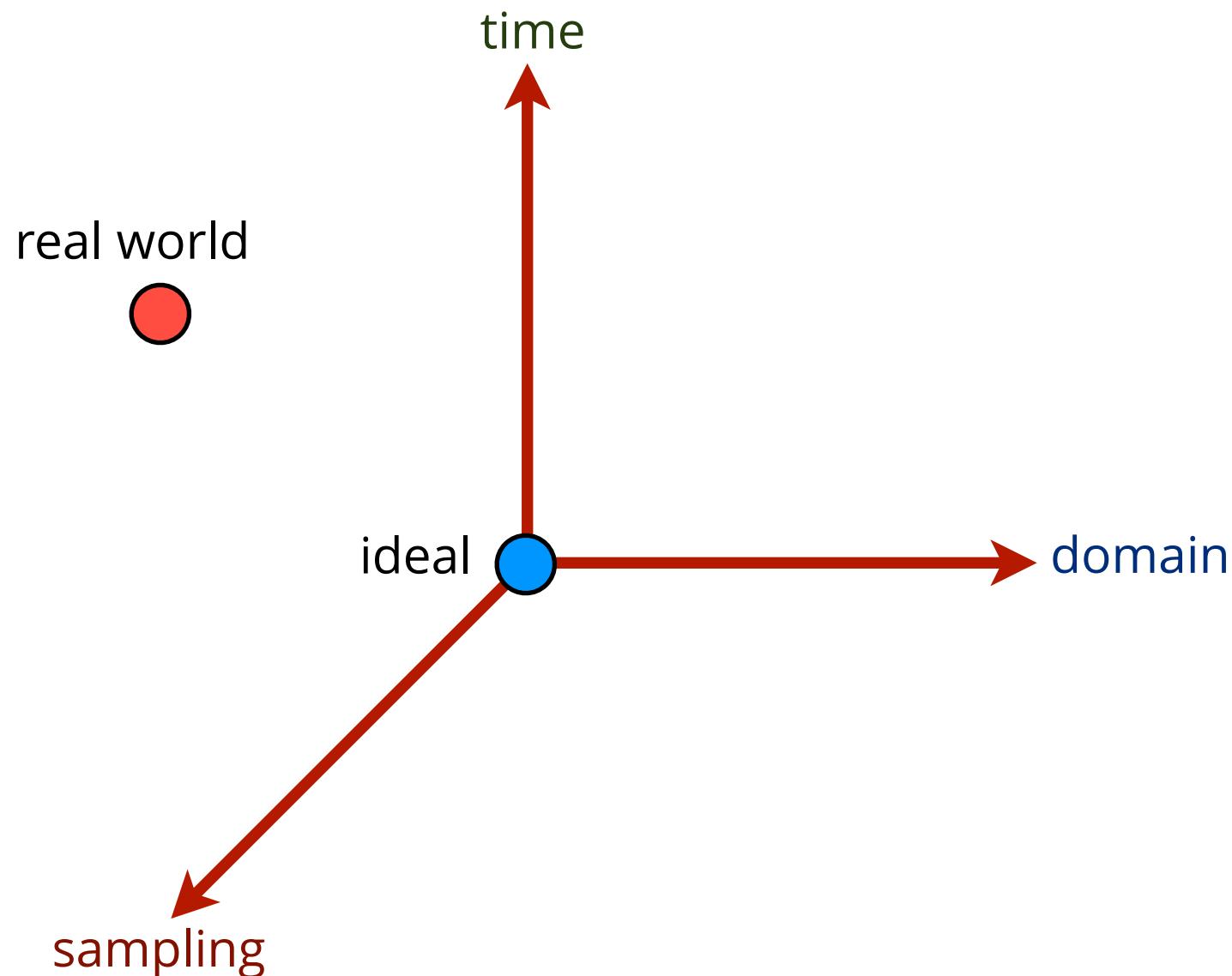
Domain Adaptation

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Non-Ideal World



Outline

- Domain adaptation.
- Multiple-source domain adaptation.

Domain Adaptation

- Sentiment analysis.
- Language modeling, part-of-speech tagging.
- Statistical parsing.
- Speech recognition.
- Computer vision.

 Solution critical for applications.

This Talk

■ Domain adaptation

- Discrepancy
- Theoretical guarantees
- Algorithm
- Enhancements

Domain Adaptation Problem

- Domains: source (Q, f_Q) , target (P, f_P) .
- Input:
 - labeled sample S drawn from source.
 - unlabeled sample T drawn from target.
- Problem: find hypothesis h in H with small expected loss with respect to target domain, that is

$$\mathcal{L}_P(h, f_P) = \mathbb{E}_{x \sim P} [L(h(x), f_P(x))].$$

Sample Bias Correction Pb

■ **Problem:** special case of domain adaptation with

- $f_Q = f_P$.
- $\text{supp}(Q) \subseteq \text{supp}(P)$.

Related Work in Theory

■ Single-source adaptation:

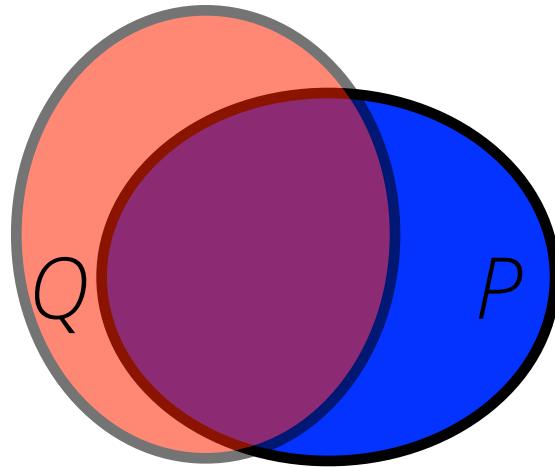
- relation between adaptation and the d_A distance (Devroye et al. (1996); Kifer et al. (2004); Ben-David et al. (2007)).
- a few negative examples of adaptation (Ben-David et al. (AISTATS 2010)).
- analysis and learning guarantees for importance weighting (Cortes, Mansour, and MM (NIPS 2010)).

Related Work in Theory

■ Multiple-source:

- same input distribution, but different labels (Crammer et al., 2005, 2006).
- theoretical analysis and method for multiple-source adaptation (Mansour, MM, Rostamizadeh, 2008).

Distribution Mismatch



Which distance should we use
to compare these distributions?

Simple Analysis

- **Proposition:** assume that the loss L is bounded by M , then

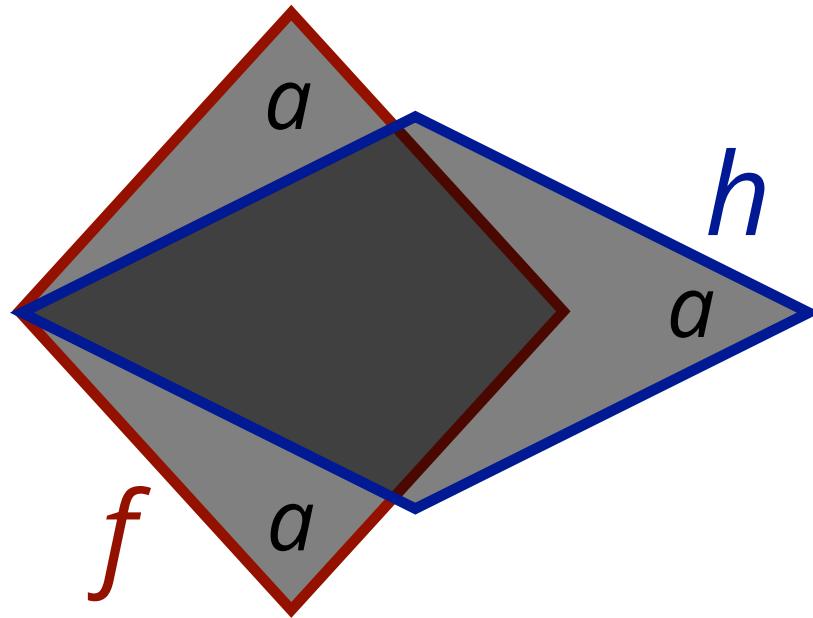
$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M L_1(Q, P).$$

- **Proof:**

$$\begin{aligned} |\mathcal{L}_P(h, f) - \mathcal{L}_Q(h, f)| &= \left| \mathbb{E}_{x \sim P} [L((h(x), f(x))] - \mathbb{E}_{x \sim Q} [L((h(x), f(x))] \right| \\ &= \left| \sum_x (P(x) - Q(x)) L((h(x), f(x)) \right| \\ &\leq M \sum_x |P(x) - Q(x)|. \end{aligned}$$

But, is this bound informative?

Example - Zero-One Loss



$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| = |Q(a) - P(a)|$$

Discrepancy

(Mansour, MM, Rostami, COLT 2009)

■ Definition:

$$\text{disc}(P, Q) = \max_{h, h' \in H} \left| \mathcal{L}_P(h, h') - \mathcal{L}_Q(h, h') \right|.$$

- symmetric, triangle inequality, in general not a distance.
- helps compare distributions for arbitrary losses, e.g. hinge loss, or L_p loss.
- generalization of d_A distance (Devroye et al. (1996); Kifer et al. (2004); Ben-David et al. (2007)).

Discrepancy - Properties

- **Theorem:** for L_q loss bounded by M , for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned} \text{disc}(P, Q) &\leq \text{disc}(\hat{P}, \hat{Q}) + 4q \left(\widehat{\mathfrak{R}}_S(H) + \widehat{\mathfrak{R}}_T(H) \right) \\ &\quad + 3M \left(\sqrt{\frac{\log \frac{4}{\delta}}{2m}} + \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \right). \end{aligned}$$

- **Proof:** Application of McDiarmid's inequality.

Discrepancy = Distance

(Cortes & MM (TCS 2013))

- **Theorem:** let K be a universal kernel (e.g., Gaussian kernel) and $H = \{h \in \mathbb{H}_K : \|h\|_K \leq \Lambda\}$. Then, for the L_2 loss, discrepancy is a distance over a compact set X .
- **Proof:** $\Psi : h \mapsto \mathbb{E}_{x \sim P}[h^2(x)] - \mathbb{E}_{x \sim Q}[h^2(x)]$ is continuous for norm $\|\cdot\|_\infty$, thus continuous on $C(X)$.
 - $\text{disc}(P, Q) = 0$ implies $\Psi(h) = 0$ for all $h \in \mathbb{H}$ since:
$$\forall h, h' \in H, \quad \left| \mathbb{E}_{x \sim P}[(h'(x) - h(x))^2] - \mathbb{E}_{x \sim Q}[(h'(x) - h(x))^2] \right| = 0.$$
 - since \mathbb{H} is dense in $C(X)$, $\Psi = 0$ over $C(X)$.
 - thus, $\mathbb{E}_P[f] - \mathbb{E}_Q[f] = 0$ for all $f \geq 0$ in $C(X)$.
 - this implies $P = Q$.

Theoretical Guarantees

■ Two types of questions:

- difference between average loss of hypothesis h on P versus Q ?
- difference of loss (measured on P) between hypothesis h obtained when training on (\hat{Q}, f_Q) versus hypothesis h' obtained when training on (\hat{P}, f_P) ?

Generalization Bound

(Mansour, MM, Rostamizadeh (COLT 2009) + MM addition)

■ Notation:

- $\mathcal{L}_Q(h_Q^*, f_Q) = \min_{h \in H} \mathcal{L}_Q(h, f_Q).$
- $\mathcal{L}_P(h_P^*, f_P) = \min_{h \in H} \mathcal{L}_P(h, f_P).$

■ Theorem: assume that L obeys the triangle inequality, then the following holds:

$$\begin{aligned}\mathcal{L}_P(h, f_P) &\leq \min_{h_Q, h_P \in H} \left\{ \mathcal{L}_Q(h, h_Q) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \right. \\ &\quad \left. + \min \{ \mathcal{L}_Q(h_Q, h_P), \mathcal{L}_P(h_Q, h_P) \} \right\}.\end{aligned}$$

Proof

$$\begin{aligned}\mathcal{L}_P(h, f_P) &\leq \min_{h_P \in H} \left\{ \mathcal{L}_P(h, h_P) + \mathcal{L}_P(h_P, f_P) \right\} && (\text{triangle ineq.}) \\ &\leq \min_{h_P \in H} \left\{ \mathcal{L}_Q(h, h_P) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \right\} \\ &\quad (\text{def. of discrepancy}) \\ &\leq \min_{h_Q, h_P \in H} \left\{ \mathcal{L}_Q(h, h_Q) + \mathcal{L}_Q(h_Q, h_P) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \right\}. \\ &\quad (\text{triangle ineq.})\end{aligned}$$

$$\begin{aligned}\mathcal{L}_P(h, f_P) &\leq \min_{h_Q, h_P \in H} \left\{ \mathcal{L}_Q(h, h_Q) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) + \min \left\{ \mathcal{L}_Q(h_Q, h_P), \mathcal{L}_P(h_Q, h_P) \right\} \right\}. \\ &\quad (\text{rerun with the opposite order of min})\end{aligned}$$

Some Natural Cases

- When $h^* = h_Q^* = h_P^*$,

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_Q(h, h^*) + \mathcal{L}_P(h^*, f_P) + \text{disc}(P, Q).$$

- When $f_P \in H$ (consistent case),

$$|\mathcal{L}_P(h, f_P) - \mathcal{L}_Q(h, f_P)| \leq \text{disc}(Q, P).$$

- Bound of (Ben-David et al., NIPS 2006) or (Blitzer et al., NIPS 2007): always worse in these cases.

Regularized ERM Algorithms

■ Objective function:

$$F_{\widehat{Q}}(h) = \lambda \|h\|_K^2 + \widehat{R}_{\widehat{Q}}(h),$$

where K is a PDS kernel;

$\lambda > 0$ is a trade-off parameter; and
 $\widehat{R}_{\widehat{Q}}(h)$ is the empirical error of h .

- broad family of algorithms including SVM, SVR, kernel ridge regression, etc.

Guarantees for Reg. ERM

(Cortes & MM (TCS 2013))

- **Theorem:** let K be a PDS kernel with $K(x, x) \leq R^2$ and L a convex loss function such that $L(\cdot, y)$ is μ -Lipschitz. Let h' be the minimizer of $F_{\widehat{P}}$ and h that of that $F_{\widehat{Q}}$, then, for all $(x, y) \in X \times Y$,

$$|L(h'(x), y) - L(h(x), y)| \leq \mu R \sqrt{\frac{\text{disc}(\widehat{P}, \widehat{Q}) + \mu \eta_H(f_P, f_Q)}{\lambda}},$$

where

$$\eta_H(f_P, f_Q) = \inf_{h \in H} \left\{ \max_{x \in \text{supp}(\widehat{P})} |f_P(x) - h(x)| + \max_{x \in \text{supp}(\widehat{Q})} |f_Q(x) - h(x)| \right\}.$$

Proof

- By the property of the minimizers, there exist subgradients such that

$$2\lambda h' = -\delta R_{\widehat{P}}(h')$$

$$2\lambda h = -\delta R_{\widehat{Q}}(h).$$

- Thus,

$$\begin{aligned} 2\lambda \|h' - h\|^2 &= -\langle h' - h, \delta R_{\widehat{P}}(h') - \delta R_{\widehat{Q}}(h) \rangle \\ &= -\langle h' - h, \delta R_{\widehat{P}}(h') \rangle + \langle h' - h, \delta R_{\widehat{Q}}(h) \rangle \\ &\leq R_{\widehat{P}}(h) - R_{\widehat{P}}(h') + R_{\widehat{Q}}(h') - R_{\widehat{Q}}(h) \\ &\leq 2\text{disc}(\widehat{P}, \widehat{Q}) + 2\mu \eta_H(f_P, f_Q). \end{aligned}$$

Proof

■ For any hypothesis h_0 , we can write:

$$\begin{aligned} 2\lambda\|h' - h\|_K^2 &\leq (\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0)) \\ &\quad + (\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)) \\ &\quad + (\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)) - (\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)). \end{aligned}$$

■ Next, by the Lipschitzness, the following holds:

$$(\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0)) \leq 2\mu \underset{x \sim \widehat{P}}{\text{E}} [|f_P(x) - h_0(x)|]$$

$$(\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)) - (\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)) \leq 2\mu \underset{x \sim \widehat{Q}}{\text{E}} [|f_Q(x) - h_0(x)|].$$

■ Since h_0 is in H , we have

$$(\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)) \leq 2 \text{disc}(\widehat{P}, \widehat{Q}).$$

Guarantees for Reg. ERM

(Cortes & MM (TCS 2013))

- **Theorem:** let K be a PDS kernel with $K(x, x) \leq R^2$ and L the L_2 loss bounded by M . Then, for all (x, y) ,

$$|L(h'(x), y) - L(h(x), y)| \leq \frac{R\sqrt{M}}{\lambda} \left(\delta + \sqrt{\delta^2 + 4\lambda \text{disc}(\hat{P}, \hat{Q})} \right),$$

where $\delta = \min_{h \in H} \left\| \mathbb{E}_{x \sim \hat{Q}} \left[(h(x) - f_Q(x)) \Phi_K(x) \right] - \mathbb{E}_{x \sim \hat{P}} \left[(h(x) - f_P(x)) \Phi_K(x) \right] \right\|_K$.

- For $f_P = f_Q = f$,
 - $\delta \leq R\epsilon$ if f is ϵ -close to H on samples.
 - $\delta = 0$ for a Gaussian kernel and f continuous.

Proof

■ For any hypothesis h_0 , we can write as for previous result:

$$\begin{aligned} 2\lambda\|h' - h\|_K^2 &\leq (\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0)) \\ &\quad + (\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)) \\ &\quad + (\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)) - (\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)). \end{aligned}$$

■ Next, for the squared loss, we have:

$$\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0) = \mathbb{E}_{x \sim \widehat{P}} [(h_0(x) - f_P(x))(2h(x) - f_P(x) - h_0(x))]$$

$$\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0) = \mathbb{E}_{x \sim \widehat{P}} [(h_0(x) - f_P(x))(2h'(x) - f_P(x) - h_0(x))].$$

■ Thus,

$$\begin{aligned} &(\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)) - (\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)) \\ &= -2 \mathbb{E}_{x \sim \widehat{Q}} [(h_0(x) - f_Q(x))(h(x) - h'(x))]. \end{aligned}$$

Proof

- As for previous theorem, we have

$$(\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)) - (\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)) \leq 2 \text{disc}(\widehat{P}, \widehat{Q}).$$

- Thus, $2\lambda \|h' - h\|_K^2 \leq 2 \text{disc}(\widehat{P}, \widehat{Q}) + 2\Delta$ with:

$$\begin{aligned}\Delta &= \left\langle h - h', \mathbb{E}_{x \sim \widehat{P}}[(h_0(x) - f_P(x))K(x, \cdot)] - \mathbb{E}_{x \sim \widehat{Q}}[(h_0(x) - f_Q(x))K(x, \cdot)] \right\rangle \\ &\leq \|h - h'\|_K \left\| \mathbb{E}_{x \sim \widehat{P}}[(h_0(x) - f_P(x))K(x, \cdot)] - \mathbb{E}_{x \sim \widehat{Q}}[(h_0(x) - f_Q(x))K(x, \cdot)] \right\|_K.\end{aligned}$$

- The result follows by solving second-degree inequality.

Empirical Discrepancy

- Discrepancy measure $\text{disc}(\hat{P}, \hat{Q})$ critical term in bounds.
- Smaller empirical discrepancy guarantees closeness of pointwise losses of h' and h .
- But, can we further reduce the discrepancy?

Algorithm - Idea

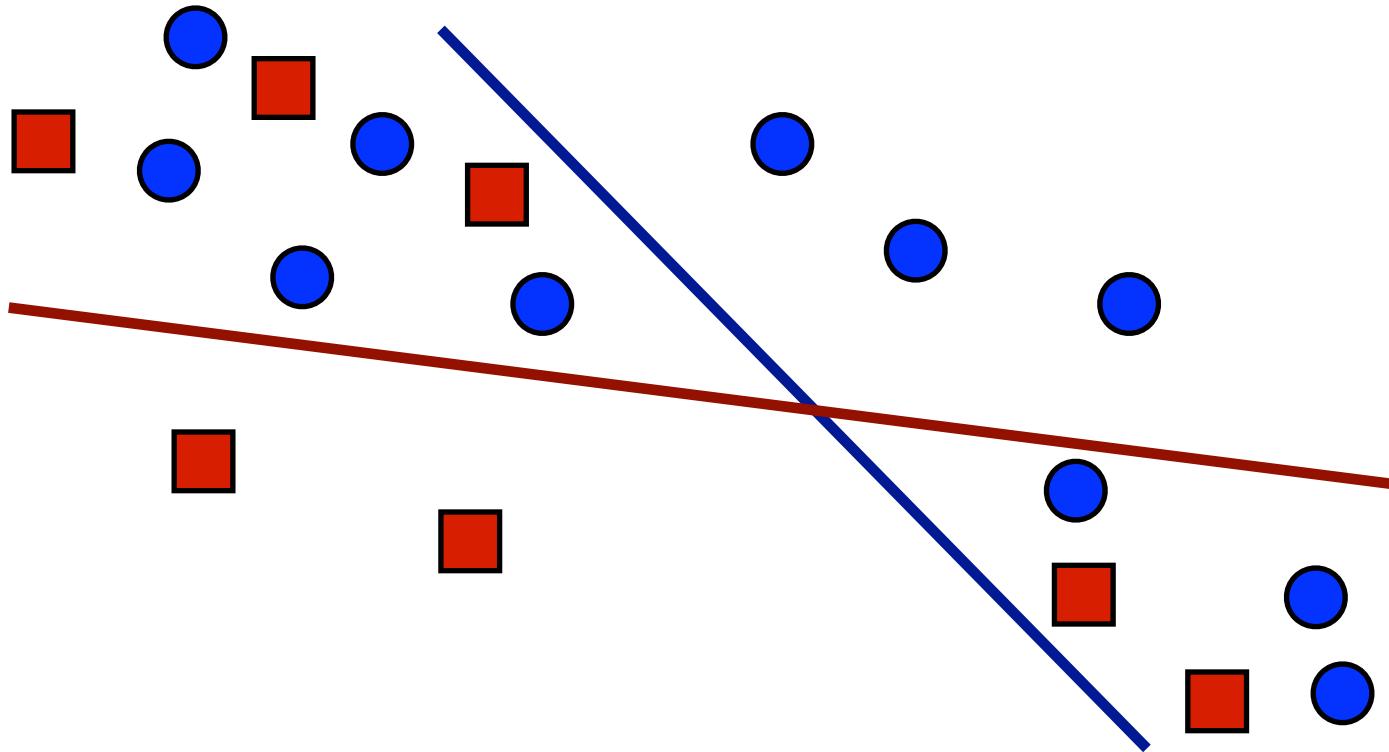
- Search for a new empirical distribution q^* with same support:

$$q^* = \operatorname{argmin}_{\text{supp}(q) \subseteq \text{supp}(\hat{Q})} \text{disc}(\hat{P}, q).$$

- Solve modified optimization problem:

$$\min_h F_{q^*}(h) = \sum_{i=1}^m q^*(x_i) L(h(x_i), y_i) + \lambda \|h\|_K^2.$$

Case of Halfspaces



Min-Max Problem

■ Reformulation:

$$\widehat{Q}' = \operatorname{argmin}_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} |\mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h)|.$$

- game theoretical interpretation.
- gives lower bound:

$$\max_{h, h' \in H} \min_{\widehat{Q}' \in \mathcal{Q}} |\mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h)| \leq$$

$$\min_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} |\mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h)|.$$

Classification - 0/1 Loss

■ Problem:

$$\min_{Q'} \max_{a \in H \Delta H} |\hat{Q}'(a) - \hat{P}(a)|$$

$$\text{subject to } \forall x \in S_Q, \hat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \hat{Q}'(x) = 1.$$

Classification - 0/1 Loss

- Linear program (LP):

$$\min_{Q'} \quad \delta$$

subject to $\forall a \in H\Delta H, \hat{Q}'(a) - \hat{P}(a) \leq \delta$

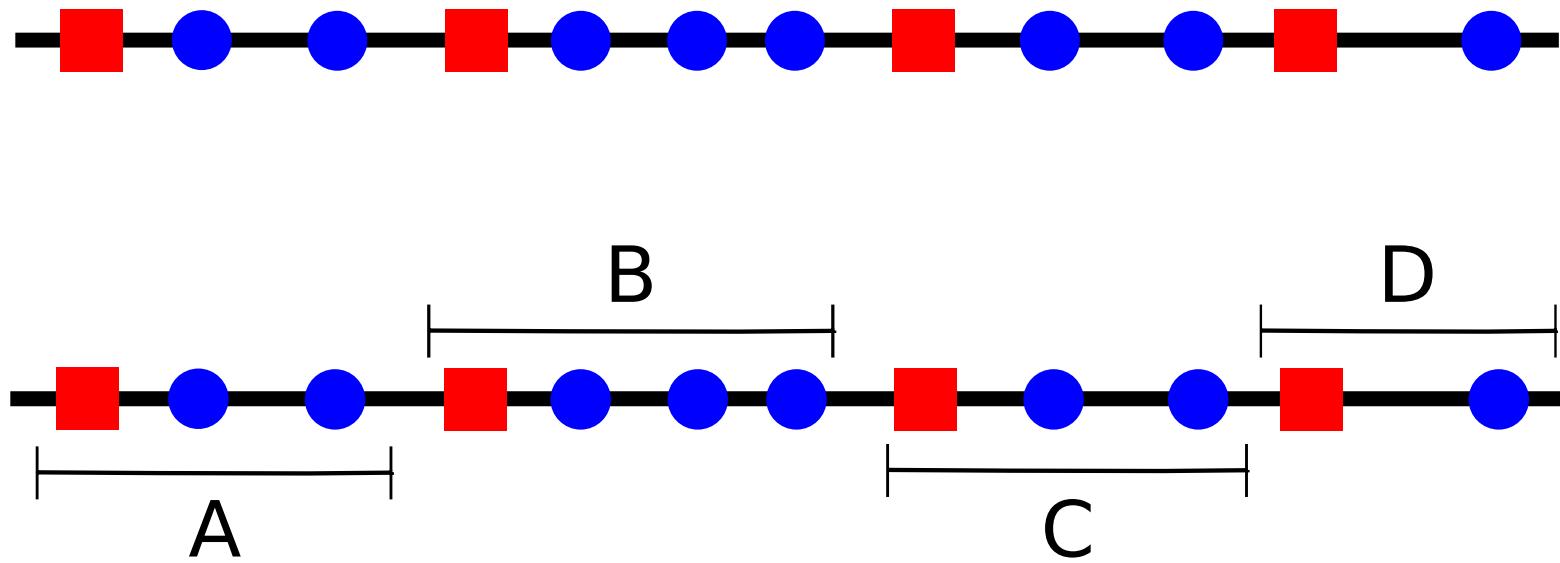
$$\forall a \in H\Delta H, \hat{P}(a) - \hat{Q}'(a) \leq \delta$$

$$\forall x \in S_Q, \hat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \hat{Q}'(x) = 1.$$

