Online Learning for Time Series Prediction

Joint work with Vitaly Kuznetsov (Google Research)