- No. of constraints bounded by shattering coefficient.

$$\Pi_{H\Delta H}(m_0 + n_0)$$

Algorithm - 1D



Regression - L2 Loss

■ Problem:

$$\min_{\hat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} \left| \mathbb{E}_{\hat{P}}[(h'(x) - h(x))^2] - \mathbb{E}_{\hat{Q}'}[(h'(x) - h(x))^2] \right|.$$

$$\min_{\hat{Q}' \in \mathcal{Q}} \max_{\substack{\|\mathbf{w}\| \leq 1 \\ \|\mathbf{w}'\| \leq 1}} \left| \mathbb{E}_{\hat{P}}[((\mathbf{w}' - \mathbf{w})^\top \mathbf{x})^2] - \mathbb{E}_{\hat{Q}'}[((\mathbf{w}' - \mathbf{w})^\top \mathbf{x})^2] \right|$$

$$= \min_{\hat{Q}' \in \mathcal{Q}} \max_{\substack{\|\mathbf{w}\| \leq 1 \\ \|\mathbf{w}'\| \leq 1}} \left| \sum_{\mathbf{x} \in S} (\hat{P}(\mathbf{x}) - \hat{Q}'(\mathbf{x}))[(\mathbf{w}' - \mathbf{w})^\top \mathbf{x}]^2 \right|$$

$$= \min_{\hat{Q}' \in \mathcal{Q}} \max_{\|\mathbf{u}\| \leq 2} \left| \sum_{\mathbf{x} \in S} (\hat{P}(\mathbf{x}) - \hat{Q}'(\mathbf{x}))[\mathbf{u}^\top \mathbf{x}]^2 \right|$$

$$= \min_{\hat{Q}' \in \mathcal{Q}} \max_{\|\mathbf{u}\| \leq 2} \left| \mathbf{u}^\top \left(\sum_{\mathbf{x} \in S} (\hat{P}(\mathbf{x}) - \hat{Q}'(\mathbf{x})) \mathbf{x} \mathbf{x}^\top \right) \mathbf{u} \right|.$$

Regression - L2 Loss

- Problem equivalent to

$$\min_{\substack{\|\mathbf{z}\|_1=1 \\ \mathbf{z} \geq 0}} \max_{\|\mathbf{u}\|=1} |\mathbf{u}^\top \mathbf{M}(\mathbf{z}) \mathbf{u}|,$$

with: $\mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{m_0} z_i \mathbf{M}_i$,

$$\mathbf{M}_0 = \sum_{\mathbf{x} \in S} P(\mathbf{x}) \mathbf{x} \mathbf{x}^\top$$

$$\mathbf{M}_i = \mathbf{s}_i \mathbf{s}_i^\top$$

elements of $\text{supp}(\hat{Q})$

Regression - L2 Loss

- Semi-definite program (SDP): linear hypotheses.

$$\min_{\mathbf{z}, \lambda} \quad \lambda$$

$$\text{subject to} \quad \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0$$

$$\lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0$$

$$\mathbf{1}^\top \mathbf{z} = 1 \wedge \mathbf{z} \geq 0,$$

where the matrix $\mathbf{M}(\mathbf{z})$ is defined by:

$$\mathbf{M}(\mathbf{z}) = \sum_{\mathbf{x} \in S} \widehat{P}(\mathbf{x}) \mathbf{x} \mathbf{x}^\top - \sum_{i=1}^{m_0} z_i \mathbf{s}_i \mathbf{s}_i^\top.$$

Regression - L2 Loss

- SDP: generalization to H RKHS for some kernel K .

$$\min_{\mathbf{z}, \lambda} \quad \lambda$$

$$\text{subject to} \quad \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0$$

$$\lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0$$

$$\mathbf{1}^\top \mathbf{z} = 1 \wedge \mathbf{z} \geq 0,$$

$$\text{with: } \mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{m_0} z_i \mathbf{M}_i$$

$$\mathbf{M}_0 = \mathbf{K}^{1/2} \operatorname{diag}(P(s_1), \dots, P(s_{p_0})) \mathbf{K}^{1/2}$$

$$\mathbf{M}_i = \mathbf{K}^{1/2} \mathbf{I}_i \mathbf{K}^{1/2}.$$

Discrepancy Min. Algorithm

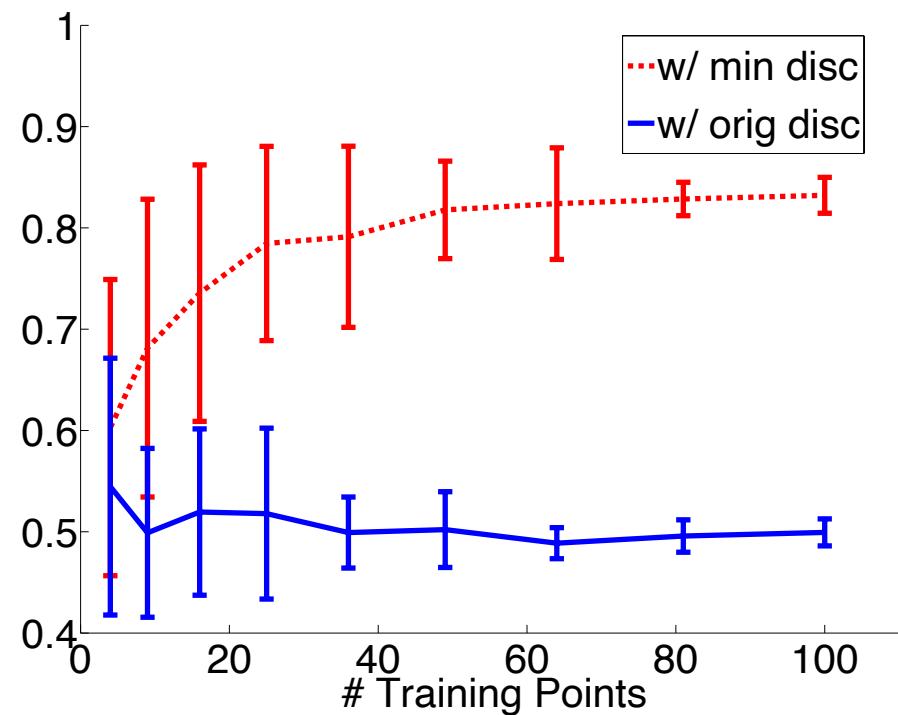
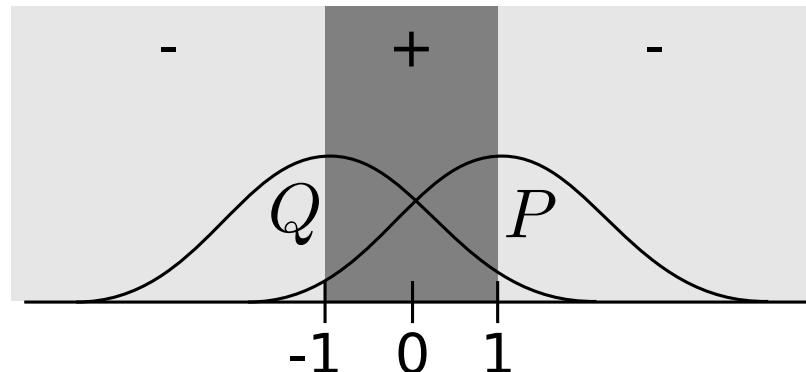
(Cortes & MM (TCS 2013))

- Convex optimization:
 - cast as semi-definite programming (SDP) prob.
 - efficient solution using smooth optimization.
- Algorithm and solution for arbitrary kernels.
- Outperforms other algorithms in experiments.

Experiments

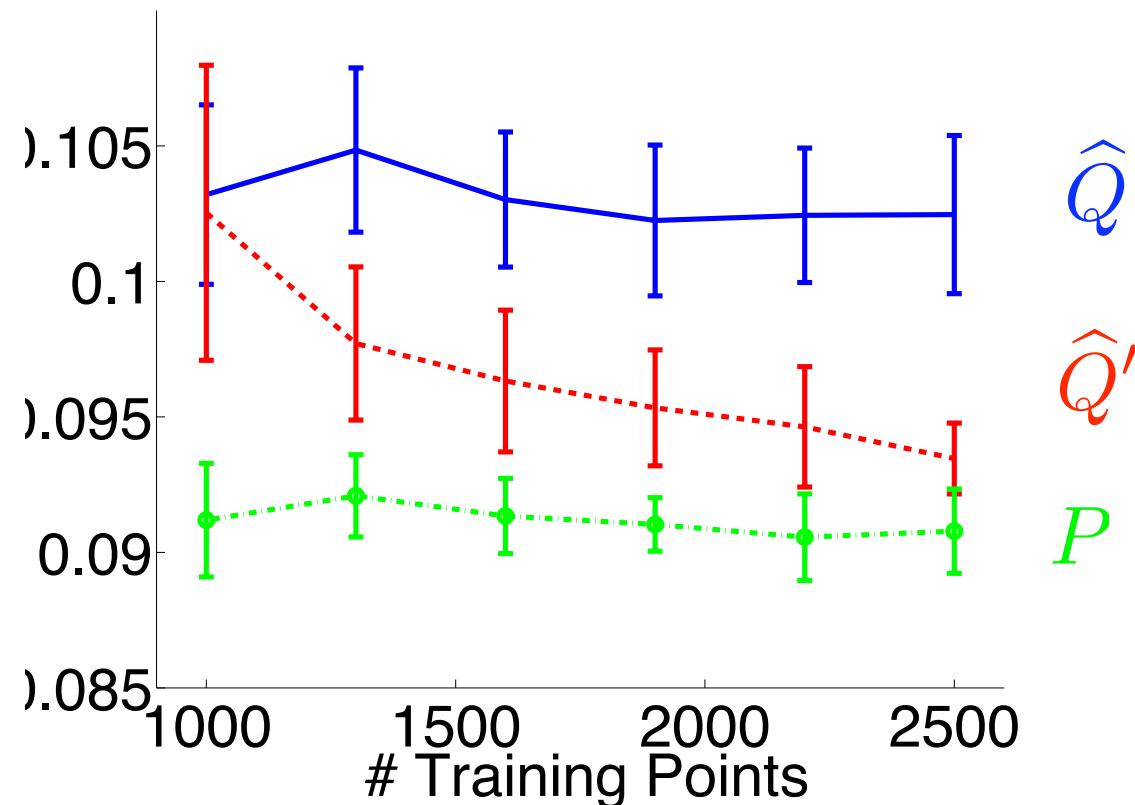
Classification:

- Q and P Gaussians.
- H : halfspaces.
- f : interval $[-1, +1]$.



Experiments

Regression:



SDP solved in about 15s using SeDuMi on 3GHz CPU with 2GB memory.

Experiments

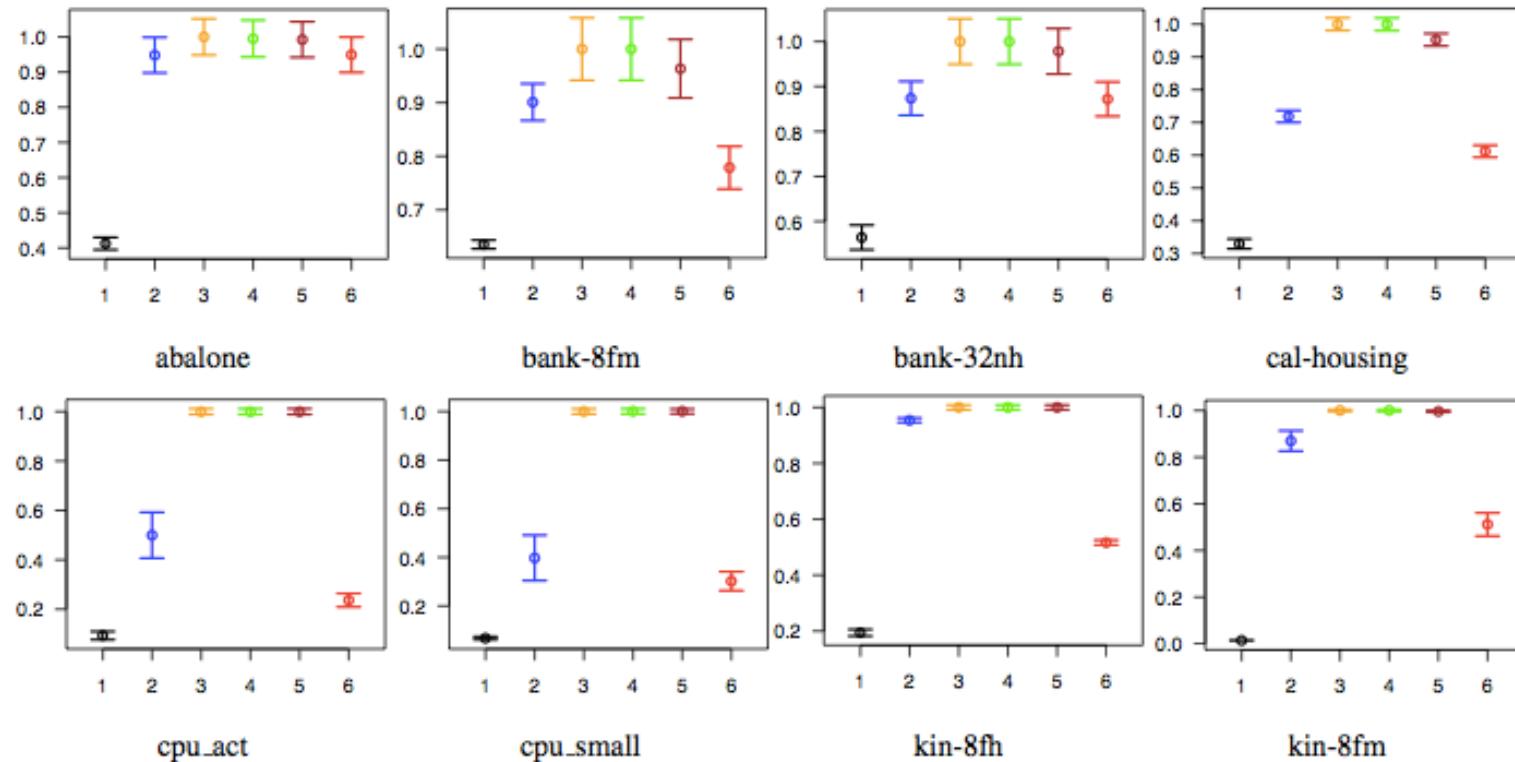


Fig. 11. Results with “easy-to-learn” biasing scheme: Relative MSE performance of (1): Optimal (in black); (2): KMM (in blue); (3): KLIEP (in orange); (4): Uniform (in green); (5): Two-Stage (in brown); and (6): DM (in red). Errors are normalized so that the average MSE of Uniform is 1.

Enhancement

(Cortes, MM, and Muñoz (2014))

■ Shortcomings:

- discrepancy depends on maximizing pair of hypotheses.
- → DM algorithm too conservative.

■ Ideas:

- finer quantity: *generalized discrepancy*, hypothesis-dependent.
- reweighting depending on hypothesis.

Algorithm

(Cortes, MM, and Muñoz (2014))

- Choose Q_h such that objectives are unif. close:

$$\lambda \|h\|_K^2 + \mathcal{L}_{Q_h}(h, f_Q)$$

$$\lambda \|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P).$$

- Ideally:

$$Q_h = \operatorname{argmin}_q |\mathcal{L}_q(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)|.$$

- Using convex surrogate H'' :

$$Q_h = \operatorname{argmin}_q \max_{h'' \in H''} |\mathcal{L}_q(h, f_Q) - \mathcal{L}(h, h'')|.$$

Optimization

(Cortes, MM, and Muñoz (2014))

$$\begin{aligned}\mathcal{L}_{Q_h}(h, f_Q) &= \operatorname{argmin}_{l \in \{\mathcal{L}_q(h, f_Q) : q \in \mathcal{F}(\mathcal{S}_X, \mathbb{R})\}} \max_{h'' \in H''} |l - \mathcal{L}_{\hat{P}}(h, h'')| \\ &= \operatorname{argmin}_{l \in \mathbb{R}} \max_{h'' \in H''} |l - \mathcal{L}_{\hat{P}}(h, h'')| \\ &= \frac{1}{2} \left(\max_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') \right).\end{aligned}$$

→ Convex optimization problem (loss jointly convex):

$$\min_h \lambda \|h\|_K^2 + \frac{1}{2} \left(\max_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') \right).$$

Convex Surrogate Hyp. Set

(Cortes, MM, and Muñoz (2014))

- Choice of H'' among balls

$$B(r) = \{h'' \in H \mid \mathcal{L}_q(h'', f_Q) \leq r^p\}.$$

- Generalization bound proven to be more favorable than DM for some choices of radius r .
- Radius r chosen via cross-validation using small amount of labeled data from target.
- Further improvement of empirical results.

Conclusion

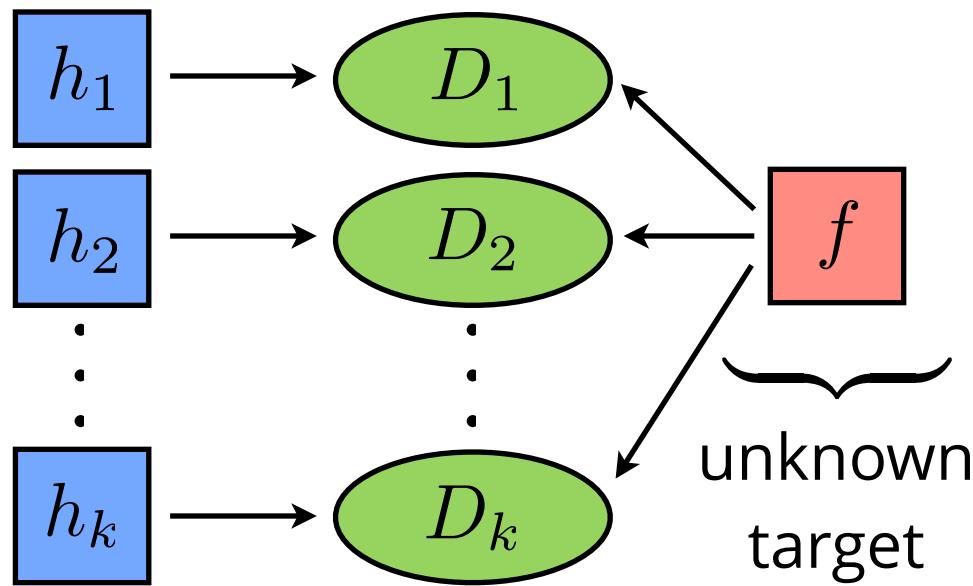
- Theory of adaptation based on discrepancy:
 - key term in analysis of adaptation and drifting.
 - discrepancy minimization algorithm DM.
 - compares favorably to other adaptation algorithms in experiments.
- Generalized discrepancy:
 - extension to hypothesis-dependent reweighting.
 - convex optimization problem.
 - further empirical improvements.
- Further generalization: [\(Awasthi, Cortes, MM, 2024\)](#).

Outline

- Domain adaptation.
- Multiple-source domain adaptation.

Problem Formulation

- Given distributions and corresponding hypotheses:



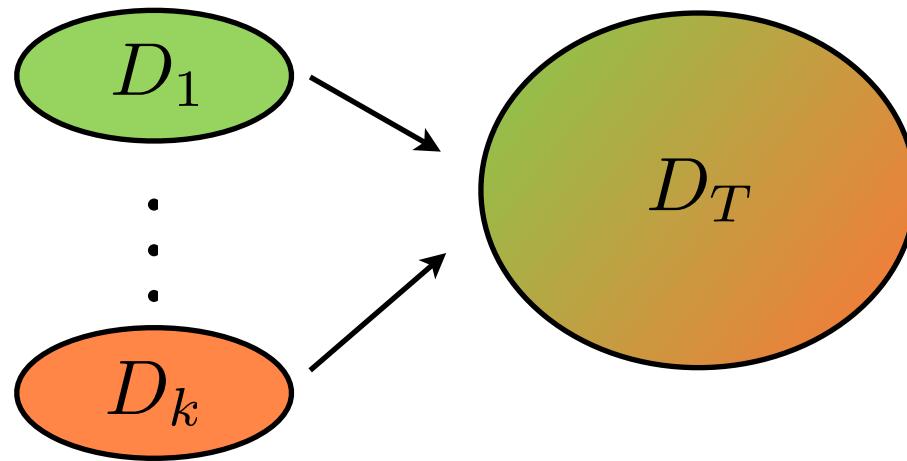
$$\forall i, \underbrace{\mathcal{L}(D_i, h_i, f) \leq \epsilon}_{\text{each hypothesis performs well in its domain.}}$$

Notation: $\mathcal{L}(D_i, h_i, f) = \underset{x \sim D_i}{\mathbb{E}} [L(h_i(x), f(x))]$.

Loss L assumed non-negative, bounded, convex and continuous.

Problem Formulation

- The **unknown** target distribution is a mixture of input distributions.



$$D_T(x) = \sum_{i=1}^k \lambda_i D_i(x)$$

- Task: choose a **hypothesis mixture** that performs well in target distribution.

$$h_z(x) = \sum_{i=1}^k z_i h_i(x)$$

convex combination rule

$$h_z(x) = \sum_{i=1}^k \frac{z_i D_i(x)}{\sum_{j=1}^k z_j D_j(x)} h_i(x)$$

distribution weighted combination

Known Target Distribution

- For some distributions, **any** convex combination performs poorly.

distribution weights

	D_T	D_0	D_1
a	0.5	I	0
b	0.5	0	I

hypothesis output

	f	h_0	h_1
a	I	I	0
b	0	I	0

- base hypotheses have no error within domain.
- any convex combination has error of 1/2!

Main Results

- Thus, although convex combinations seem natural, they can perform very poorly.
- We will show that distribution weighted combinations seem to define the “right” combination rule.
- There exists a single “robust” distribution weighted hypothesis, that does well for any target mixture.

$$\forall f, \exists z, \forall \lambda, \mathcal{L}(D_\lambda, h_z, f) \leq \epsilon.$$

Known Target Distribution

- If distribution is known, distribution weighted rule will always do well. Choose: $z = \lambda$.

$$h_\lambda(x) = \sum_{i=1}^k \frac{\lambda_i D_i(x)}{\sum_{j=1}^k \lambda_j D_j(x)} h_i(x).$$

- Proof:

$$\begin{aligned}\mathcal{L}(D_T, h_\lambda, f) &= \sum_{x \in X} L(h_\lambda(x), f(x)) D_T(x) \\ &\leq \sum_{x \in X} \sum_{i=1}^k \frac{\lambda_i D_i(x)}{D_T(x)} L(h_i(x), f(x)) D_T(x) \\ &= \sum_{i=1}^k \lambda_i \mathcal{L}(D_i, h_i(x), f(x)) \leq \sum_{i=1}^k \lambda_i \epsilon = \epsilon.\end{aligned}$$

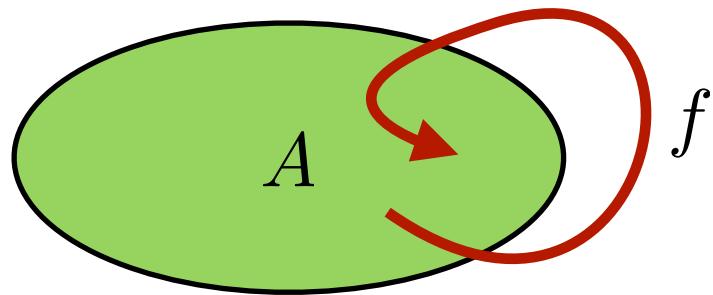
Unknown Target Mixture

■ Zero-sum game:

- NATURE: select a target distribution D_i .
 - LEARNER: select a z , i.e. a distribution weighted hypothesis h_z .
 - Payoff: $\mathcal{L}(D_i, h_z, f)$.
 - Already shown: game value is at most ϵ .
- ## ■ Minimax theorem (modulo discretization of z): there exists a mixture $\sum_j \alpha_j h_{z_j}$ of distribution weighted hypothesis that does well for any distribution mixture.

Balancing Losses

- Brouwer's Fixed Point theorem: for any compact, convex, non-empty set A and any continuous function $f: A \rightarrow A$, there exists x such that: $f(x) = x$.



Notation:

$$\mathcal{L}_i^z := \mathcal{L}(D_i, h_z, f).$$

- Define mapping ϕ by: $[\phi(z)]_i = \frac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z}$.
- By fixed point theorem (modulo continuity):

$$\exists z: \forall i, z_i = \frac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z} \implies \forall i, \mathcal{L}_i^z = \sum_j z_j \mathcal{L}_j^z =: \gamma.$$

Bounding Loss

- For fixed point z ,

$$\begin{aligned}\mathcal{L}(D_z, h_z, f) &= \sum_{x \in X} L(h_z(x), f(x)) \left(\sum_{i=1}^k z_i D_i(x) \right) \\ &= \sum_{i=1}^k z_i \sum_{x \in X} D_i(x) L(h_z(x), f(x)) \\ &= \sum_{i=1}^k z_i \mathcal{L}_i^z = \sum_{i=1}^k z_i \gamma = \gamma.\end{aligned}$$

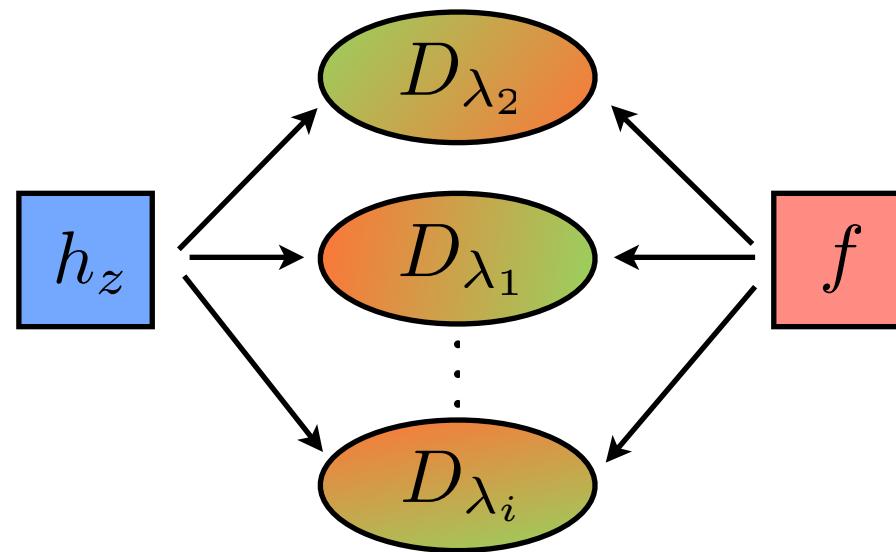
- Also, by convexity,

$$\gamma = \mathcal{L}(D_z, h_z, f) \leq \sum_{x \in X} \sum_{i=1}^k \frac{z_i D_i(x)}{D_z(x)} L(h_i(x), f(x)) D_z(x) = \sum_{i=1}^k z_i \mathcal{L}(D_i, h_i, f) \leq \epsilon.$$

Bounding Loss

- Thus, $\gamma \leq \epsilon$ and for any mixture λ ,

$$\mathcal{L}(D_\lambda, h_z, f) = \sum_{i=1}^k \lambda_i \mathcal{L}(D_i, h_z, f) \leq \sum_{i=1}^k \lambda_i \gamma = \gamma \leq \epsilon.$$



Details

- To deal with **non-continuity** refine hypotheses:

$$h_z^\eta(x) = \sum_{i=1}^k \frac{z_i D_i(x) + \eta/k}{\sum_{j=1}^k z_j D_j(x) + \eta} h_i(x).$$

- **Theorem:** for any target function f and any $\delta > 0$,

$$\exists \eta > 0, z: \forall \lambda, \mathcal{L}(D_\lambda, h_z^\eta, f) \leq \epsilon + \delta.$$

- If loss obeys triangle inequality:

$$\forall \delta > 0, \exists z, \eta > 0, \forall \lambda, f \in \mathcal{F}, \mathcal{L}(D_\lambda, h_z^\eta, f) \leq 3\epsilon + \delta.$$

holds for **all admissible target functions**.

A Simple Algorithm

- A simple **constructive algorithm**, choose z with uniform weights:

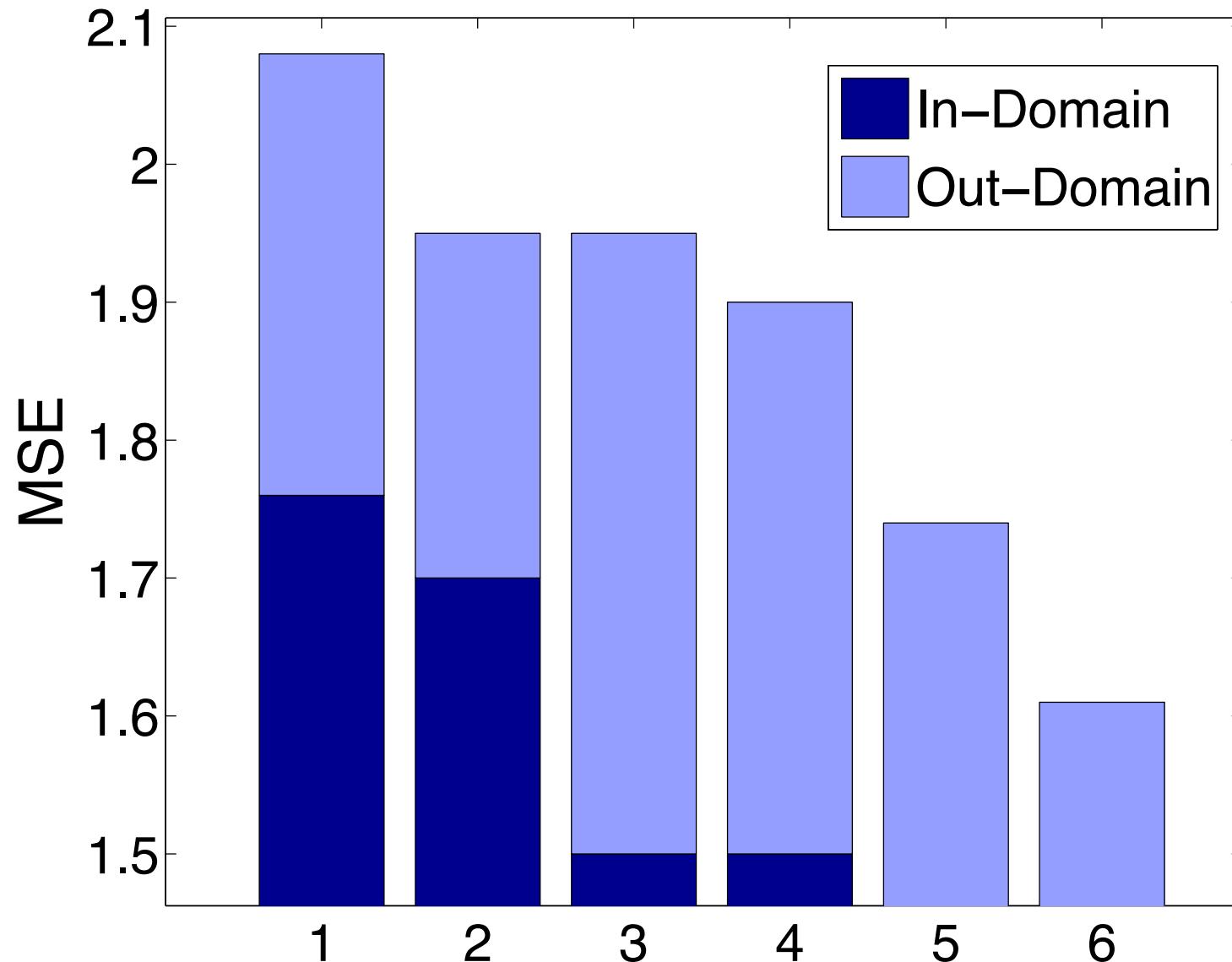
$$\begin{aligned}\mathcal{L}(D_\lambda, h_u, f) &= \sum_x D_\lambda(x) L \left(\sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x) \right) \\ &= \sum_x \left(\sum_{m=1}^k \lambda_m D_m(x) \right) L \left(\sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x) \right) \\ &\leq \sum_x \underbrace{\frac{\sum_{m=1}^k \lambda_m D_m(x)}{\sum_{j=1}^k D_j(x)}}_{\leq 1} \sum_{i=1}^k D_i(x) L(h_i(x), f(x)) \\ &\leq \sum_{i=1}^k \sum_x D_i(x) L(h_i(x), f(x)) = \sum_{i=1}^k \mathcal{L}(D_i, h_i, f) = \sum_{i=1}^k \epsilon_i \boxed{\leq k\epsilon.}\end{aligned}$$

Preliminary Empirical Results

- Sentiment Analysis - given a product review (text string), predict a rating (between 1.0 and 5.0).
- 4 Domains: Books, DVDs, Electronics and Kitchen Appliances.
- Base hypotheses are trained within each domain (Support Vector Regression).
- We are not given the distributions. We model each distribution using a bag of words model.
- We then test the distribution combination rule on known target mixture domains.

Empirical Results

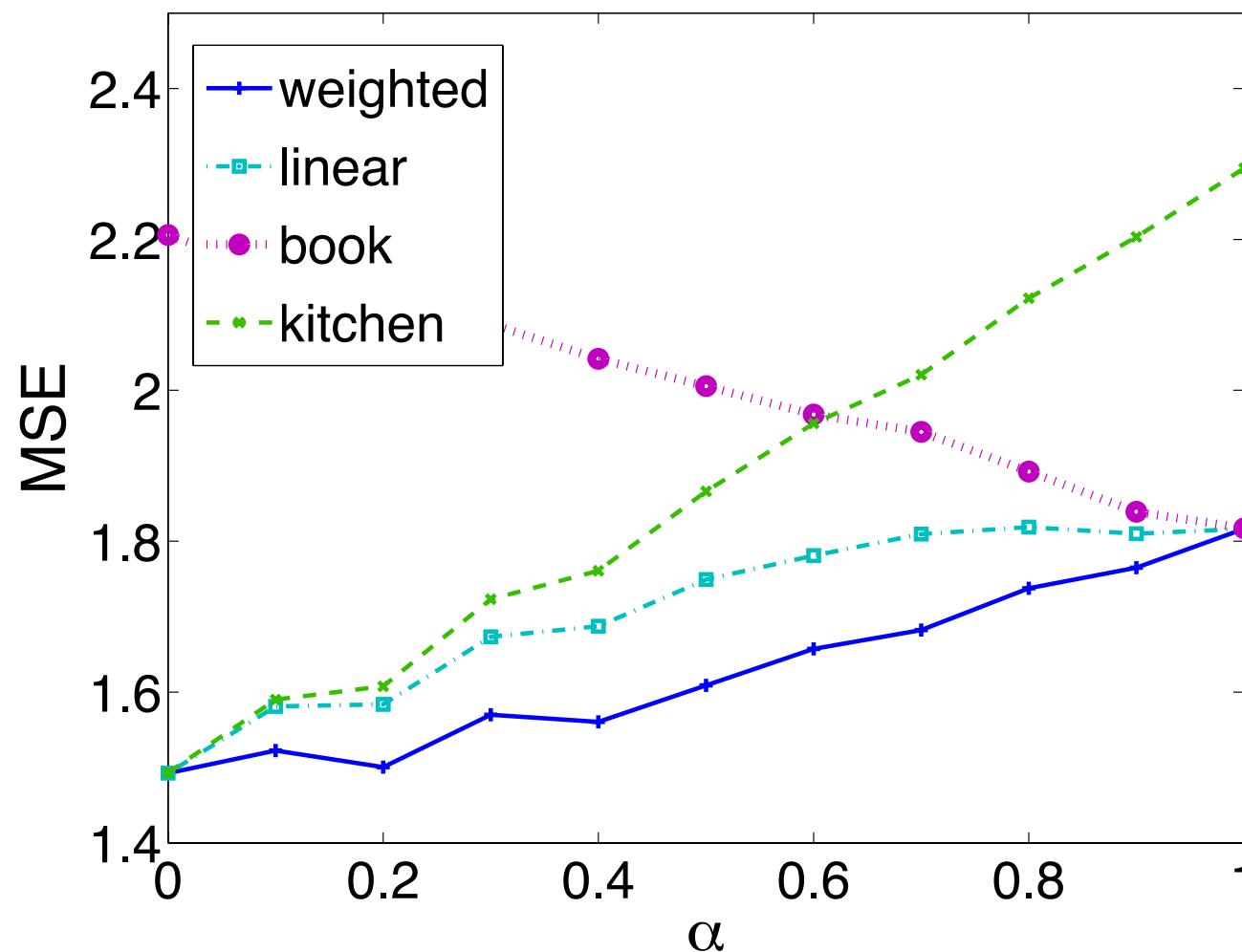
Uniform Mixture Over 4 Domains



Empirical Results

■ 2 class

$$\text{Mixture} = \alpha \text{ book} + (1 - \alpha) \text{ kitchen}$$



Conclusion

- Formulation of the multiple source adaptation problem.
- Theoretical analysis for mixture distributions.
- Efficient algorithm for finding distribution weighted combination hypothesis?
- Beyond mixture distributions?

Rényi Divergences

■ **Definition:** for $\alpha \geq 0$,

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_x P(x) \left[\frac{P(x)}{Q(x)} \right]^{\alpha-1}.$$

- $\alpha = 1$: coincides with relative entropy.
- $\alpha = 2$: logarithm of expected probability ratio;

$$D_\alpha(P\|Q) = \log \mathbb{E}_{x \sim P} \left[\frac{P(x)}{Q(x)} \right].$$

- $\alpha = +\infty$: logarithm of maximum probability ratio;

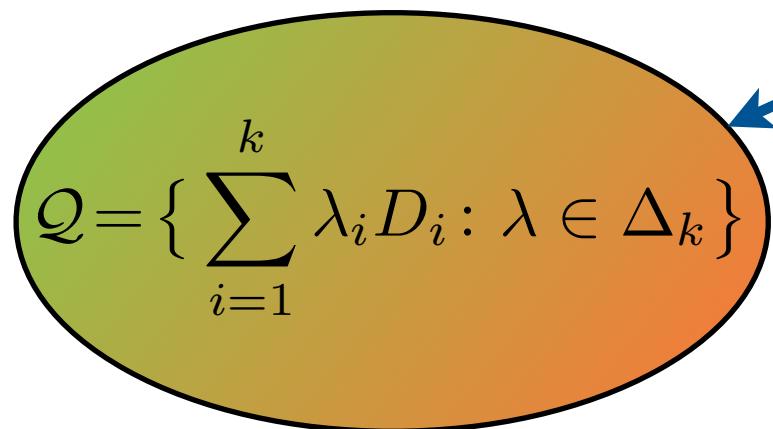
$$D_\alpha(P\|Q) = \log \sup_{x \sim P} \left[\frac{P(x)}{Q(x)} \right].$$

Extensions - Arbitrary Target

(Mansour, MM, and Rostami, 2009)

- **Theorem:** for any $\delta > 0$ and $\alpha > 1$,

$$\exists \eta, z: \forall P, \mathcal{L}(P, h_z^\eta, f) \leq [d_\alpha(P||Q)(\epsilon + \delta)]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.$$



P

measured in terms of Rényi divergence,

$$d_\alpha(P, Q) = \left[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \right]^{\frac{1}{\alpha-1}}.$$

Proof

■ By Hölder's inequality, for any hypothesis h ,

$$\begin{aligned}\mathcal{L}(P, h, f) &= \sum_x \frac{P(x)}{Q^{\frac{\alpha-1}{\alpha}}(x)} Q^{\frac{\alpha-1}{\alpha}}(x) L(h(x), f(x)) \\ &\leq \left[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \right]^{\frac{1}{\alpha}} \left[\sum_x Q(x) L^{\frac{\alpha}{\alpha-1}}(h(x), f(x)) \right]^{\frac{\alpha-1}{\alpha}} \\ &= (d_\alpha(P\|Q))^{\frac{\alpha-1}{\alpha}} \left[\underset{x \sim Q}{\text{E}} [L^{\frac{\alpha}{\alpha-1}}(h(x), f(x))] \right]^{\frac{\alpha-1}{\alpha}} \\ &= (d_\alpha(P\|Q))^{\frac{\alpha-1}{\alpha}} \left[\underset{x \sim Q}{\text{E}} [L(h(x), f(x)) L^{\frac{1}{\alpha-1}}(h(x), f(x))] \right]^{\frac{\alpha-1}{\alpha}} \\ &\leq (d_\alpha(P\|Q))^{\frac{\alpha-1}{\alpha}} \left[\mathcal{L}(Q, h, f) M^{\frac{1}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}}.\end{aligned}$$

Other Extensions

(Mansour, MM, and Rostami, 2009)

- Approximate distributions (estimated):
 - similar results shown depending on divergence between true and estimated distributions.
- Different source target functions f_i :
 - similar results when target functions close to f on target distribution.

References

- S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NIPS*. 2007.
- S. Ben-David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. *Journal of Machine Learning Research-Proceedings Track*, 9:129–136, 2010.
- S. Bickel, M. Bruckner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009.
- J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *NIPS*. 2008.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *NIPS*. 2010.
- Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519, 2014.

References

- Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In Proceedings of *ALT*. 2008.
- Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from Data of Variable Quality. In Proceedings of *NIPS*, 2006.
- Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. In Proceedings of *NIPS*, 2007.
- Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.
- M. Dredze, J. Blitzer, P. P. Talukdar, K. Ganchev, J. Graca, and F. Pereira. Frustratingly Hard Domain Adaptation for Parsing. In *CoNLL*, 2007.
- J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, volume 19, pages 601–608. 2006.

References

- Kifer D., Ben-David S., Gehrke J. Detecting change in data streams. In: Proceedings of VLDB, pp 180–191. 2004.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*. 2009.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In Advances in *NIPS*, pages 1041-1048. 2009.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the Rényi divergence. In Proceedings of *UAI*. 2009.
- Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In Proceedings of *ALT*, volume 7568, pages 124-138. 2012.

References

- H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227– 244, 2000.
- M. Sugiyama, S. Nakajima, H. Kashima, P. von Bunau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*. 2008.
- M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bunau, and M.Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Advanced Machine Learning

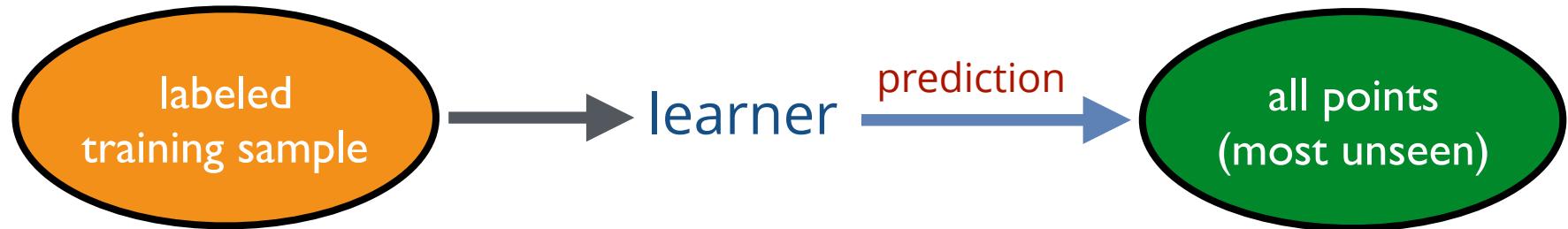
Transduction

MEHRYAR MOHRI MOHRI@

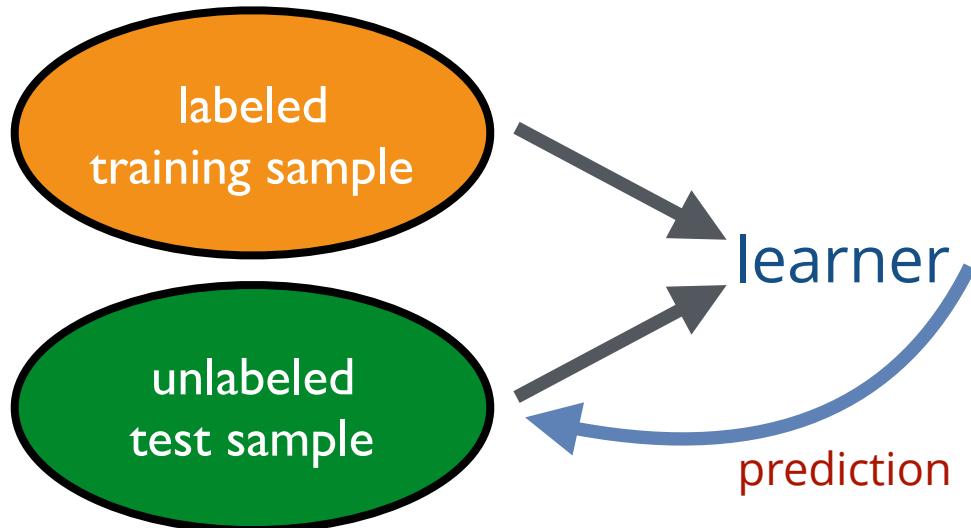
COURANT INSTITUTE & GOOGLE RESEARCH

Induction vs Transduction

- Inductive scenario:

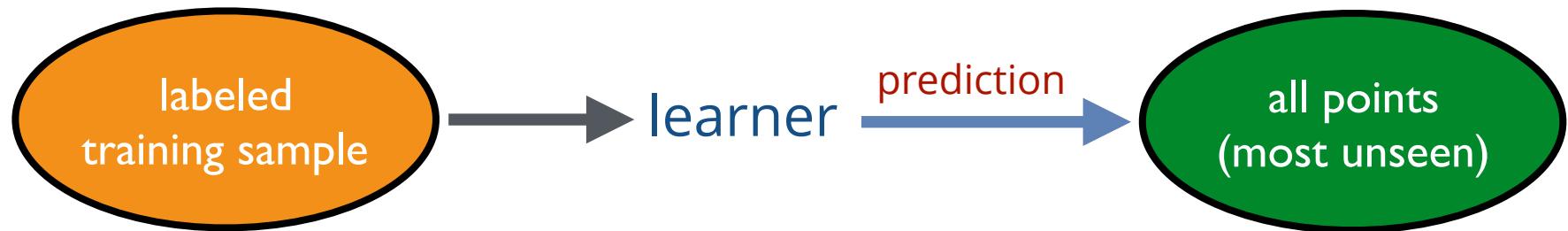


- Transductive scenario (Vapnik, 1998):

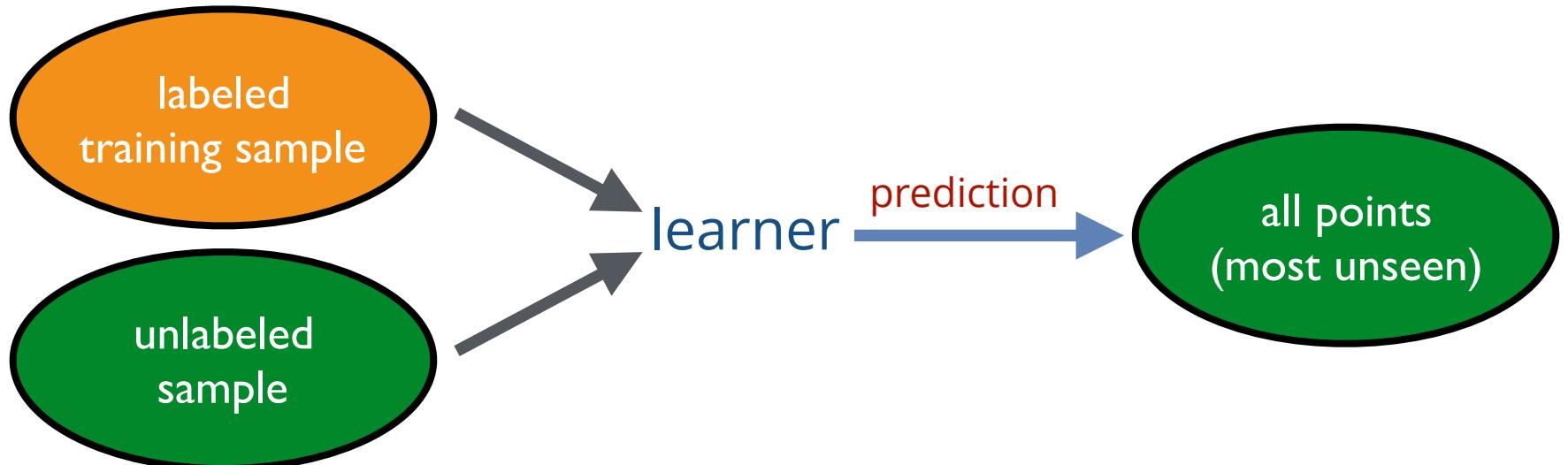


Semi-Supervised Learning

- Inductive scenario:



- Semi-supervised learning scenario:



Motivation

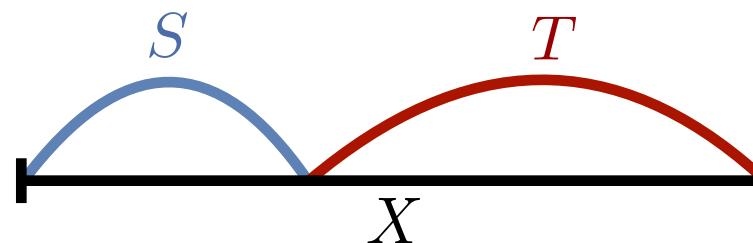
- Common scenario in many applications:
 - network predictions in computational biology.
 - web graph predictions.
 - NLP applications.
- Seemingly more favorable scenario than induction:
 - but can we (provably) benefit from that?
 - analysis of generalization in transductive setting.
 - transductive learning algorithms.

Outline

- Transduction scenario.
- Generalization bounds.
- Examples of algorithms.

Setting One

- A full sample X of size $(m + u)$ is fixed.
- The learner receives:
 - a sample $S = (x_1, \dots, x_m)$ drawn uniformly without replacement from X as well as the labels (y_1, \dots, y_m) .
 - an unlabeled test sample $T = (x_{m+1}, \dots, x_{m+u})$ formed by the remaining points of X .



Setting One

- Loss function L taking values in $[0, 1]$.
- Hypothesis set H .
- Errors: for a hypothesis $h \in H$,
 - training error: $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m L(h(x_i), y_i)$.
 - test error: $R_T(h) = \frac{1}{u} \sum_{i=1}^u L(h(x_{m+i}), y_{m+i})$.
 - full sample error (not a random variable):

$$R(h) = \frac{1}{m+u} \sum_{i=1}^{m+u} L(h(x_i), y_i) = \frac{1}{m+u} \left[m\hat{R}_S(h) + uR_T(h) \right].$$

Setting Two

- Distribution D over input space X .
- The learner receives:
 - a sample S of size m drawn i.i.d. from D^m as well as the corresponding labels.
 - a sample T of size u drawn i.i.d. from D^u .

Relationship btw Settings

- Any generalization bound for setting one implies a generalization bound for setting two by taking the expectation:

$$\mathbb{E}_{S \sim D^m, T \sim D^u} [1_{\{\sup_{h \in H} R_T(h) - \hat{R}_S(h) > \epsilon\}}] = \mathbb{E}_{X \sim D^{m+u}} \left[\mathbb{E}_{(S,T)=X} [1_{\{\sup_{h \in H} R_T(h) - \hat{R}_S(h) > \epsilon\}}] \right].$$

→ we will study generalization in setting one.

Outline

- Transduction scenario.
- Generalization bounds.
- Examples of algorithms.

Generalization Bounds

- VC-dimension bounds (Vapnik, 1998; Cortes and MM 2007).
- PAC-Bayesian bounds (Derbelko, El-Yaniv, and Meir, 2004).
- Stability bounds (El-Raniv and Pechyony 2008; Cortes, MM, Pechyony, Rastogi, 2008 and 2009).
- Rademacher complexity bounds (El-Raniv and Pechyony 2007).

McDiarmid's Inequality

(McDiarmid, 1989; corollary 6.10)

- **Theorem:** let X_1, \dots, X_m be random variables taking values in X and let $\Phi: X^m \rightarrow \mathbb{R}$ be a measurable function. Assume that there exist constants c_1, \dots, c_m such that

$$\left| \mathbb{E} [\Phi(X_1^m) | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x_i] - \mathbb{E} [\Phi(X_1^m) | X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x'_i] \right| \leq c_i,$$

for all $i \in [1, m]$ and $x_1^m, x'_1 \in X^m$. Then, for any $\epsilon > 0$,

$$\Pr[|\Phi(X_1^m) - \mathbb{E}[\Phi(X_1^m)]| > \epsilon] \leq 2 \exp \left[\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2} \right].$$

Sampling w/o Replacement

(Cortes, MM, Pechyony, Rastogi, 2008 & 2009)

- **Theorem:** let X_1, \dots, X_m be a sequence of r.v.'s distributed according to the uniform distribution without replacement from a set X of size $m + u$ and let $\Phi: X^m \rightarrow \mathbb{R}$ be a symmetric measurable function. Assume that there exists a constant c such that

$$|\Phi(x_1^m) - \Phi(x_1^{i-1}, x'_i, x_{i+1}^m)| \leq c$$

for all $i \in [1, m]$ and $x_1^m, x'_1 \in X^m$. Then, for any $\epsilon > 0$,

$$\Pr[|\Phi(X_1^m) - \mathbb{E}[\Phi(X_1^m)]| > \epsilon] \leq 2 \exp \left[\frac{-2\epsilon^2}{\alpha(m, u)c^2} \right],$$

with $\alpha(m, u) = \frac{mu}{m+u-1/2} \frac{1}{1-1/(2 \max\{m, u\})}$.

Proof

- For any $i \in [1, m]$,

$$\begin{aligned}
& \mathbb{E} [\Phi(X_1^m) | X_1^i = x_1^i] - \mathbb{E} [\Phi(X_1^m) | X_1^{i-1} = x_1^{i-1}, X_i = x'_i] \\
&= \sum_{x_{i+1}^m} \Pr[X_{i+1}^m = x_{i+1}^m | X_1^i = x_1^i] \Phi(x_1^{i-1}, x_i, x_{i+1}^m) \\
&\quad - \sum_{x_{i+1}'^m} \Pr[X_{i+1}^m = x_{i+1}'^m | X_1^{i-1} = x_1^{i-1}, X_i = x'_i] \Phi(x_1^{i-1}, x'_i, x_{i+1}'^m) \\
&= \left[\prod_{k=i}^{m-1} \frac{1}{m+u-k} \right] \left[\sum_{x_{i+1}^m} \Phi(x_1^{i-1}, x_i, x_{i+1}^m) - \sum_{x_{i+1}'^m} \Phi(x_1^{i-1}, x'_i, x_{i+1}'^m) \right] \\
&= \frac{u!}{(m+u-i)!} \left[\sum_{x_{i+1}^m} \Phi(x_1^{i-1}, x_i, x_{i+1}^m) - \sum_{x_{i+1}'^m} \Phi(x_1^{i-1}, x'_i, x_{i+1}'^m) \right].
\end{aligned}$$

Proof

■ Two cases:

- x'_{i+1}^m contains x_i : then there is (a unique) x_{i+1}^m such that $\{x'_i x'_{i+1}^m\} = \{x_i x_{i+1}^m\}$ and the corresponding terms cancel out by the symmetry of Φ .
- x'_{i+1}^m does not contain x_i : then there is (a unique) x_{i+1}^m such that $\{x'_i x'_{i+1}^m\}$ differs from $\{x_i x_{i+1}^m\}$ by $x_i \neq x'_i$. By assumption, the corresponding terms differ in absolute value by at most c .

■ Second case instances: number of x'_{i+1}^m permutations chosen out of the set $X - \{x_1^{i-1}, x_i, x'_i\}$:

$$\frac{(m+u-i-1)!}{(m+u-i-1-(m-i))!} = \frac{(m+u-i-1)!}{(u-1)!}.$$

Proof

■ Thus,

$$\begin{aligned} & \left| \mathbb{E} [\Phi(X_1^m) | X_1^i = x_1^i] - \mathbb{E} [\Phi(X_1^m) | X_1^{i-1} = x_1^{i-1}, X_i = x'_i] \right| \\ & \leq \frac{u!}{(m+u-i)!} \frac{(m+u-i-1)!}{(u-1)!} c = \frac{uc}{m+u-i}. \end{aligned}$$

■ The term in McDiarmid's inequality is bounded as follows:

$$\sum_{i=1}^m \frac{u^2 c^2}{(m+u-i)^2} = \sum_{j=u}^{m+u-1} \frac{u^2 c^2}{j^2} \leq \int_{u-1/2}^{m+u-1/2} \frac{u^2 c^2 dx}{x^2} = \frac{m u c^2}{m+u-1/2} \frac{u}{u-1/2}.$$

■ The theorem follows by observing that m and u can be permuted by the symmetry of Φ .

Generalization Bound

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$,

$$R_T(h) \leq \widehat{R}_S(h) + \mathbb{E}[\Phi(S)] + \sqrt{\frac{\eta}{2} \left[\frac{1}{m} + \frac{1}{u} \right] \log \frac{1}{\delta}},$$

where $\eta = \frac{m+u}{m+u-\frac{1}{2}} \frac{1}{1 - \frac{1}{2 \max\{m, u\}}}$.

- Proof: apply concentration bound to

$$\Phi(S) = \sup_{h \in H} R_T(h) - \widehat{R}_S(h).$$

- observe that

$$|\Phi(S') - \Phi(S)| \leq \frac{1}{m} + \frac{1}{u} = \frac{m+u}{mu}.$$

Rademacher Complexity

- Define random variable σ_i as taking value
 - $\frac{m+u}{u}$ with probability $\frac{u}{m+u}$.
 - $-\frac{m+u}{m}$ with probability $\frac{m}{m+u}$.
- **Definition:** the transductive Rademacher complexity of G is

$$\mathfrak{R}_{m+u}(G) = \frac{1}{m+u} \mathbf{E}_{\sigma} \left[\sup_{g \in G} \sum_{i=1}^{m+u} \sigma_i g(x_i) \right].$$

- note: simpler definition than [\(El-Yaniv and Pechyony 2007\)](#).

Analysis

- For any $N \in \left[-\frac{(m+u)^2}{m}, \frac{u(m+u)^2}{u} \right]$, define

$$R(N) = \frac{1}{m+u} \mathbb{E}_{\sigma} \left[\sup_{g \in G} \sum_{i=1}^{m+u} \sigma_i g(x_i) \middle| \sum_{i=1}^{m+u} \sigma_i = N \right].$$

- Observe that if $\sum_{i=1}^{m+u} \sigma_i = 0$ and $n \sigma_i$ s take value $\frac{m+u}{u}$, then

$$n \frac{m+u}{u} - (m+u-n) \frac{m+u}{m} = 0 \Leftrightarrow n = u.$$

Thus, $\mathbb{E}_S[\Phi(S)] = R(0)$.

Analysis

■ For any n

$$N = \sum_{i=1}^{m+u} \sigma_i = n \frac{m+u}{u} - (m+u-n) \frac{m+u}{m} = \frac{(m+u)^2}{mu} (n-u).$$

■ Let $n_2 \geq n_1$,

$$R(N_1) = \frac{1}{m+u} \mathbb{E} \left[\sup_{g \in G} \sum_{i=1}^{n_1} \frac{m+u}{u} g(x_i) - \sum_{i=n_1+1}^{m+u} \frac{m+u}{m} g(x_i) \right].$$

$$\begin{aligned} R(N_2) &= \frac{1}{m+u} \mathbb{E} \left[\sup_{g \in G} \sum_{i=1}^{n_1} \frac{m+u}{u} g(x_i) - \sum_{i=n_1+1}^{m+u} \frac{m+u}{m} g(x_i) \right. \\ &\quad \left. + \sum_{i=n_1+1}^{n_2} \left[\frac{m+u}{u} + \frac{m+u}{m} \right] g(x_i) \right]. \end{aligned}$$

Analysis

- Lipschitz property:

$$|R(N_2) - R(N_1)| \leq |n_2 - n_1| \left(\frac{1}{m} + \frac{1}{u} \right) = \frac{|N_2 - N_1|}{m + u}.$$

- Thus, for $N = \sum_{i=1}^{m+u} \sigma_i$,

$$\Pr \left[|R(N) - R(\mathbb{E}[N])| > \epsilon \right] \leq \Pr \left[|N - \mathbb{E}[N]| > (m + u)\epsilon \right].$$

Transductive Rad. Comp. Bound

- **Theorem:** let H_L denote $\{x \mapsto L(h(x), f(x)) : h \in H\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$,

$$R_T(h) \leq \hat{R}_S(h) + \mathfrak{R}_{m+u}(H_L) + O\left(\sqrt{\min\{m, u\}} \left[\frac{1}{m} + \frac{1}{u}\right]\right) + \sqrt{\frac{\eta}{2} \left[\frac{1}{m} + \frac{1}{u}\right] \log \frac{1}{\delta}},$$

where $\eta = \frac{m+u}{m+u-\frac{1}{2}} \frac{1}{1 - \frac{1}{2 \max\{m, u\}}}$.

Notes

- For large m , the bound varies only as $O(\frac{1}{\sqrt{u}})$: quite different from the induction scenario.
- H can be selected after measuring $\mathfrak{R}_{m+u}(H_L)$ since the full sample is accessible.

Transductive Stability Bound

- **Theorem:** let L be a loss function taking values in $[0, 1]$ and let \mathcal{A} be a uniformly β -stable algorithm returning $h_S \in H$ when trained using labeled sample S . Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$R_T(h_S) \leq \hat{R}_S(h_S) + \beta + \left(2\beta + \frac{(m+u)}{mu}\right) \sqrt{\frac{\alpha(m,u) \log \frac{1}{\delta}}{2}}.$$

Proof

- Define for any $h \in H$, $\Phi(S, h) = R_T(h) - \hat{R}_S(h)$.
- Assume that S and S' differ by one point. Then,

$$\Phi(S', h_{S'}) - \Phi(S, h_S)$$

$$\begin{aligned} &= \frac{1}{u} \sum_{i=1}^{u-1} L(h_{S'}(x_{m+i}), y_{m+i}) - L(h_S(x_{m+i}), y_{m+i}) \\ &\quad + \frac{1}{m} \sum_{i=1}^{m-1} L(h_{S'}(x_i), y_i) - L(h_S(x_i), y_i) \\ &\quad + \frac{1}{u} (L(h_{S'}(x'_{m+i}), y'_{m+i}) - L(h_S(x_{m+i}), y_{m+i})) \\ &\quad + \frac{1}{m} (L(h_{S'}(x'_m), y'_m) - L(h_S(x_m), y_m)). \end{aligned}$$

Proof

■ Thus,

$$\left| \Phi(S', h_{S'}) - \Phi(S, h_S) \right| \leq \frac{\beta(u-1)}{u} + \frac{\beta(m-1)}{m} + \frac{1}{u} + \frac{1}{m} \leq 2\beta + \frac{1}{u} + \frac{1}{m}.$$

■ Bounding the expectation:

$$\begin{aligned} \underset{S}{\mathbb{E}}[\Phi(S, h_S)] &= \frac{1}{u} \sum_{i=1}^u \underset{S}{\mathbb{E}}[L(h_S(x_{m+i}), y_{m+i})] - \frac{1}{m} \sum_{i=1}^m \underset{S}{\mathbb{E}}[L(h_S(x_i), y_i)] \\ &= \underset{S, x' \notin S}{\mathbb{E}}[L(h_S(x'), y_{x'})] - \underset{S, x \in S}{\mathbb{E}}[L(h_S(x), y_x)] \\ &= \underset{S, x' \notin S}{\mathbb{E}}[L(h_{S-\{x\} \cup \{x'\}}(x), y_x) - L(h_S(x), y_x)] \leq \beta. \end{aligned}$$

Outline

- Transduction scenario.
- Generalization bounds.
- Examples of algorithms.

Transductive SVM (TSVM)

(Vapnik, 2008), see also (Joachims, 1999)

■ Optimization problem:

$$\min_{\mathbf{w}, b, \mathbf{y}_{m+1}^{m+u}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m L(\mathbf{w} \cdot \mathbf{x}_i + b, y_i) + C' \sum_{i=1}^u L(\mathbf{w} \cdot \mathbf{x}_{m+i} + b, y_{m+i})$$

- classification: hinge loss.
- regression: trivial solution, last term vanishes! (Cortes and MM, 2007).
- theoretical guarantee: unclear.
- computational complexity: exponential.
- experiments: issue of uniform labeling of test points in high dimension (Joachims, 1999); poor results (Tong and Oles, 1999).

Local Transductive Regression

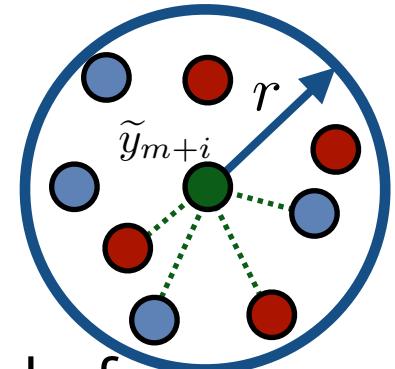
(Cortes, MM, Pechyony, Rastogi, 2008 & 2009)

■ Optimization problem (LTR):

$$\min_{h \in \mathbb{H}} \|h\|_K^2 + C \sum_{i=1}^m L(h(x_i), y_i) + C' \sum_{i=1}^u L(h(x_{m+i}), \tilde{y}_{m+i}),$$

with K a PDS kernel, and

\tilde{y}_{m+i} 's pseudo-labels obtained via local weighted average or any other local regression algorithm from neighborhood of radius r .



Stability Guarantee

- **Theorem:** assume that for all $x \in X$, $|y(x)| \leq M$ and that the local estimator has score-stability β_{loc} . Then, LTR has uniform stability

$$\beta \leq 2(C_0M)^2r^2 \left[\frac{C}{m} + \frac{C'}{u} + \sqrt{\left(\frac{C}{m} + \frac{C'}{u} \right)^2 + \frac{2C'\beta_{\text{loc}}}{C_0Mr^2u}} \right],$$

with $r^2 = \sup_{x \in X} K(x, x)$ and $C_0 = 1 + r\sqrt{C + C'}$.

Graph Regularization Algo.

■ Set-up:

- weighted directed graph $G = (X, E)$.
- hypothesis $h: X \rightarrow \mathbb{R}$ in H identified with vector

$$\mathbf{h} = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_{m+u}) \end{bmatrix}.$$

- PSD matrix $\mathbf{L} \in \mathbb{R}^{(m+u) \times (m+u)}$ (similarity matrix).

Graph Regularization Algo.

■ Optimization problem:

$$\min_{\mathbf{h} \in H} \mathbf{h}^\top \mathbf{L} \mathbf{h} + \frac{C}{m} (\mathbf{h}_S - \mathbf{y}_S)^\top (\mathbf{h}_S - \mathbf{y}_S)$$
$$\text{s.t.: } \mathbf{h}^\top \mathbf{u} = 0,$$

where \mathbf{h}_S is the restriction of \mathbf{h} to the training sample S and \mathbf{y}_S the vector of training labels, and \mathbf{u} a constant vector in \mathbb{R}^{m+u} .

Graph Regularization Algo.

■ Example (Belkin et al., 2004):

- graph assumed connected.
- \mathbf{L} is the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where

$$\mathbf{D} = \text{diag} \left(\sum_{i=1}^n w_{1i}, \dots, \sum_{i=1}^n w_{ni} \right).$$

Then, $\mathbf{h}^\top \mathbf{L} \mathbf{h} = \sum_{i \sim j} w_{ij} (h(x_i) - h(x_j))^2$.

- $\mathbf{u} = (1, \dots, 1)^\top$.
- data assumed centered: $\mathbf{u}^\top \mathbf{y} = 0$, and graph connected.
- → zero eigenvalue of Laplacian has multiplicity one and the solutions \mathbf{h} in $\text{range}(\mathbf{L})$.

Graph Regularization Algo.

- Lagrangian:

$$\mathcal{L} = \mathbf{h}^\top \mathbf{L} \mathbf{h} + \frac{C}{m} (\mathbf{h}_S - \mathbf{y}_S)^\top (\mathbf{h}_S - \mathbf{y}_S) + \beta \mathbf{h}^\top \mathbf{u}.$$

- Differentiating and applying orthogonal projection to \mathbf{u} :

$$\mathbf{P} \left(\mathbf{L} + \frac{C}{m} \mathbf{I}_S \right) \mathbf{h} = \frac{C}{m} \mathbf{P} \mathbf{y}_S - \beta \mathbf{P} \mathbf{u} = \frac{C}{m} \mathbf{P} \mathbf{y}_S$$

$$\Rightarrow \mathbf{h} = \left[\mathbf{P} \left(\frac{m}{C} \mathbf{L} + \mathbf{I}_S \right) \right]^{-1} \mathbf{P} \mathbf{y}_S. \quad \left(\mathbf{P} \left(\frac{m}{C} \mathbf{L} + \mathbf{I}_S \right) \text{ invertible} \right)$$

Stability Guarantee

(Cortes, MM, Pechyony, Rastogi, 2009)

- **Theorem:** assume that for all $h \in H$ and $x \in X$, $|h(x) - y_x| \leq M$. Then, the graph Laplacian regularization algorithm has uniform stability

$$\beta \leq \frac{4CM^2}{m} \min \left\{ \frac{1}{\lambda_2}, \rho_G \right\},$$

where λ_2 is the smallest non-trivial eigenvalue of \mathbf{L} and ρ_G the diameter of the graph (longest shortest path).

Proof

- The graph Laplacian algorithm can be shown to coincide with LTR with the kernel matrix $\mathbf{K} = \mathbf{L}^+$: for all $\mathbf{h} \in \text{range}(\mathbf{L})$,

$$\mathbf{K}\mathbf{L}\mathbf{h} = \mathbf{L}^+\mathbf{L}\mathbf{h} = \mathbf{h}$$

$$\mathbf{h}'^\top \mathbf{L} \mathbf{K} \mathbf{L} \mathbf{h} = \mathbf{h}'^\top \mathbf{L} \mathbf{h}.$$

- The result follows by applying the stability bound for LTR with the bound on the $K(x, x)$ in terms of λ_2 and ρ_G .

Notes

- For a hypercube, $\lambda_2 = 2$.
- Does not perform well in experiments in comparison with LTR.

References

- Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In COLT, pages 624–638. Springer, 2004.
- Olivier Chapelle, Vladimir Vapnik, and Jason Weston. Transductive inference for estimating values of functions. In NIPS, pages 421–427. 1999.
- Corinna Cortes and Mehryar Mohri. On transductive regression. In NIPS, pages 305–312. 2007.
- Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability of transductive regression algorithms. In ICML. 2008.
- Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability analysis and learning bounds for transductive regression algorithms. ArXiv 0904.0814. 2009.
- Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. J. Artif. Intell. Res. (JAIR), 22:117–142, 2004.

References

- Thorsten Joachims. Transductive Inference for Text Classification using Support Vector Machines. ICML 1999: 200-209.
- Colin McDiarmid. On the method of bounded differences. In Surveys in Combinatorics, pages 148–188. Cambridge University Press, Cambridge, 1989.
- Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. In COLT, 2007.
- Vladimir N. Vapnik. Statistical Learning Theory. Wiley-Interscience, New York, 1998.
- Zhang, Tong and Frank Oles. A probability analysis on the value of unlabeled data for classification problems. In ICML, pp. 1191–1198, 2000.

Advanced Machine Learning

Active Learning

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Active Learning Setup

■ Passive learning:

- IID sample $((x_1, y_1), \dots, (x_m, y_m)) \sim D^m$ is drawn.
- learner receives full labeled sample.

■ Active learning:

- IID sample $((x_1, y_1), \dots, (x_m, y_m)) \sim D^m$ is drawn.
- learner has access to (x_1, \dots, x_m) .
- learner can request the label y_i of point x_i .
- objective: fewer label requests than in passive learning.

Key Active Learning Problem

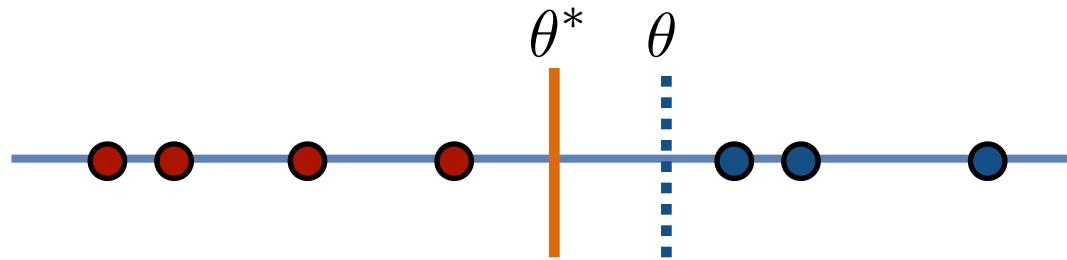
■ Tension:

- requesting label of new point to gain more information.
- sample bias induced by the label queries.

Favorable Example

■ Binary classification problem in \mathbb{R} :

- H : threshold functions.
- data assumed separable.



■ Sample complexity for determining θ^* within ϵ :

- supervised learner needs $O(\frac{1}{\epsilon})$ samples since at least one point is needed in $[\theta^* - \epsilon, \theta^* + \epsilon]$.
 - active learner needs only $O(\log \frac{1}{\epsilon})$ using binary search.
- exponential improvement!

Negative Result

(Kääriäinen, 2006)

■ Non-realizable case:

- stochastic or deterministic labels.
- if Bayes error is $\beta > 0$, the sample complexity of any active learning algorithm is at least

$$\Omega\left(\frac{\beta^2}{\epsilon^2}\right).$$

- thus, lower bound matches passive learning upper bound $O\left(\frac{1}{\epsilon^2}\right)$.

CAL Algorithm

(Cohn, Atlas, and Ladner, 1994)

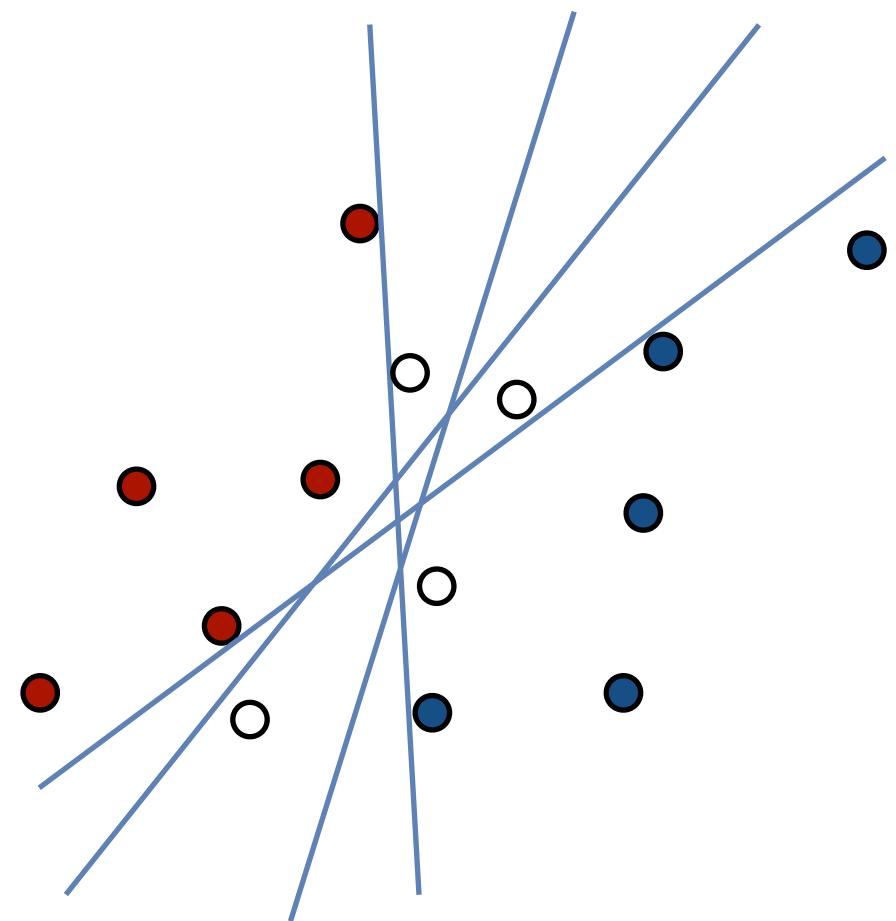
- Assume realizable case with hypothesis set H .

$\text{CAL}(H)$

```
1   $H_1 \leftarrow H$ 
2  for  $t \leftarrow 1$  to  $T$  do
3      if  $(\exists h, h' \in H_t : h(x_t) \neq h'(x_t))$  then
4           $y_t \leftarrow \text{QUERYLABEL}(x_t)$ 
5           $H_{t+1} \leftarrow \{h \in H_t : h(x_t) = y_t\}$ 
6      else  $H_{t+1} \leftarrow H_t$ 
7  return  $H_{T+1}$ 
```

CAL Algorithm

- Simple algorithm, but:
 - Computational cost of maintaining H_tS .
 - Separability requirement.



Definitions

(Hanneke, 2009)

- Region of disagreement:

$$\text{DIS}(H) = \{x \in X \mid \exists h, h' \in H: h(x) \neq h'(x)\}.$$

- Disagreement metric:

$$d(h, h') = \Pr_{x \sim D} [h(x) \neq h'(x)].$$

- Disagreement ball:

$$B(h, r) = \left\{ h' \in H : d(h, h') \leq r \right\}.$$

- Disagreement coefficient (rate of disagreement decrease):

$$\theta = \limsup_{r \rightarrow 0} \frac{\Pr \left(\text{DIS}(B(h^*, r)) \right)}{r}.$$

Disagreement Coefficient

(Hanneke, 2009)

- Property: for all $r > 0$, $\text{DIS}(B(h^*, r)) \leq \theta r$.
- Examples:
 - threshold functions: $\theta \leq 2$.
 - let $t \in B(t^*, r)$, then $t \in [t^* - \epsilon, t^* + \epsilon']$ where
$$\epsilon = \underset{\epsilon > 0}{\text{argmax}} \{ \Pr([t^* - \epsilon, t^*]) \leq r \} \quad \epsilon' = \underset{\epsilon > 0}{\text{argmax}} \{ \Pr([t^*, t^* + \epsilon]) \leq r \}.$$
 - thus, $\text{DIS}(B(h^*, r)) \leq 2r$.
 - finite hypothesis sets: $\theta \leq |H|$.
 - linear separators going through the origin and uniform distribution: $\theta \leq \pi\sqrt{N}$.

CAL Guarantees

- **Theorem:** let H be a hypothesis set with $\text{VCdim}(H) = d$ and assume that the data is separable with disagreement coefficient θ . Then, the label complexity of CAL is bounded by

$$\tilde{O}\left(\theta d \log \frac{1}{\epsilon}\right).$$

DHM Algorithm

(Dasgupta, Hsu, and Monteleoni, 2007)

- $\mathcal{A}(S, T)$ returns hypothesis in H consistent with S with minimum error on T when it exists, `NIL` otherwise.

$\text{DHM}((x_1, \dots, x_T))$

```
1   $S \leftarrow \emptyset$      $\triangleleft$  labels inferred
2   $T \leftarrow \emptyset$      $\triangleleft$  labels queried
3  for  $t \leftarrow 1$  to  $T$  do
4       $h_+ \leftarrow \mathcal{A}(S \cup (x_t, +1), T)$ 
5       $h_- \leftarrow \mathcal{A}(S \cup (x_t, -1), T)$ 
6      if ( $h_+ = \text{NIL}$ ) then
7           $S \leftarrow S \cup \{(x_t, -1)\}$ 
8      elseif ( $h_- = \text{NIL}$ ) then
9           $S \leftarrow S \cup \{(x_t, +1)\}$ 
10     elseif  $\hat{R}_{S \cup T}(h_+) - \hat{R}_{S \cup T}(h_-) > \Delta_t$  then
11          $S \leftarrow S \cup \{(x_t, -1)\}$ 
12     elseif  $\hat{R}_{S \cup T}(h_-) - \hat{R}_{S \cup T}(h_+) > \Delta_t$  then
13          $S \leftarrow S \cup \{(x_t, +1)\}$ 
14     else  $y_t \leftarrow \text{QUERYLABEL}(x_t)$ 
15          $T \leftarrow T \cup \{(x_t, y_t)\}$ 
16 return  $H_{T+1}$ 
```

Notes

- $S \cup T$ not an i.i.d. labeled sample drawn according to D .
- Δ_t is defined by $\Delta_t = \beta_t^2 + \beta_t \left(\sqrt{\hat{R}_t(h_+)} + \sqrt{\hat{R}_t(h_-)} \right)$,

with $\beta_t = 2\sqrt{\frac{\log((8t^2 + t)\Pi_{2t}^2(H)) + \log \frac{1}{\delta}}{t}} = \tilde{O}\left(\sqrt{\frac{d \log t}{t}}\right)$.

DHM Guarantees

- **Theorem:** let H be a hypothesis set with $\text{VCdim}(H) = d$ and disagreement coefficient θ . Then, the label complexity of DHM is bounded by

$$\tilde{O}\left(\theta\left(d \log^2 \frac{1}{\epsilon} + \frac{d\nu^2}{\epsilon^2}\right)\right),$$

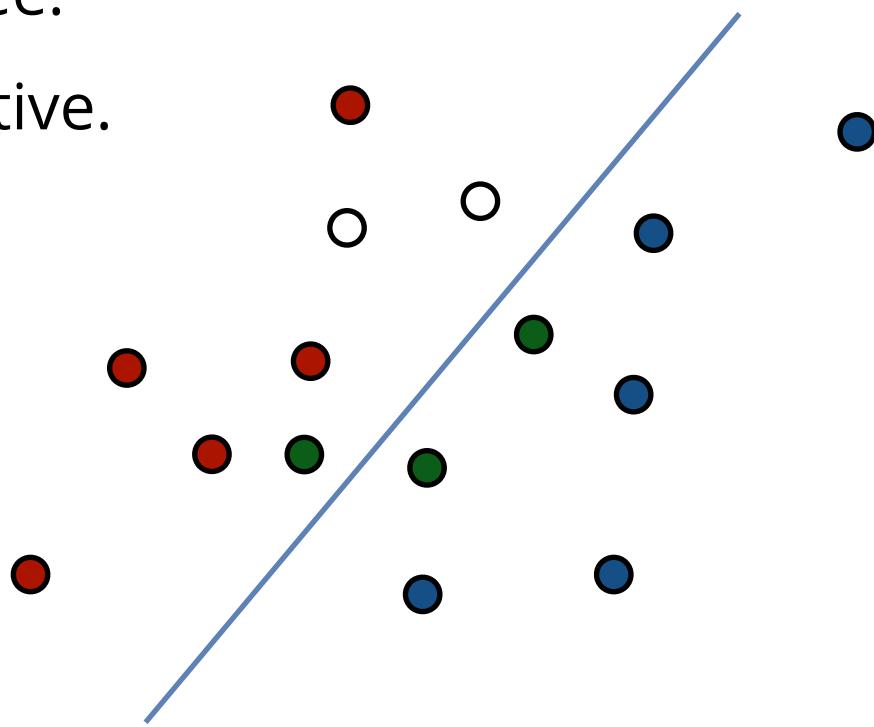
where $\nu = R(h^*)$.

Heuristics

(see for example (Tong and Koller, 2002))

Idea:

- select points close to the decision surface.
- poor theory: no guarantee.
- experiments: often effective.



Recent Algorithms

- ‘Margin-based active learning’ (Balcan, Broder, and Zhang, 2007; Balcan and Long, 2013; Awasthi, Balcan, and Long, 2014): improvement over disagreement-based for
 - uniformly distributed linear classifiers.
 - log-concave distributions.
- Confidence-rated predictors (Zhang and K. Chaudhuri, 2014):
 - better sample complexity than disagreement-based ones (term better than dis. coeff.).
 - more general than margin-based techniques.
 - however, computationally inefficient.

Empirical Results

(Guyon, Cawley, Dror and Lemaire, 2011)

- Active learning challenge (2011):
 - algorithms allowed to query labels with a budget.
 - performance measured in terms of AUC.
 - disappointing results compared to baseline passive learning algorithms.

References

- P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In STOC, 2014.
- M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In ICML, 2006.
- M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In COLT, 2007.
- M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In COLT, 2013.
- A. Beygelzimer, S. Dasgupta, and J. Langford (2009) Importance-weighted active learning, Twenty-Sixth International Conference on Machine Learning (ICML).

References

- D. Cohn, L. Atlas, and R. Ladner (1994) Improving generalization with active learning, *Machine Learning*, 15, 201–221.
- S. Dasgupta (2011) Two faces of active learning, *Theoretical Computer Science*, 412, 1767–1781.
- S. Dasgupta and D.J. Hsu. Hierarchical sampling for active learning, in ICML, 2008.
- S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In NIPS, 2007.
- S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning, *Journal of Machine Learning Research*, 10, 281–299, 2009.

References

- I. Guyon, G. Cawley, G. Dror and V. Lemaire: Results of the Active Learning Challenge. Active Learning and Experimental Design at AISTATS, 19-45. 2011.
- S. Hanneke. Theoretical Foundations of Active Learning, Ph.D. Thesis, CMU Machine Learning Department, 2009.
- M. Kääriäinen. Active learning in the non-realizable case. In COLT, 2006.
- V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. Journal of Machine Learning, 11:2457–2485, 2010.
- S. Tong and D. Koller. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2:45–66, 2002.
- C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In NIPS, 2014.

Advanced Machine Learning

Time Series Prediction

VITALY KUZNETSOV

KUZNETSOV@

GOOGLE RESEARCH

MEHRYAR MOHRI

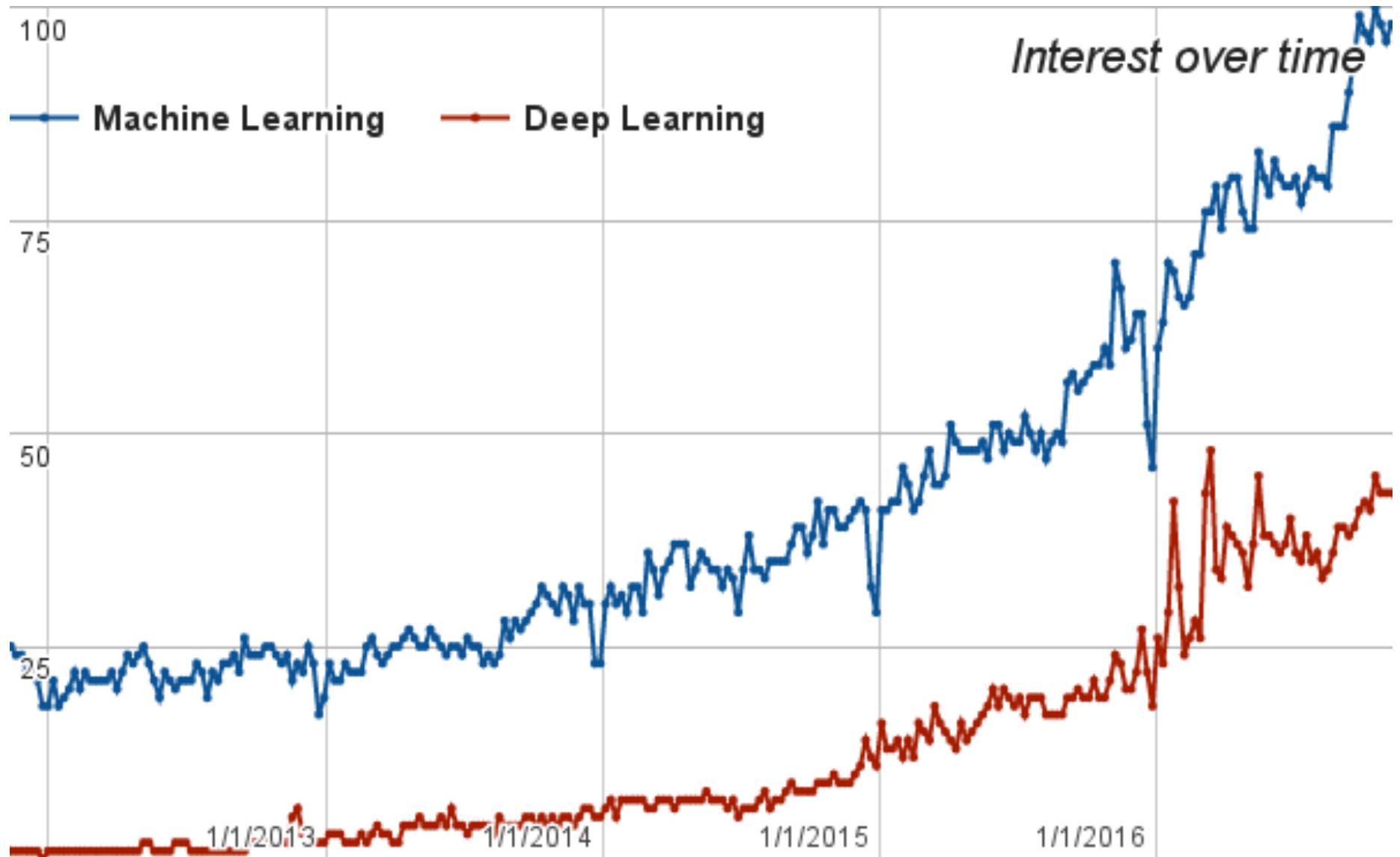
MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

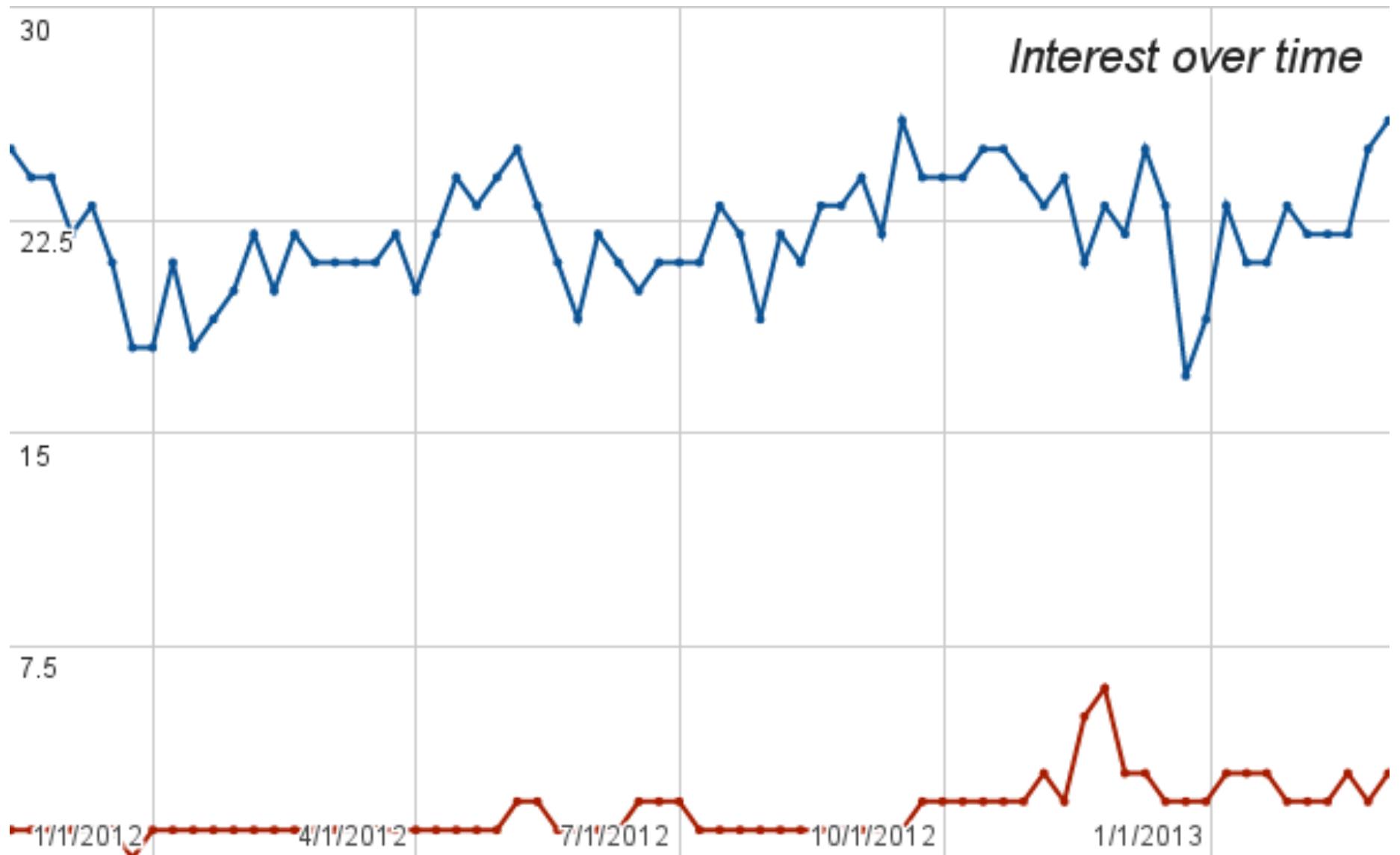
Motivation

- Time series prediction:
 - stock values.
 - economic variables.
 - weather: e.g., local and global temperature.
 - sensors: Internet-of-Things.
 - earthquakes.
 - energy demand.
 - signal processing.
 - sales forecasting.

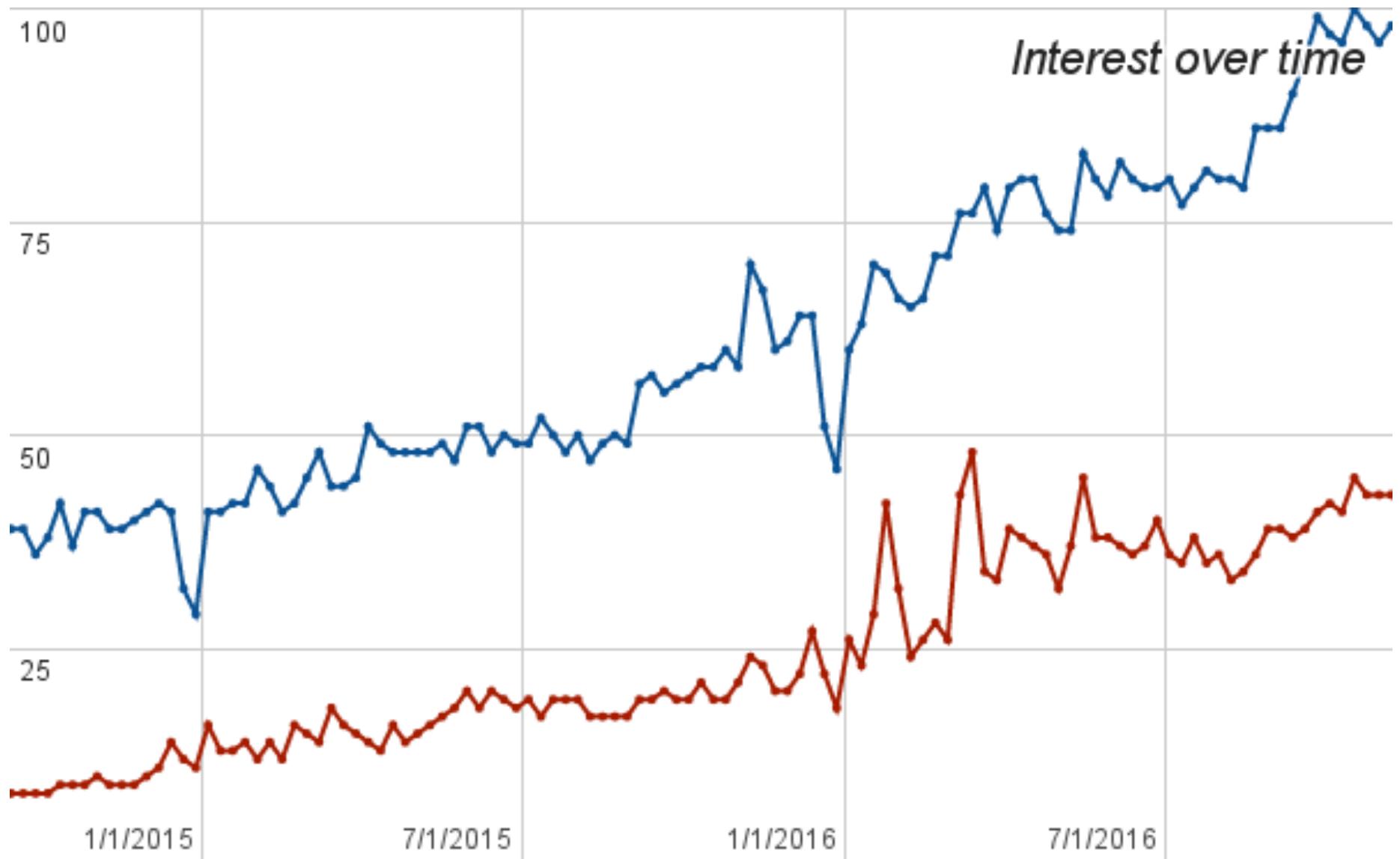
Google Trends



Google Trends



Google Trends



Challenges

■ Standard Supervised Learning:

- IID assumption.
- Same distribution for training and test data.
- Distributions fixed over time (stationarity).

→ none of these assumptions holds
for time series!

Outline

- Introduction to time series analysis.
- Learning theory for forecasting non-stationary time series.
- Algorithms for forecasting non-stationary time series.
- Time series prediction and on-line learning.

Why not Online-to-Batch?

Answer: because we are not in the i.i.d. setting

Introduction to Time Series Analysis

Classical Framework

- Postulate a particular form of a parametric model that is assumed to generate data.
- Use given sample to estimate unknown parameters of the model.
- Use estimated model to make predictions.

Autoregressive (AR) Models

- **Definition:** AR(p) model is a linear generative model based on the p th order Markov assumption:

$$\forall t, Y_t = \sum_{i=1}^p a_i Y_{t-i} + \epsilon_t$$

where

- ϵ_t s are zero mean uncorrelated random variables with variance σ .
- a_1, \dots, a_p are autoregressive coefficients.
- Y_t is observed stochastic process.

Moving Averages (MA)

- **Definition:** MA(q) model is a linear generative model for the noise term based on the q th order Markov assumption:

$$\forall t, Y_t = \epsilon_t + \sum_{j=1}^q b_j \epsilon_{t-j}$$

where

- b_1, \dots, b_q are moving average coefficients.

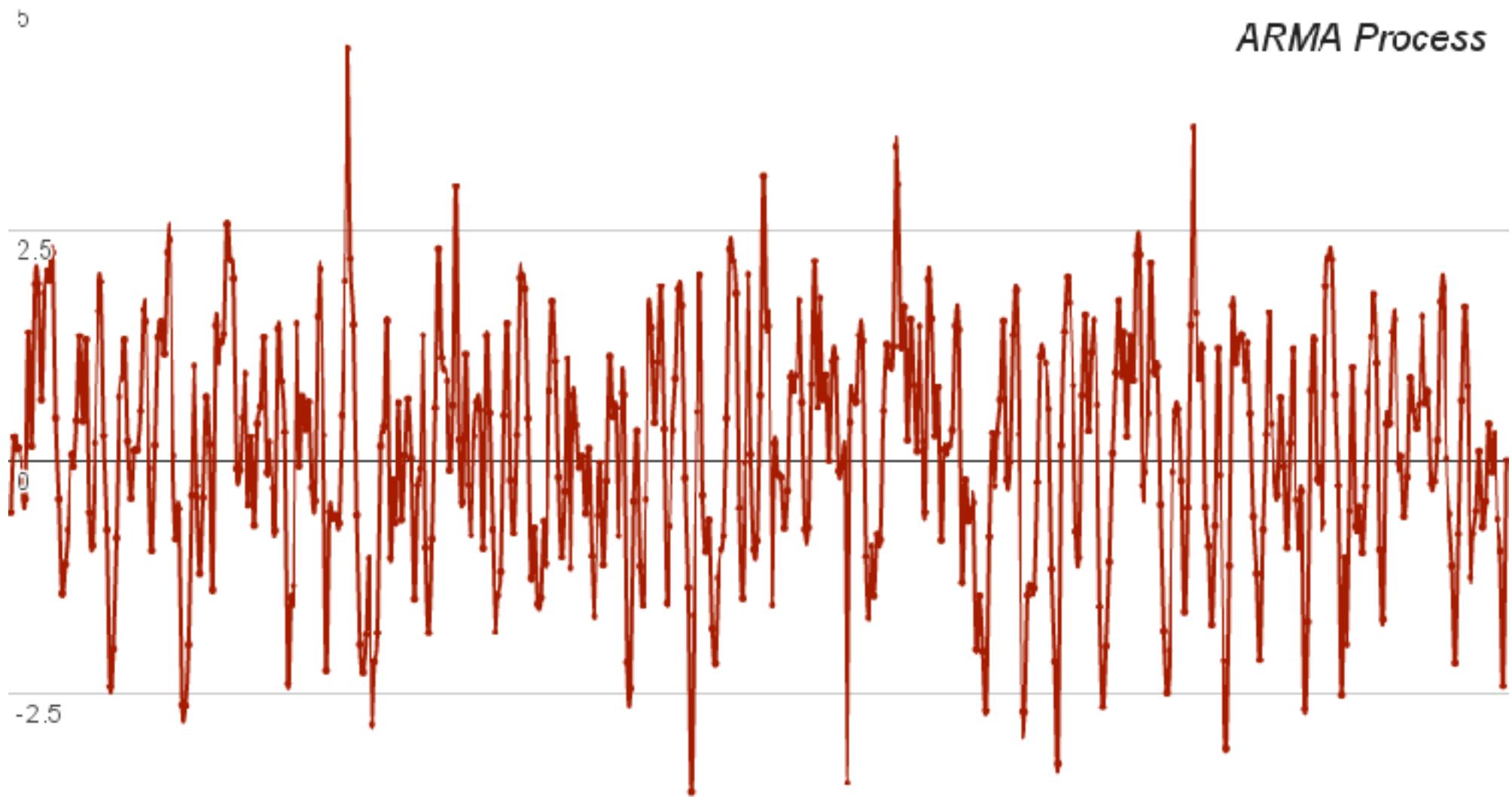
ARMA model

(Whittle, 1951; Box & Jenkins, 1971)

- **Definition:** ARMA(p, q) model is a generative linear model that combines AR(p) and MA(q) models:

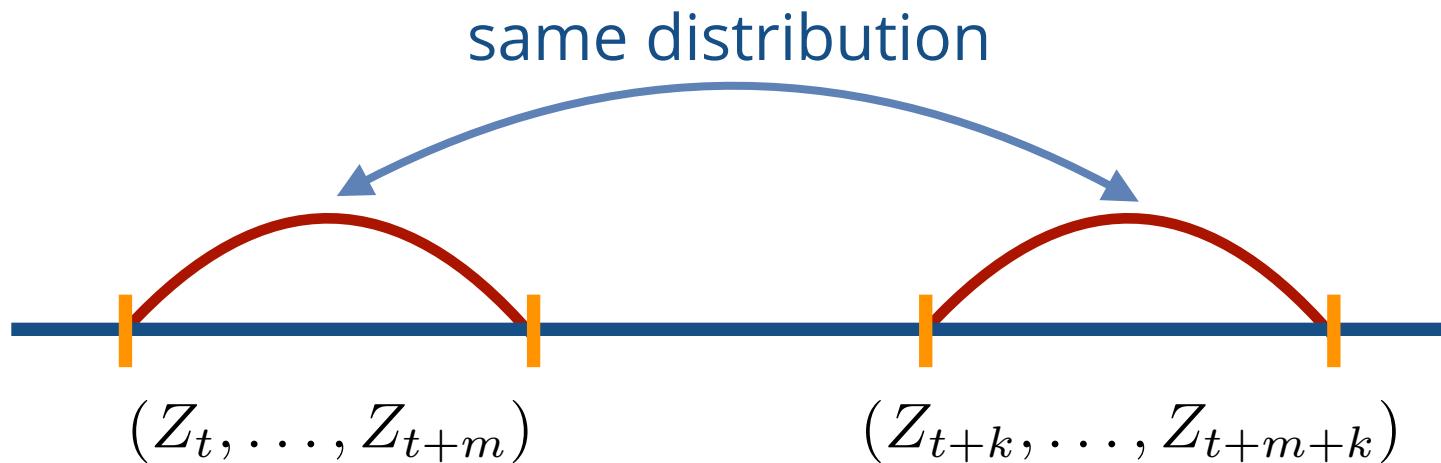
$$\forall t, Y_t = \sum_{i=1}^p a_i Y_{t-i} + \epsilon_t + \sum_{j=1}^q b_j \epsilon_{t-j}.$$

ARMA



Stationarity

- **Definition:** a sequence of random variables $\mathbf{Z} = \{Z_t\}_{-\infty}^{+\infty}$ is stationary if its distribution is invariant to shifting in time.



Weak Stationarity

- **Definition:** a sequence of random variables $\mathbf{Z} = \{Z_t\}_{-\infty}^{+\infty}$ is **weakly stationary** if its first and second moments are invariant to shifting in time, that is,
 - $\mathbb{E}[Z_t]$ is independent of t .
 - $\mathbb{E}[Z_t Z_{t-j}] = f(j)$ for some function f .

Lag Operator

- Lag operator \mathcal{L} is defined by $\mathcal{L}Y_t = Y_{t-1}$.
- ARMA model in terms of the lag operator:

$$\left(1 - \sum_{i=1}^p a_i \mathcal{L}^i \right) Y_t = \left(1 + \sum_{j=1}^q b_j \mathcal{L}^j \right) \epsilon_t$$

- Characteristic polynomial

$$P(z) = 1 - \sum_{i=1}^p a_i z^i$$

can be used to study properties of this stochastic process.

Weak Stationarity of ARMA

- **Theorem:** an ARMA(p, q) process is weakly stationary if the roots of the characteristic polynomial $P(z)$ are outside the unit circle.

Proof

- If roots of the characteristic polynomial are outside the unit circle then:

$$\begin{aligned} P(z) &= 1 - \sum_{i=1}^p a_i z^i = c(\psi_1 - z) \cdots (\psi_p - z) \\ &= c'(1 - \psi_1^{-1}z) \cdots (1 - \psi_p^{-1}z) \end{aligned}$$

where $|\psi_i| > 1$ for all $i = 1, \dots, p$ and c, c' are constants.

Proof

- Therefore, the ARMA(p,q) process

$$\left(1 - \sum_{i=1}^p a_i \mathcal{L}^i\right) Y_t = \left(1 + \sum_{j=1}^q b_j \mathcal{L}^j\right) \epsilon_t$$

admits MA(∞) representation:

$$Y_t = \left(1 - \psi_1^{-1} \mathcal{L}\right)^{-1} \cdots \left(1 - \psi_p^{-1} \mathcal{L}\right)^{-1} \left(1 + \sum_{j=1}^q b_j \mathcal{L}^j\right) \epsilon_t$$

where

$$\left(1 - \psi_i^{-1} \mathcal{L}\right)^{-1} = \sum_{k=0}^{\infty} \left(-\psi_i^{-1} \mathcal{L}\right)^k$$

is well-defined since $|\psi_i^{-1}| < 1$.

Proof

- Therefore, it suffices to show that

$$Y_t = \sum_{j=0}^{\infty} \phi_j \epsilon_{t-j}$$

is weakly stationary.

- The mean is constant

$$\mathbb{E}[Y_t] = \sum_{j=0}^{\infty} \phi_j \mathbb{E}[\epsilon_{t-j}] = 0.$$

- Covariance function $\mathbb{E}[Y_t Y_{t-l}]$ only depends on the lag l :

$$\mathbb{E}[Y_t Y_{t-l}] = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \phi_k \phi_j \mathbb{E}[\epsilon_{t-j} \epsilon_{t-l-k}] = \sum_{j=0}^{\infty} \phi_j \phi_{j+l}.$$

ARIMA

- Non-stationary processes can be modeled using processes whose characteristic polynomial has unit roots.
- Characteristic polynomial with unit roots can be factored:

$$P(z) = R(z)(1 - z)^D$$

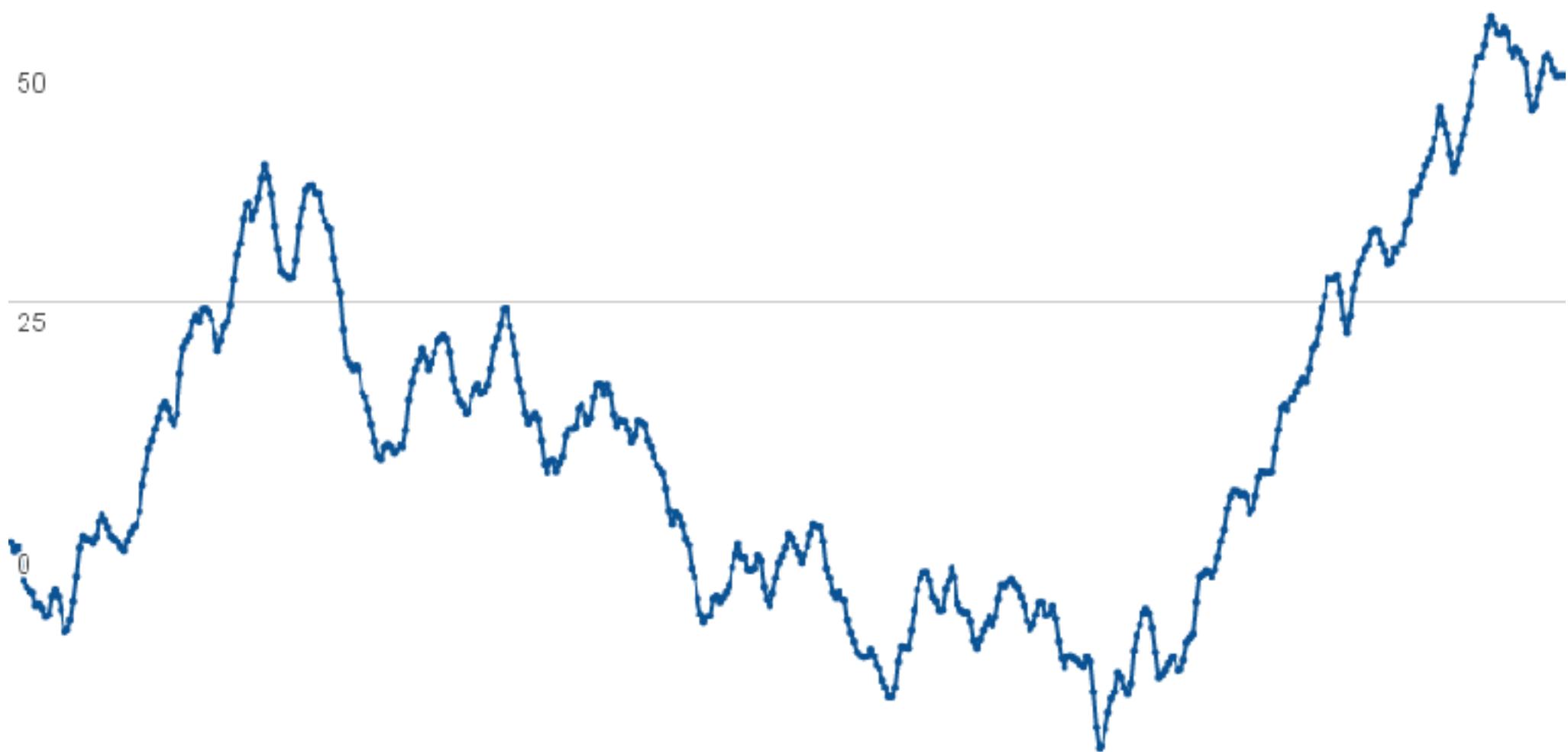
where $R(z)$ has no unit roots.

- **Definition:** ARIMA(p, D, q) model is an ARMA(p, q) model for $(1 - \mathcal{L})^D Y_t$:

$$\left(1 - \sum_{i=1}^p a_i \mathcal{L}^i\right) \left(1 - \mathcal{L}\right)^D Y_t = \left(1 + \sum_{j=1}^q b_j \mathcal{L}^j\right) \epsilon_t.$$

ARIMA

ARIMA Process



Other Extensions

■ Further variants:

- models with seasonal components (SARIMA).
- models with side information (ARIMAX).
- models with long-memory (ARFIMA).
- multi-variate time series models (VAR).
- models with time-varying coefficients.
- other non-linear models.

Modeling Variance

(Engle, 1982; Bollerslev, 1986)

- **Definition:** the generalized autoregressive conditional heteroscedasticity GARCH(p, q) model is an ARMA(p, q) model for the variance σ_t of the noise term ϵ_t :

$$\forall t, \sigma_{t+1}^2 = \omega + \sum_{i=0}^{p-1} \alpha_i \sigma_{t-i}^2 + \sum_{j=0}^{q-1} \beta_j \epsilon_{t-j}^2$$

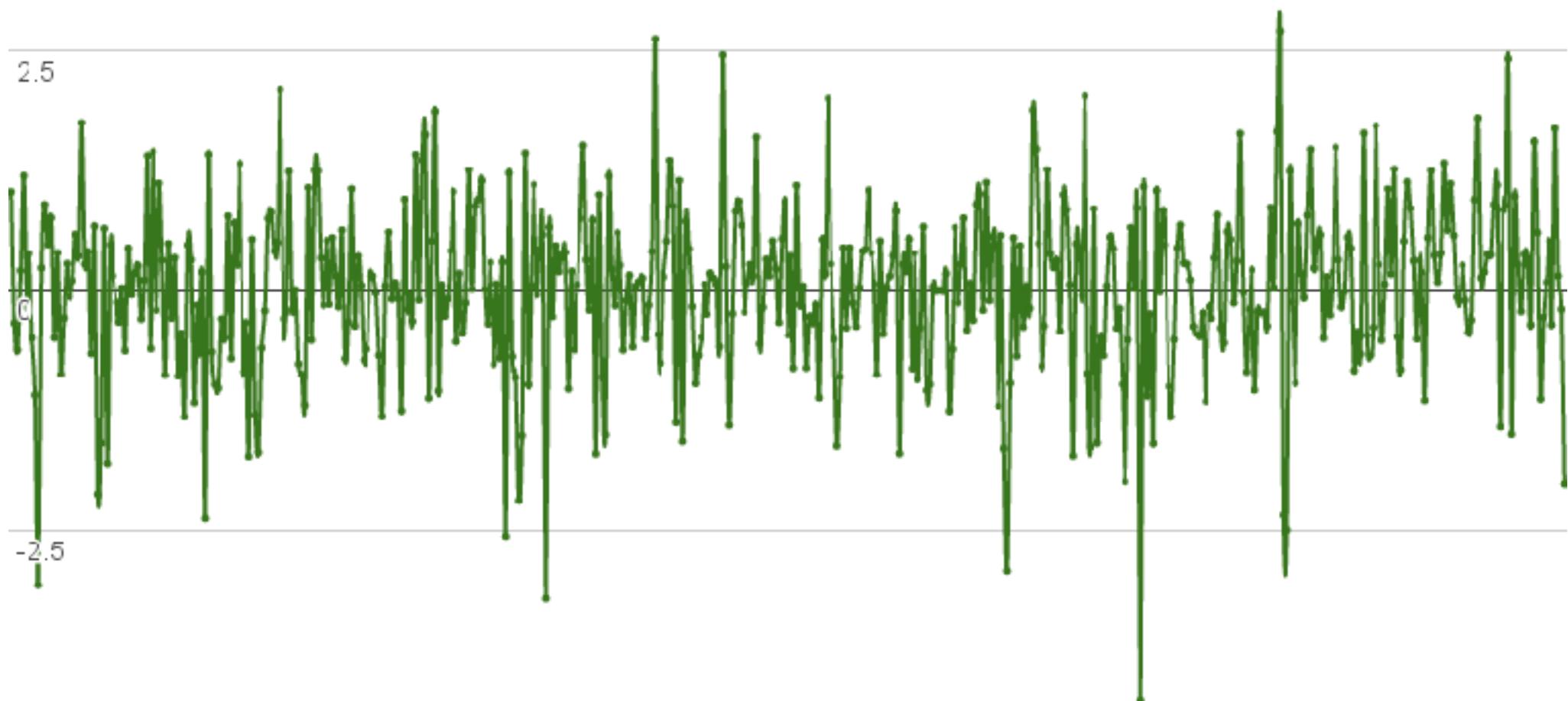
where

- ϵ_t s are zero mean Gaussian random variables with variance σ_t conditioned on $\{Y_{t-1}, Y_{t-2}, \dots\}$.
- $\omega > 0$ is the mean parameter.

GARCH Process

ε

GARCH Process



State-Space Models

- Continuous state space version of Hidden Markov Models:

$$\mathbf{X}_{t+1} = \mathbf{B}\mathbf{X}_t + \mathbf{U}_t,$$

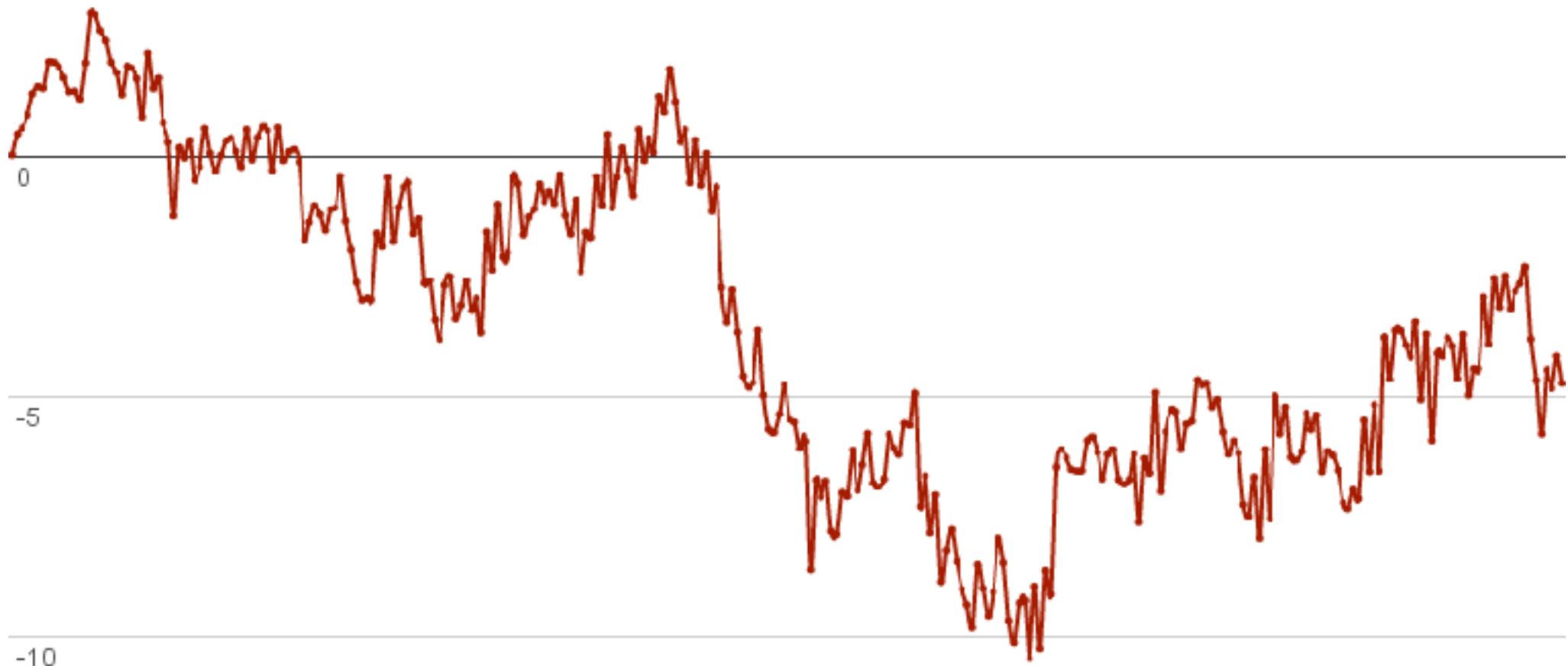
$$Y_t = \mathbf{A}\mathbf{X}_t + \epsilon_t$$

where

- \mathbf{X}_t is an n -dimensional state vector.
- Y_t is an observed stochastic process.
- \mathbf{A} and \mathbf{B} are model parameters.
- \mathbf{U}_t and ϵ_t are noise terms.

State-Space Models

Local level state-space model



Estimation

- Different methods for estimating model parameters:
 - Maximum likelihood estimation:
 - Requires further parametric assumptions on the noise distribution (e.g. Gaussian).
 - Method of moments (Yule-Walker estimator).
 - Conditional and unconditional least square estimation.
 - Restricted to certain models.

Invertibility of ARMA

- **Definition:** an ARMA(p,q) process is invertible if the roots of the polynomial

$$Q(z) = 1 + \sum_{j=1}^q b_j z^j$$

are outside the unit circle.

Learning guarantee

- **Theorem:** assume $Y_t \sim \text{ARMA}(p,q)$ is weakly stationary and invertible. Let $\hat{\mathbf{a}}_T$ denote the least square estimate of $\mathbf{a} = (a_1, \dots, a_p)$ and assume that p is known. Then, $\|\hat{\mathbf{a}}_T - \mathbf{a}\|$ converges in probability to zero.
- Similar results hold for other estimators and other models.

Notes

- Many other generative models exist.
- Learning guarantees are asymptotic.
- Model needs to be correctly specified.
- Non-stationarity needs to be modeled explicitly.

Theory

Time Series Forecasting

- **Training data:** finite sample realization of some stochastic process,

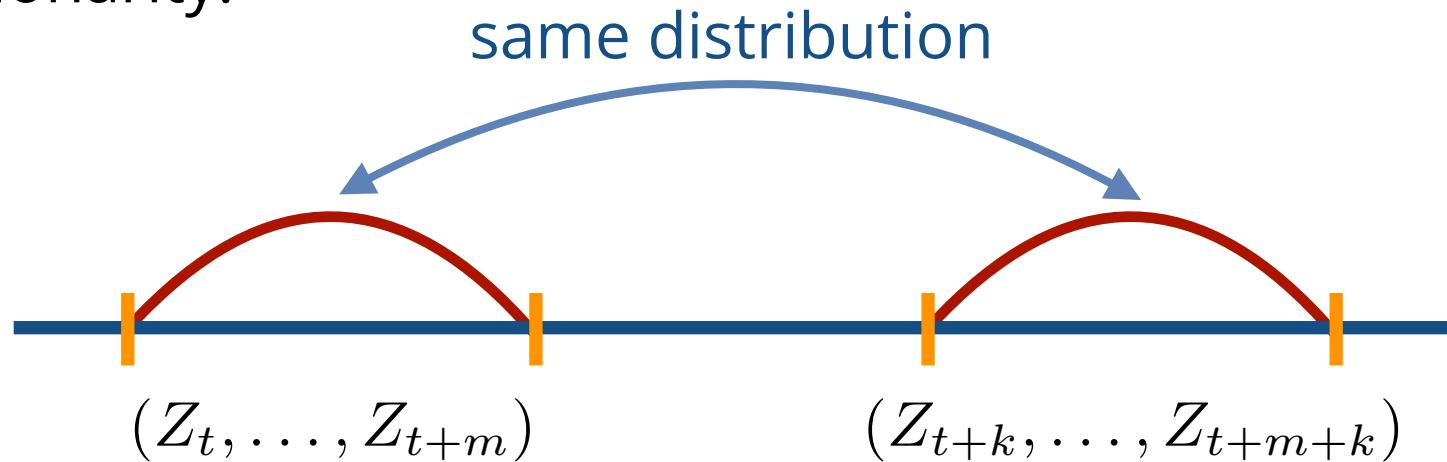
$$(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}.$$

- **Loss function:** $L: H \times \mathcal{Z} \rightarrow [0, 1]$, where H is a hypothesis set of functions mapping from \mathcal{X} to \mathcal{Y} .
- **Problem:** find $h \in H$ with small path-dependent expected loss,

$$\mathcal{L}(h, \mathbf{Z}_1^T) = \mathbb{E}_{Z_{T+1}} [L(h, Z_{T+1}) | \mathbf{Z}_1^T].$$

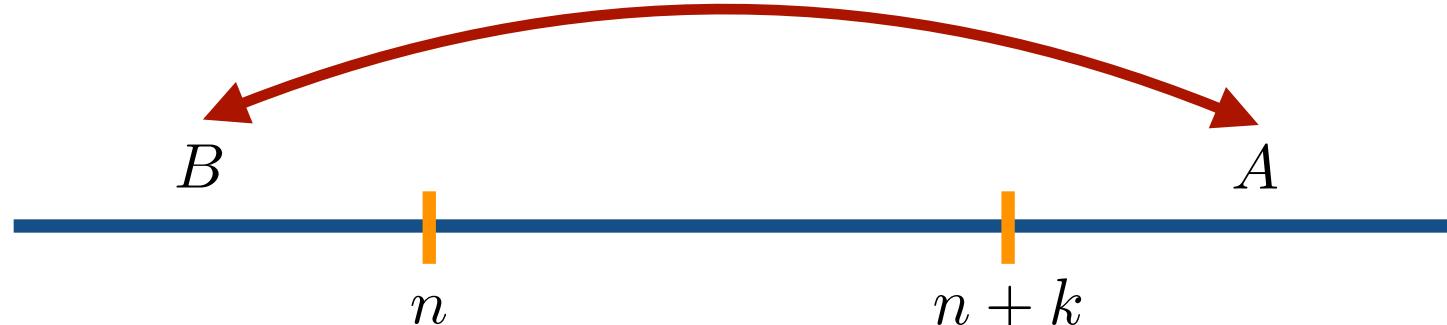
Standard Assumptions

- Stationarity:



- Mixing:

dependence between events decaying with k .

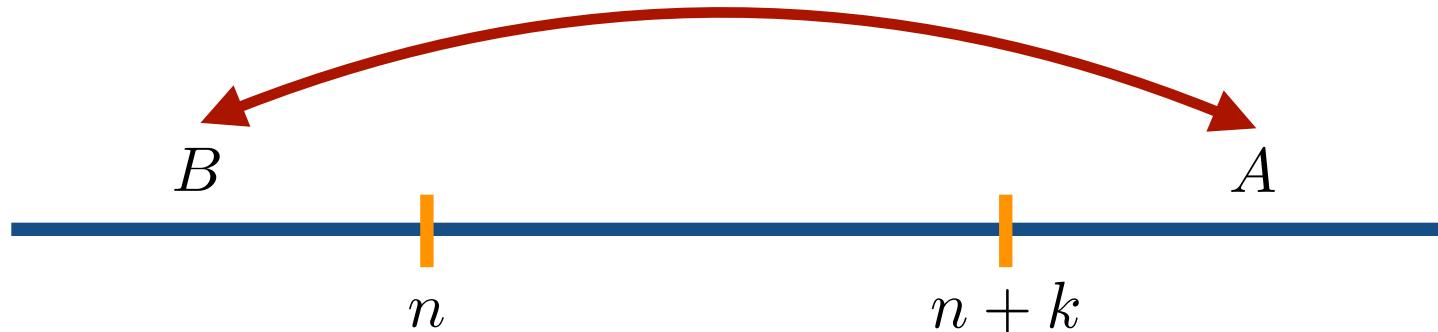


β -Mixing

- **Definition:** a sequence of random variables $\mathbf{Z} = \{Z_t\}_{-\infty}^{+\infty}$ is β -mixing if

$$\beta(k) = \sup_n \mathbb{E}_{B \in \sigma_{-\infty}^n} \left[\sup_{A \in \sigma_{n+k}^\infty} \left| \mathbb{P}[A | B] - \mathbb{P}[A] \right| \right] \rightarrow 0.$$

dependence between events decaying with k .



Learning Theory

- Stationary and β -mixing process: generalization bounds.
 - PAC-learning preserved in that setting (Vidyasagar, 1997).
 - VC-dimension bounds for binary classification (Yu, 1994).
 - covering number bounds for regression (Meir, 2000).
 - Rademacher complexity bounds for general loss functions (MM and Rostamizadeh, 2000).
 - PAC-Bayesian bounds (Alquier et al., 2014).

Learning Theory

- Stationarity and mixing: algorithm-dependent bounds.
 - AdaBoost ([Lozano et al., 1997](#)).
 - general stability bounds ([MM and Rostamizadeh, 2010](#)).
 - regularized ERM ([Steinwart and Christmann, 2009](#)).
 - stable on-line algorithms ([Agarwal and Duchi, 2013](#)).

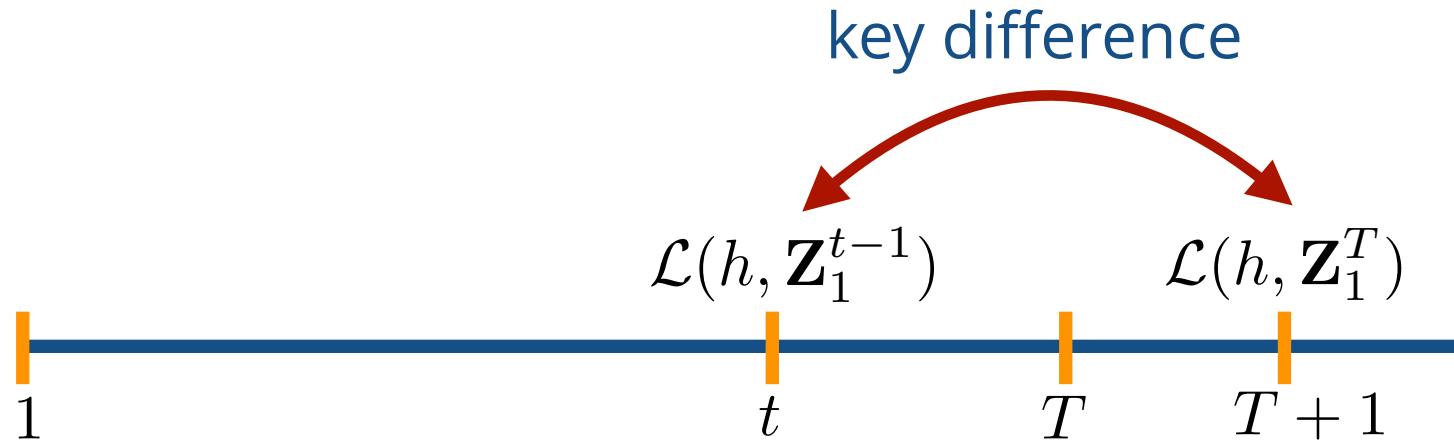
Problem

- Stationarity and mixing assumptions:
 - often do not hold (think trend or periodic signals).
 - not testable.
 - estimating mixing parameters can be hard, even if general functional form known.
 - hypothesis set and loss function ignored.

Questions

- Is learning with general (non-stationary, non-mixing) stochastic processes possible?
- Can we design algorithms with theoretical guarantees?
→ need a new tool for the analysis.

Key Quantity - Fixed h



→ Key average quantity: $\left| \frac{1}{T} \sum_{t=1}^T [\mathcal{L}(h, \mathbf{z}_1^T) - \mathcal{L}(h, \mathbf{z}_1^{t-1})] \right|$.

Discrepancy

■ Definition:

$$\Delta = \sup_{h \in H} \left| \mathcal{L}(h, \mathbf{Z}_1^T) - \frac{1}{T} \sum_{t=1}^T \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right|.$$

- captures hypothesis set and loss function.
- can be estimated from data, under mild assumptions.
- $\Delta = 0$ in IID case or for weakly stationary processes with linear hypotheses and squared loss (K and MM, 2014).

Weighted Discrepancy

- **Definition:** extension to weights $(q_1, \dots, q_T) = \mathbf{q}$.

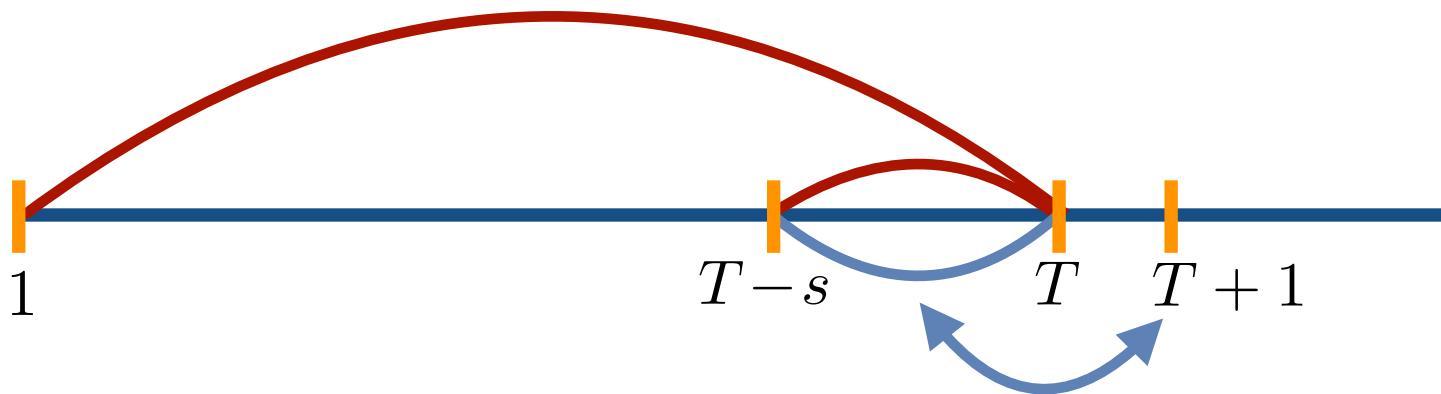
$$\Delta(\mathbf{q}) = \sup_{h \in H} \left| \mathcal{L}(h, \mathbf{Z}_1^T) - \sum_{t=1}^T q_t \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right|.$$

- strictly extends discrepancy definition in drifting (MM and Muñoz Medina, 2012) or domain adaptation (Mansour, MM, Rostamizadeh 2009; Cortes and MM 2011, 2014); or for binary loss (Devroye et al., 1996; Ben-David et al., 2007).
- admits upper bounds in terms of relative entropy, or in terms of ϕ -mixing coefficients of asymptotic stationarity for an asymptotically stationary process.

Estimation

- Decomposition: $\Delta(\mathbf{q}) \leq \Delta_0(\mathbf{q}) + \Delta_s$.

$$\begin{aligned}\Delta(\mathbf{q}) &\leq \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^T \mathcal{L}(h, \mathbf{Z}_1^{t-1}) - \sum_{t=1}^T q_t \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right) \\ &\quad + \sup_{h \in H} \left(\mathcal{L}(h, \mathbf{Z}_1^T) - \frac{1}{s} \sum_{t=T-s+1}^T \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right).\end{aligned}$$



Learning Guarantee

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$ and $\alpha > 0$,

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t L(h, Z_t) + \Delta(\mathbf{q}) + 2\alpha + \|\mathbf{q}\|_2 \sqrt{2 \log \frac{\mathbb{E}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where $\mathcal{G} = \{z \mapsto L(h, z) : h \in H\}$.

Bound with Emp. Discrepancy

- **Corollary:** for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$ and $\alpha > 0$,

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^T q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + 4\alpha \\ &\quad + \left[\|\mathbf{q}\|_2 + \|\mathbf{q} - \mathbf{u}_s\|_2 \right] \sqrt{2 \log \frac{2 \mathbb{E}_{\mathbf{z}} [\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},\end{aligned}$$

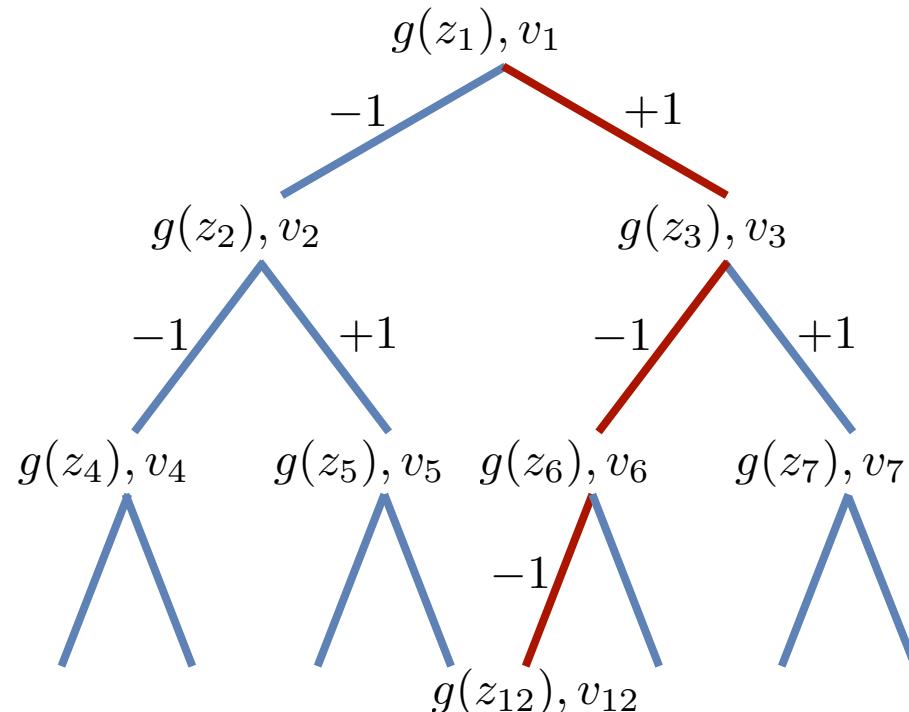
where $\left\{ \begin{array}{l} \widehat{\Delta}(\mathbf{q}) = \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^T L(h, Z_t) - \sum_{t=1}^T q_t L(h, Z_t) \right) \\ \mathbf{u}_s \text{ unif. dist. over } [T-s, T] \\ \mathcal{G} = \{z \mapsto L(h, z) : h \in H\}. \end{array} \right.$

Weighted Sequential α -Cover

(Rakhlin et al., 2010; K and MM, 2015)

- **Definition:** let \mathbf{z} be a \mathcal{Z} -valued full binary tree of depth T . Then, a set of trees \mathcal{V} is an l_1 -norm \mathbf{q} -weighted α -cover of a function class \mathcal{G} on \mathbf{z} if

$$\forall g \in \mathcal{G}, \forall \boldsymbol{\sigma} \in \{\pm 1\}^T, \exists \mathbf{v} \in \mathcal{V}: \sum_{t=1}^T |v_t(\boldsymbol{\sigma}) - g(z_t(\boldsymbol{\sigma}))| \leq \frac{\alpha}{\|\mathbf{q}\|_\infty}.$$



$$\left\| \begin{bmatrix} v_1 - g(z_1) \\ v_3 - g(z_3) \\ v_6 - g(z_6) \\ v_{12} - g(z_{12}) \end{bmatrix} \right\|_1 \leq \frac{\alpha}{\|\mathbf{q}\|_\infty}.$$

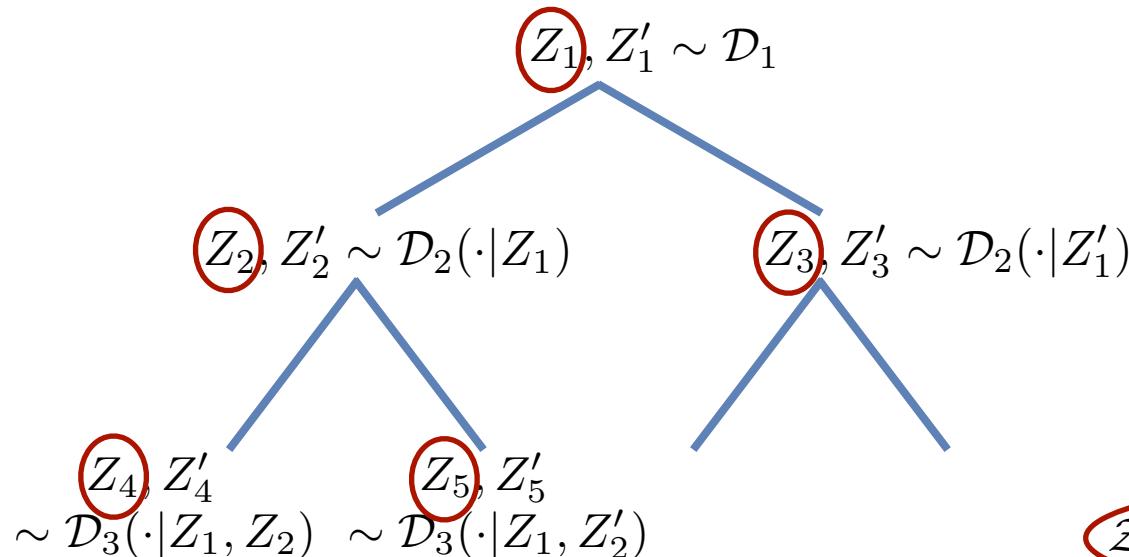
Sequential Covering Numbers

■ Definitions:

- sequential covering number:

$$\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z}) = \min\{|\mathcal{V}| : \mathcal{V} \text{ } l_1\text{-norm } \mathbf{q}\text{-weighted } \alpha\text{-cover of } \mathcal{G}\}.$$

- expected sequential covering number: $\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}_T} [\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]$.



\mathcal{Z}_T : distribution based on Z_t s.

Proof

■ Key quantities: $\Phi(\mathbf{Z}_1^T) = \sup_{h \in H} \left(\mathcal{L}(h, Z_T) - \sum_{t=1}^T q_t L(h, Z_t) \right)$

$$\Delta(\mathbf{q}) = \sup_{h \in H} \left| \mathcal{L}(h, \mathbf{Z}_1^T) - \sum_{t=1}^T q_t \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right|.$$

■ Chernoff technique: for any $t > 0$,

$$\begin{aligned} & \mathbb{P} [\Phi(\mathbf{Z}_1^T) - \Delta(\mathbf{q}) > \epsilon] \\ & \leq \mathbb{P} \left[\sup_{h \in H} \sum_{t=1}^T q_t [\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)] > \epsilon \right] \quad (\text{sub-add. of sup}) \\ & = \mathbb{P} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t [\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)] \right) > e^{t\epsilon} \right] \quad (t > 0) \\ & \leq e^{-t\epsilon} \mathbb{E} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t [\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)] \right) \right]. \quad (\text{Markov's ineq.}) \end{aligned}$$

Symmetrization

- Key tool: decoupled tangent sequence \mathbf{Z}'_1^T associated to \mathbf{Z}_1^T .
 - Z_t and Z'_t i.i.d. given \mathbf{Z}_1^{t-1} .

$$\begin{aligned} & \mathbb{P} [\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)] \\ & \leq e^{-t\epsilon} \mathbb{E} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t [\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)] \right) \right] \\ & = e^{-t\epsilon} \mathbb{E} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t [\mathbb{E}[L(h, Z'_t) | \mathbf{Z}_1^{t-1}] - L(h, Z_t)] \right) \right] \quad (\text{tangent seq.}) \\ & = e^{-t\epsilon} \mathbb{E} \left[\exp \left(t \sup_{h \in H} \mathbb{E} \left[\sum_{t=1}^T q_t [L(h, Z'_t) - L(h, Z_t)] \mid \mathbf{Z}_1^T \right] \right) \right] \quad (\text{lin. of expectation}) \\ & \leq e^{-t\epsilon} \mathbb{E} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t [L(h, Z'_t) - L(h, Z_t)] \right) \right]. \quad (\text{Jensen's ineq.}) \end{aligned}$$

Symmetrization

$$\begin{aligned}
& \mathbb{P} [\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)] \\
& \leq e^{-t\epsilon} \mathbb{E} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t [L(h, Z'_t) - L(h, Z_t)] \right) \right] \\
& = e^{-t\epsilon} \mathbb{E}_{(\mathbf{z}, \mathbf{z}')} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t [L(h, z'_t(\boldsymbol{\sigma})) - L(h, z_t(\boldsymbol{\sigma}))] \right) \right] \quad (\text{tangent seq. prop.}) \\
& = e^{-t\epsilon} \mathbb{E}_{(\mathbf{z}, \mathbf{z}')} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left(t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t L(h, z'_t(\boldsymbol{\sigma})) + t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t L(h, z_t(\boldsymbol{\sigma})) \right) \right] \quad (\text{sub-add. of sup}) \\
& \leq e^{-t\epsilon} \mathbb{E}_{(\mathbf{z}, \mathbf{z}')} \mathbb{E}_{\boldsymbol{\sigma}} \left[\frac{1}{2} \exp \left(2t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t L(h, z'_t(\boldsymbol{\sigma})) \right) \right. \\
& \quad \left. + \frac{1}{2} \exp \left(2t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t L(h, z_t(\boldsymbol{\sigma})) \right) \right] \quad (\text{convexity. of exp}) \\
& = e^{-t\epsilon} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left(2t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t L(h, z_t(\boldsymbol{\sigma})) \right) \right].
\end{aligned}$$

Covering Number

$$\begin{aligned}
& \mathbb{P} [\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)] \\
& \leq e^{-t\epsilon} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left(2t \sup_{h \in H} \sum_{t=1}^T q_t \sigma_t L(h, z_t(\boldsymbol{\sigma})) \right) \right] \\
& \leq e^{-t\epsilon} \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left(2t \left[\max_{\mathbf{v} \in \mathcal{V}} \sum_{t=1}^T q_t \sigma_t v_t(\boldsymbol{\sigma}) + \alpha \right] \right) \right] \quad (\alpha\text{-covering}) \\
& \leq e^{-t(\epsilon-2\alpha)} \mathbb{E}_{\mathbf{z}} \left[\sum_{\mathbf{v} \in \mathcal{V}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp \left(2t \sum_{t=1}^T q_t \sigma_t v_t(\boldsymbol{\sigma}) \right) \right] \right] \quad (\text{monotonicity of exp}) \\
& \leq e^{-t(\epsilon-2\alpha)} \mathbb{E}_{\mathbf{z}} \left[\sum_{\mathbf{v} \in \mathcal{V}} \exp \left(\frac{t^2 \|\mathbf{q}\|^2}{2} \right) \right] \quad (\text{Hoeffding's ineq.}) \\
& \leq \mathbb{E}_{\mathbf{z}} [\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})] \exp \left[-t(\epsilon - 2\alpha) + \frac{t^2 \|\mathbf{q}\|^2}{2} \right].
\end{aligned}$$

Algorithms

Review

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$ and $\alpha > 0$,

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^T q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + 4\alpha \\ &\quad + \left[\|\mathbf{q}\|_2 + \|\mathbf{q} - \mathbf{u}_s\|_2 \right] \sqrt{2 \log \frac{2 \mathbb{E}_{\mathbf{z}} [\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}}.\end{aligned}$$

- This bound can be extended to hold uniformly over \mathbf{q} at the price of the additional term:

$$\tilde{O}(\|\mathbf{q} - \mathbf{u}\|_1 \sqrt{\log_2 \log_2(1 - \|\mathbf{q} - \mathbf{u}\|)^{-1}}).$$

- Data-dependent learning guarantee.

Discrepancy-Risk Minimization

- Key Idea: directly optimize the upper bound on generalization over q and h .
- This problem can be solved efficiently for some L and H .

Kernel-Based Regression

- Squared loss function: $L(y, y') = (y - y')^2$

- Hypothesis set: for PDS kernel K ,

$$H = \left\{ x \mapsto \mathbf{w} \cdot \Phi_K(x) : \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda \right\}.$$

- Complexity term can be bounded by

$$O\left((\log^{3/2} T)\Lambda \sup_x K(x, x)\|\mathbf{q}\|_2\right).$$

Instantaneous Discrepancy

- Empirical discrepancy can be further upper bounded in terms of instantaneous discrepancies:

$$\widehat{\Delta}(\mathbf{q}) \leq \sum_{t=1}^T q_t d_t + M \|\mathbf{q} - \mathbf{u}\|_1$$

where $M = \sup_{y,y'} L(y, y')$ and

$$d_t = \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^T L(h, Z_t) - L(h, Z_t) \right).$$

Proof

- By sub-additivity of supremum

$$\begin{aligned}\widehat{\Delta}(\mathbf{q}) &= \sup_{h \in H} \left\{ \frac{1}{s} \sum_{t=T-s+1}^T L(h, Z_t) - \sum_{t=1}^T q_t L(h, Z_t) \right\} \\ &= \sup_{h \in H} \left\{ \sum_{t=1}^T q_t \left(\frac{1}{s} \sum_{t=T-s+1}^T L(h, Z_t) - L(h, Z_t) \right) \right. \\ &\quad \left. + \sum_{t=1}^T \left(\frac{1}{T} - q_t \right) \frac{1}{s} \sum_{t=T-s+1}^T L(h, Z_t) \right\} \\ &\leq \sum_{t=1}^T q_t \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^T L(h, Z_t) - q_t L(h, Z_t) \right) + M \|\mathbf{u} - \mathbf{q}\|_1.\end{aligned}$$

Computing Discrepancies

- Instantaneous discrepancy for kernel-based hypothesis with squared loss:

$$d_t = \sup_{\|\mathbf{w}'\| \leq \Lambda} \left(\sum_{s=1}^T u_s (\mathbf{w}' \cdot \Phi_K(x_s) - y_s)^2 - (\mathbf{w}' \cdot \Phi_K(x_t) - y_t)^2 \right).$$

- Difference of convex (DC) functions.
- Global optimum via DC-programming: (Tao and Ahn, 1998).

Discrepancy-Based Forecasting

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all kernel-based hypothesis $h \in H$ and all $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^T q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s \\ &\quad + \tilde{O}\left(\log^{3/2} T \sup_x K(x, x) \Lambda + \|\mathbf{q} - \mathbf{u}\|_1\right).\end{aligned}$$

- Corresponding optimization problem:

$$\min_{\mathbf{q} \in [0,1]^T, \mathbf{w}} \left\{ \sum_{t=1}^T q_t (\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2 + \lambda_1 \sum_{t=1}^T q_t d_t + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \lambda_3 \|\mathbf{q} - \mathbf{u}\|_1 \right\}.$$

Discrepancy-Based Forecasting

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all kernel-based hypothesis $h \in H$ and all $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \boxed{\sum_{t=1}^T q_t L(h, Z_t)} + \widehat{\Delta}(\mathbf{q}) + \Delta_s \\ &\quad + \tilde{O}\left(\log^{3/2} T \sup_x K(x, x) \Lambda + \|\mathbf{q} - \mathbf{u}\|_1\right).\end{aligned}$$

- Corresponding optimization problem:

$$\min_{\mathbf{q} \in [0,1]^T, \mathbf{w}} \left\{ \boxed{\sum_{t=1}^T q_t (\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2} + \lambda_1 \sum_{t=1}^T q_t d_t + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \lambda_3 \|\mathbf{q} - \mathbf{u}\|_1 \right\}.$$

Discrepancy-Based Forecasting

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all kernel-based hypothesis $h \in H$ and all $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^T q_t L(h, Z_t) + \boxed{\widehat{\Delta}(\mathbf{q})} + \Delta_s \\ &\quad + \tilde{O}\left(\log^{3/2} T \sup_x K(x, x) \Lambda + \|\mathbf{q} - \mathbf{u}\|_1\right).\end{aligned}$$

- Corresponding optimization problem:

$$\min_{\mathbf{q} \in [0,1]^T, \mathbf{w}} \left\{ \sum_{t=1}^T q_t (\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2 + \boxed{\lambda_1 \sum_{t=1}^T q_t d_t} + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \lambda_3 \|\mathbf{q} - \mathbf{u}\|_1 \right\}.$$

Discrepancy-Based Forecasting

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all kernel-based hypothesis $h \in H$ and all $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^T q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s \\ &\quad + \tilde{O}\left(\log^{3/2} T \sup_x K(x, x) \Lambda + \|\mathbf{q} - \mathbf{u}\|_1\right).\end{aligned}$$

- Corresponding optimization problem:

$$\min_{\mathbf{q} \in [0,1]^T, \mathbf{w}} \left\{ \sum_{t=1}^T q_t (\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2 + \lambda_1 \sum_{t=1}^T q_t d_t + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \lambda_3 \|\mathbf{q} - \mathbf{u}\|_1 \right\}.$$

Discrepancy-Based Forecasting

- **Theorem:** for any $\delta > 0$, with probability at least $1 - \delta$, for all kernel-based hypothesis $h \in H$ and all $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\begin{aligned}\mathcal{L}(h, \mathbf{Z}_1^T) &\leq \sum_{t=1}^T q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s \\ &\quad + \tilde{O}\left(\log^{3/2} T \sup_x K(x, x) \Lambda + \boxed{\|\mathbf{q} - \mathbf{u}\|_1}\right).\end{aligned}$$

- Corresponding optimization problem:

$$\min_{\mathbf{q} \in [0,1]^T, \mathbf{w}} \left\{ \sum_{t=1}^T q_t (\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2 + \lambda_1 \sum_{t=1}^T q_t d_t + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \boxed{\lambda_3 \|\mathbf{q} - \mathbf{u}\|_1} \right\}.$$

Convex Problem

- Change of variable: $r_t = 1/q_t$.
- Upper bound: $|r_t^{-1} - 1/T| \leq T^{-1}|r_t - T|$.

$$\min_{\mathbf{r} \in \mathcal{D}, \mathbf{w}} \left\{ \sum_{t=1}^T \frac{(\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2 + \lambda_1 d_t}{r_t} + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \lambda_3 \sum_{t=1}^T |r_t - T| \right\}.$$

- where $\mathcal{D} = \{\mathbf{r}: r_t \geq 1\}$.
- convex optimization problem.

Two-Stage Algorithm

- Minimize empirical discrepancy $\hat{\Delta}(\mathbf{q})$ over \mathbf{q} (convex optimization).
- Solve (weighted) kernel-ridge regression problem:

$$\min_{\mathbf{w}} \left\{ \sum_{t=1}^T q_t^* (\mathbf{w} \cdot \Psi_K(x_t) - y_t)^2 + \lambda \|\mathbf{w}\|_{\mathbb{H}} \right\}$$

where \mathbf{q}^* is the solution to discrepancy minimization problem.

Preliminary Experiments

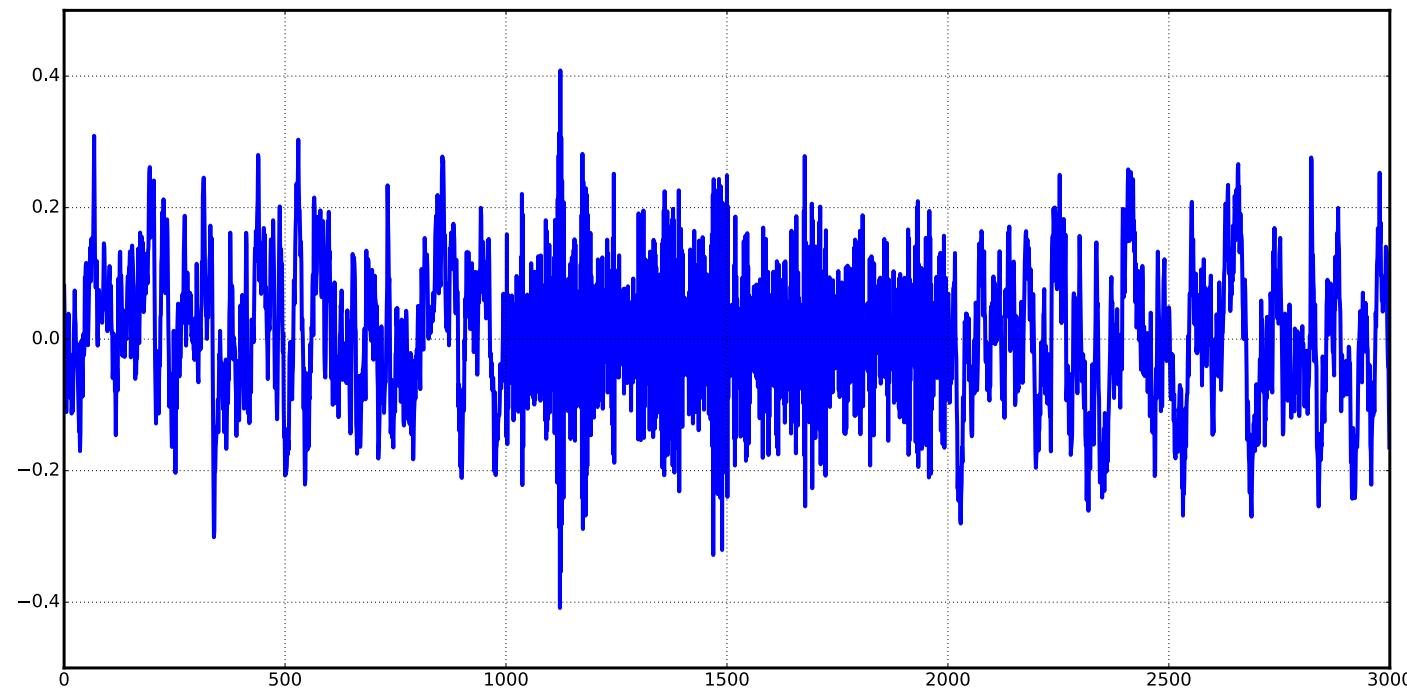
■ Artificial data sets:

ads1: $Y_t = \alpha_t Y_{t-1} + \epsilon_t$, $\alpha_t = -0.9$ if $t \in [1000, 2000]$ and 0.9 otherwise,

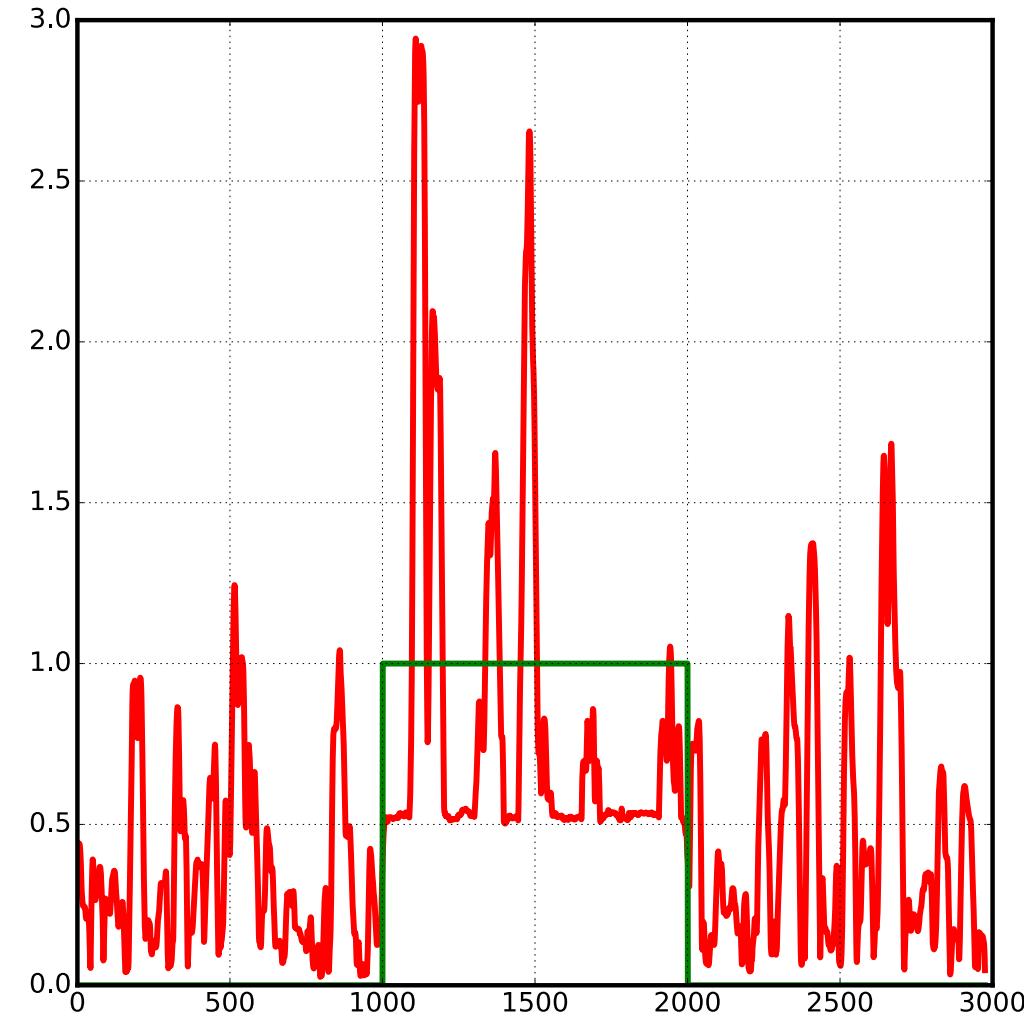
ads2: $Y_t = \alpha_t Y_{t-1} + \epsilon_t$, $\alpha_t = 1 - (t/1500)$,

ads3: $Y_t = \alpha_{i(t)} Y_{t-1} + \epsilon_t$, $\alpha_1 = -0.5$ and $\alpha_2 = 0.9$,

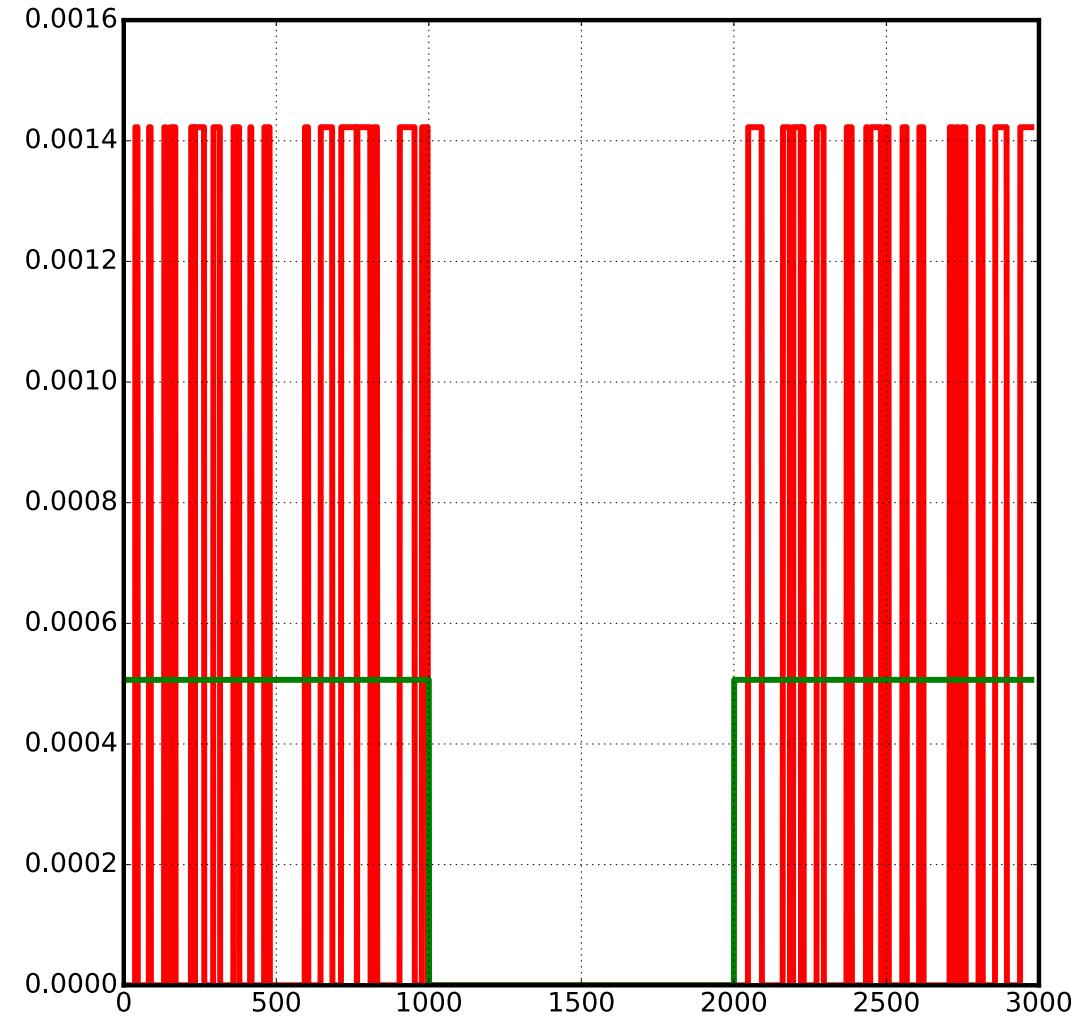
ads4: $Y_t = -0.5Y_{t-1} + \epsilon_t$,



True vs. Empirical Discrepancies

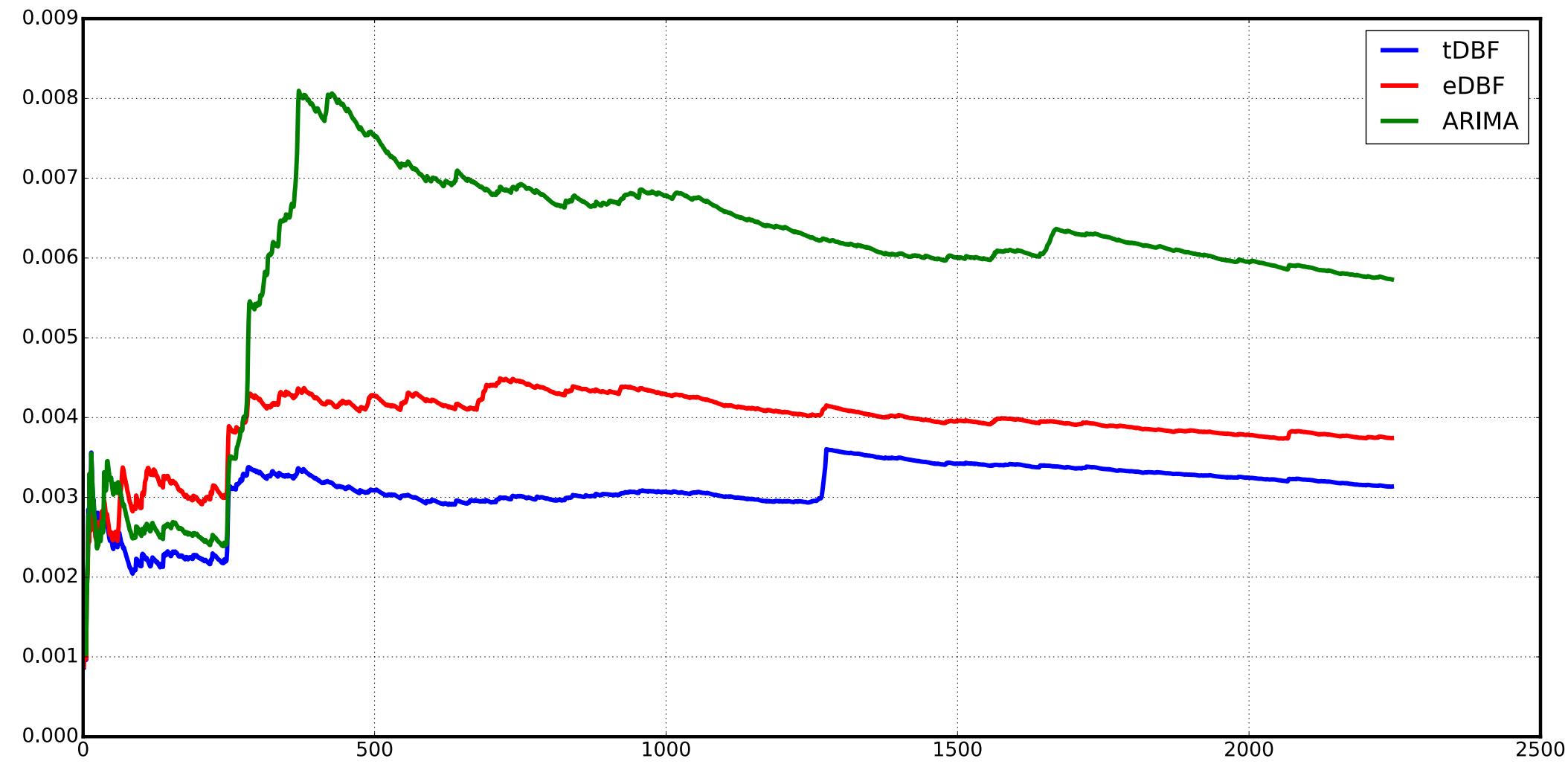


Discrepancies



Weights

Running MSE



Real-world Data

- Commodity prices, exchange rates, temperatures & climate.

Dataset	DBF	ARIMA
bitcoin	4.400×10^{-3} (26.500×10^{-3})	4.900×10^{-3} (29.990×10^{-3})
coffee	3.080×10^{-3} (6.570×10^{-3})	3.260×10^{-3} (6.390×10^{-3})
eur/jpy	7.100×10^{-5} (16.900×10^{-5})	7.800×10^{-5} (24.200×10^{-5})
jpy/usd	9.770×10^{-1} (25.893×10^{-1})	10.004×10^{-1} (27.531×10^{-1})
mso	32.876×10^0 (55.586×10^0)	32.193×10^0 (51.109×10^0)
silver	7.640×10^{-4} (46.65×10^{-4})	34.180×10^{-4} (158.090×10^{-4})
soy	5.071×10^{-2} (9.938×10^{-2})	5.003×10^{-2} (10.097×10^{-2})
temp	6.418×10^0 (9.958×10^0)	6.461×10^0 (10.324×10^0)

Time Series Prediction & On-line Learning

Two Learning Scenarios

- Stochastic scenario:
 - distributional assumption.
 - performance measure: expected loss.
 - guarantees: generalization bounds.
- On-line scenario:
 - no distributional assumption.
 - performance measure: regret.
 - guarantees: regret bounds.
 - active research area: (Cesa-Bianchi and Lugosi, 2006; Anava et al. 2013, 2015, 2016; Bousquet and Warmuth, 2002; Herbster and Warmuth, 1998, 2001; Koolen et al., 2015).

On-Line Learning Setup

- Adversarial setting with hypothesis/action set H .
- For $t = 1$ to T do
 - player receives $x_t \in \mathcal{X}$.
 - player selects $h_t \in H$.
 - adversary selects $y_t \in \mathcal{Y}$.
 - player incurs loss $L(h_t(x_t), y_t)$.
- **Objective:** minimize (external) regret

$$\text{Reg}_T = \sum_{t=1}^T L(h_t(x_t), y_t) - \min_{h \in H^*} \sum_{t=1}^T L(h(x_t), y_t).$$

Example: Exp. Weights (EW)

- Expert set $H^* = \{\mathcal{E}_1, \dots, \mathcal{E}_N\}$, $H = \text{conv}(H^*)$.

$\text{EW}(\{\mathcal{E}_1, \dots, \mathcal{E}_N\})$

```
1  for  $i \leftarrow 1$  to  $N$  do
2       $w_{1,i} \leftarrow 1$ 
3  for  $t \leftarrow 1$  to  $T$  do
4      RECEIVE( $x_t$ )
5       $h_t \leftarrow \frac{\sum_{i=1}^N w_{t,i} \mathcal{E}_i}{\sum_{i=1}^N w_{t,i}}$ 
6      RECEIVE( $y_t$ )
7      INCUR-LOSS( $L(h_t(x_t), y_t)$ )
8      for  $i \leftarrow 1$  to  $N$  do
9           $w_{t+1,i} \leftarrow w_{t,i} e^{-\eta L(\mathcal{E}_i(x_t), y_t)}$      $\triangleright$  (parameter  $\eta > 0$ )
10 return  $h_T$ 
```

EW Guarantee

- **Theorem:** assume that L is convex in its first argument and takes values in $[0, 1]$. Then, for any $\eta > 0$ and any sequence $y_1, \dots, y_T \in \mathcal{Y}$, the regret of EW at time T satisfies

$$\text{Reg}_T \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

For $\eta = \sqrt{8 \log N / T}$,

$$\text{Reg}_T \leq \sqrt{(T/2) \log N}.$$

$$\frac{\text{Reg}_T}{T} = O\left(\sqrt{\frac{\log N}{T}}\right).$$

EW - Proof

■ Potential: $\Phi_t = \log \sum_{i=1}^N w_{t,i}$.

■ Upper bound:

$$\begin{aligned}\Phi_t - \Phi_{t-1} &= \log \frac{\sum_{i=1}^N w_{t-1,i} e^{-\eta L(\mathcal{E}_i(x_t), y_t)}}{\sum_{i=1}^N w_{t-1,i}} \\ &= \log \left(\mathbb{E}_{w_{t-1}} [e^{-\eta L(\mathcal{E}_i(x_t), y_t)}] \right) \\ &= \log \left(\mathbb{E}_{w_{t-1}} \left[\exp \left(-\eta \left(L(\mathcal{E}_i(x_t), y_t) - \mathbb{E}_{w_{t-1}} [L(\mathcal{E}_i(x_t), y_t)] \right) - \eta \mathbb{E}_{w_{t-1}} [L(\mathcal{E}_i(x_t), y_t)] \right) \right] \right) \\ &\leq -\eta \mathbb{E}_{w_{t-1}} [L(\mathcal{E}_i(x_t), y_t)] + \frac{\eta^2}{8} \quad (\text{Hoeffding's ineq.}) \\ &\leq -\eta L(\mathbb{E}_{w_{t-1}} [\mathcal{E}_i(x_t)], y_t) + \frac{\eta^2}{8} \quad (\text{convexity of first arg. of } L) \\ &= -\eta L(h_t(x_t), y_t) + \frac{\eta^2}{8}.\end{aligned}$$

EW - Proof

- Upper bound: summing up the inequalities yields

$$\Phi_T - \Phi_0 \leq -\eta \sum_{t=1}^T L(h_t(x_t), y_t) + \frac{\eta^2 T}{8}.$$

- Lower bound:

$$\begin{aligned}\Phi_T - \Phi_0 &= \log \sum_{i=1}^N e^{-\eta \sum_{t=1}^T L(\mathcal{E}_i(x_t), y_t)} - \log N \\ &\geq \log \max_{i=1}^N e^{-\eta \sum_{t=1}^T L(\mathcal{E}_i(x_t), y_t)} - \log N \\ &= -\eta \min_{i=1}^N \sum_{t=1}^T L(\mathcal{E}_i(x_t), y_t) - \log N.\end{aligned}$$

- Comparison:

$$\sum_{t=1}^T L(h_t(x_t), y_t) - \min_{i=1}^N \sum_{t=1}^T L(\mathcal{E}_i(x_t), y_t) \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

Questions

- Can we exploit both batch and on-line to
 - design flexible algorithms for time series prediction with stochastic guarantees?
 - tackle notoriously difficult time series problems e.g., model selection, learning ensembles?

Model Selection

- **Problem:** given N time series models, how should we use sample \mathbf{Z}_1^T to select a single best model?
 - in i.i.d. case, cross-validation can be shown to be close to the structural risk minimization solution.
 - but, how do we select a validation set for general stochastic processes?
 - use most recent data?
 - use the most distant data?
 - use various splits?
 - models may have been pre-trained on \mathbf{Z}_1^T .

Learning Ensembles

- **Problem:** given a hypothesis set H and a sample \mathbf{Z}_1^T , find accurate convex combination $h = \sum_{t=1}^T q_t h_t$ with $\mathbf{h} \in H_A$ and $\mathbf{q} \in \Delta$.
 - in most general case, hypotheses may have been pre-trained on \mathbf{Z}_1^T .
- on-line-to-batch conversion for general non-stationary non-mixing processes.

On-Line-to-Batch (OTB)

- **Input:** sequence of hypotheses $\mathbf{h} = (h_1, \dots, h_T)$ returned after T rounds by an on-line algorithm \mathcal{A} minimizing general regret

$$\text{Reg}_T = \sum_{t=1}^T L(h_t, Z_t) - \inf_{\mathbf{h}^* \in \mathbf{H}^*} \sum_{t=1}^T L(\mathbf{h}^*, Z_t).$$

On-Line-to-Batch (OTB)

- **Problem:** use $\mathbf{h} = (h_1, \dots, h_T)$ to derive a hypothesis $h \in H$ with small path-dependent expected loss,

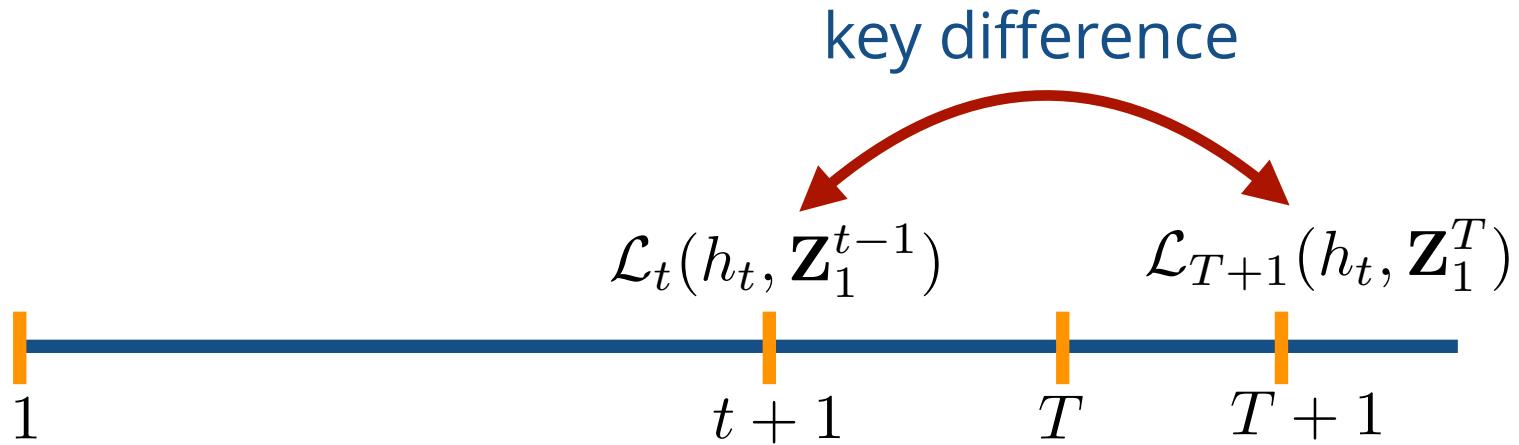
$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathbb{E}_{Z_{T+1}} [L(h, Z_{T+1}) | \mathbf{Z}_1^T].$$

- i.i.d. problem is standard: (Littlestone, 1989), (Cesa-Bianchi et al., 2004).
- but, how do we design solutions for the general time-series scenario?

Questions

- Is OTB with general (non-stationary, non-mixing) stochastic processes possible?
- Can we design algorithms with theoretical guarantees?
→ need a new tool for the analysis.

Relevant Quantity



→ Average difference: $\frac{1}{T} \sum_{t=1}^T [\mathcal{L}_{T+1}(h_t, \mathbf{z}_1^T) - \mathcal{L}_t(h_t, \mathbf{z}_1^{t-1})]$.

On-line Discrepancy

■ Definition:

$$\text{disc}(\mathbf{q}) = \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \right|.$$

- $\mathbf{H}_{\mathcal{A}}$: sequences that \mathcal{A} can return.
- $\mathbf{q} = (q_1, \dots, q_T)$: arbitrary weight vector.
- natural measure of non-stationarity or dependency.
- captures hypothesis set and loss function.
- can be efficiently estimated under mild assumptions.
- generalization of definition of [\(Kuznetsov and MM, 2015\)](#) .

Discrepancy Estimation

- Batch discrepancy estimation method.
- Alternative method:
 - assume that the loss is μ -Lipschitz.
 - assume that there exists an accurate hypothesis h^* :

$$\eta = \inf_{h^*} \mathbb{E} \left[L(Z_{T+1}, h^*(X_{T+1})) | \mathbf{Z}_1^T \right] \ll 1.$$

Discrepancy Estimation

- **Lemma:** fix sequence \mathbf{Z}_1^T in \mathcal{Z} . Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\alpha > 0$:

$$\text{disc}(\mathbf{q}) \leq \widehat{\text{disc}}_{H^T}(\mathbf{q}) + \mu\eta + 2\alpha + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{\mathbb{E}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where

$$\widehat{\text{disc}}_H(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[L(h_t(X_{T+1}), h(X_{T+1})) - L(h_t, Z_t) \right] \right|.$$

Proof Sketch

$$\begin{aligned}
\text{disc}(\mathbf{q}) &= \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \right| \\
&\leq \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathbb{E} \left[L(h_t(X_{T+1}), h^*(X_{T+1})) \mid \mathbf{Z}_1^T \right] \right] \right| \\
&\quad + \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[\mathbb{E} \left[L(h_t(X_{T+1}), h^*(X_{T+1})) \mid \mathbf{Z}_1^T \right] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \right| \\
&\leq \mu \sup_{\mathbf{h} \in H_{\mathcal{A}}} \sum_{t=1}^T q_t \mathbb{E} \left[L(h^*(X_{T+1}), Y_{T+1}) \mid \mathbf{Z}_1^T \right] \\
&\quad + \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[\mathbb{E} \left[L(h_t(X_{T+1}), h^*(X_{T+1})) \mid \mathbf{Z}_1^T \right] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \right| \\
&= \mu \sup_{\mathbf{h} \in H_{\mathcal{A}}} \mathbb{E} \left[L(h^*(X_{T+1}), Y_{T+1}) \mid \mathbf{Z}_1^T \right] \\
&\quad + \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^T q_t \left[\mathbb{E} \left[L(h_t(X_{T+1}), h^*(X_{T+1})) \mid \mathbf{Z}_1^T \right] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \right|.
\end{aligned}$$

Learning Guarantee

- **Lemma:** let L be a convex loss bounded by M and \mathbf{h}_1^T a hypothesis sequence adapted to \mathbf{Z}_1^T . Fix $\mathbf{q} \in \Delta$. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$ for the hypothesis $h = \sum_{t=1}^T q_t h_t$:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t L(h_t, Z_t) + \text{disc}(\mathbf{q}) + M \|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

Proof

- By definition of the on-line discrepancy,

$$\sum_{t=1}^T q_t \left[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \leq \text{disc}(\mathbf{q}).$$

- $A_t = q_t \left[\mathcal{L}_t(h_t, Z_1^{t-1}) - L(h_t, Z_t) \right]$ is a martingale difference, thus by Azuma's inequality, whp,

$$\sum_{t=1}^T q_t \mathcal{L}_t(h_t, Z_1^{t-1}) \leq \sum_{t=1}^T q_t L(h_t, Z_t) + \|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

- By convexity of the loss:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T).$$

Learning Guarantee

- **Theorem:** let L be a convex loss bounded by M and H^* a set of hypothesis sequences adapted to \mathbf{Z}_1^T . Fix $\mathbf{q} \in \Delta$. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$ for the hypothesis $h = \sum_{t=1}^T q_t h_t$:

$$\begin{aligned} & \mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \\ & \leq \inf_{\mathbf{h}^* \in H} \sum_{t=1}^T \mathcal{L}_{T+1}(h^*, \mathbf{Z}_1^T) + 2\text{disc}(\mathbf{q}) + \frac{\text{Reg}_T}{T} \\ & \quad + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}}. \end{aligned}$$

Conclusion

- Time series forecasting:
 - key learning problem in many important tasks.
 - very challenging: theory, algorithms, applications.
 - new and general data-dependent learning guarantees for non-mixing non-stationary processes.
 - algorithms with guarantees.
- Time series prediction and on-line learning:
 - proof for flexible solutions derived via OTB.
 - application to model selection.
 - application to learning ensembles.

References

- T. M. Adams and A. B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367, 2010.
- A. Agrawal, J. Duchi. The generalization ability of online algorithms for dependent data. *Information Theory, IEEE Transactions on*, 59(1):573–587, 2013.
- P. Alquier, X. Li, O. Wintenberger. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modelling*, 1:65–93, 2014.
- Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. *COLT*, 2013.

References

- O. Anava, E. Hazan, and A. Zeevi. Online time series prediction with missing data. *ICML*, 2015.
- O. Anava and S. Mannor. Online Learning for heteroscedastic sequences. *ICML*, 2016.
- P. L. Bartlett. Learning with a slowly changing distribution. *COLT*, 1992.
- R. D. Barve and P. M. Long. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):101–123, 1997.

References

- P. Berti and P. Rigo. A Glivenko-Cantelli theorem for exchangeable random variables. *Statistics & Probability Letters*, 32(4):385 – 391, 1997.
- S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NIPS*. 2007.
- T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J Econometrics*, 1986.
- G. E. P. Box, G. Jenkins. (1990) . *Time Series Analysis, Forecasting and Control*.

References

- O. Bousquet and M. K. Warmuth. Tracking a small set of experts by mixing past posteriors. *COLT*, 2001.
- N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. on Inf. Theory* , 50(9), 2004.
- N. Cesa-Bianchi and C. Gentile. Tracking the best hyperplane with a simple budget perceptron. *COLT*, 2006.
- C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519, 2014.

References

- V. H. De la Pena and E. Gine. (1999) *Decoupling: from dependence to independence: randomly stopped processes, U-statistics and processes, martingales and beyond. Probability and its applications*. Springer, NY.
- P. Doukhan. (1994) *Mixing: properties and examples. Lecture notes in statistics*. Springer-Verlag, New York.
- R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

References

- D. P. Helmbold and P. M. Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14(1): 27-46, 1994.
- M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2), 1998.
- M. Herbster and M. K. Warmuth. Tracking the best linear predictor. *JMLR*, 2001.
- D. Hsu, A. Kontorovich, and C. Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. *NIPS*, 2015.

References

- W. M. Koolen, A. Malek, P. L. Bartlett, and Y. Abbasi. Minimax time series prediction. *NIPS*, 2015.
- V. Kuznetsov, M. Mohri. Generalization bounds for time series prediction with non-stationary processes. In *ALT*, 2014.
- V. Kuznetsov, M. Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *NIPS*, 2015.
- V. Kuznetsov, M. Mohri. Time series prediction and on-line learning. In *COLT*, 2016.

References

- A. C. Lozano, S. R. Kulkarni, and R. E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β -mixing observations. In *NIPS*, pages 819–826, 2006.
- Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*. 2009.
- R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, pages 5–34, 2000.

References

- D. Modha, E. Masry. Memory-universal prediction of stationary random processes. *Information Theory, IEEE Transactions on*, 44(1):117–133, Jan 1998.
- M. Mohri, A. Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *ALT*, 2012.
- M. Mohri, A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *NIPS*, 2009.
- M. Mohri, A. Rostamizadeh. Stability bounds for stationary φ -mixing and β -mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

References

- V. Pestov. Predictive PAC learnability: A paradigm for learning from exchangeable input data. *GRC*, 2010.
- A. Rakhlin, K. Sridharan, A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010.
- L. Ralaivola, M. Szafranski, G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary beta-mixing processes. *JMLR* 11:1927–1956, 2010.
- C. Shalizi and A. Kontorovitch. Predictive PAC-learning and process decompositions. *NIPS*, 2013.

References

- A. Rakhlin, K. Sridharan, A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *NIPS*, 2010.
- I. Steinwart, A. Christmann. Fast learning from non-i.i.d. observations. *NIPS*, 2009.
- M. Vidyasagar. (1997). *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer-Verlag New York, Inc.
- V. Vovk. Competing with stationary prediction strategies. *COLT* 2007.

References

- B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.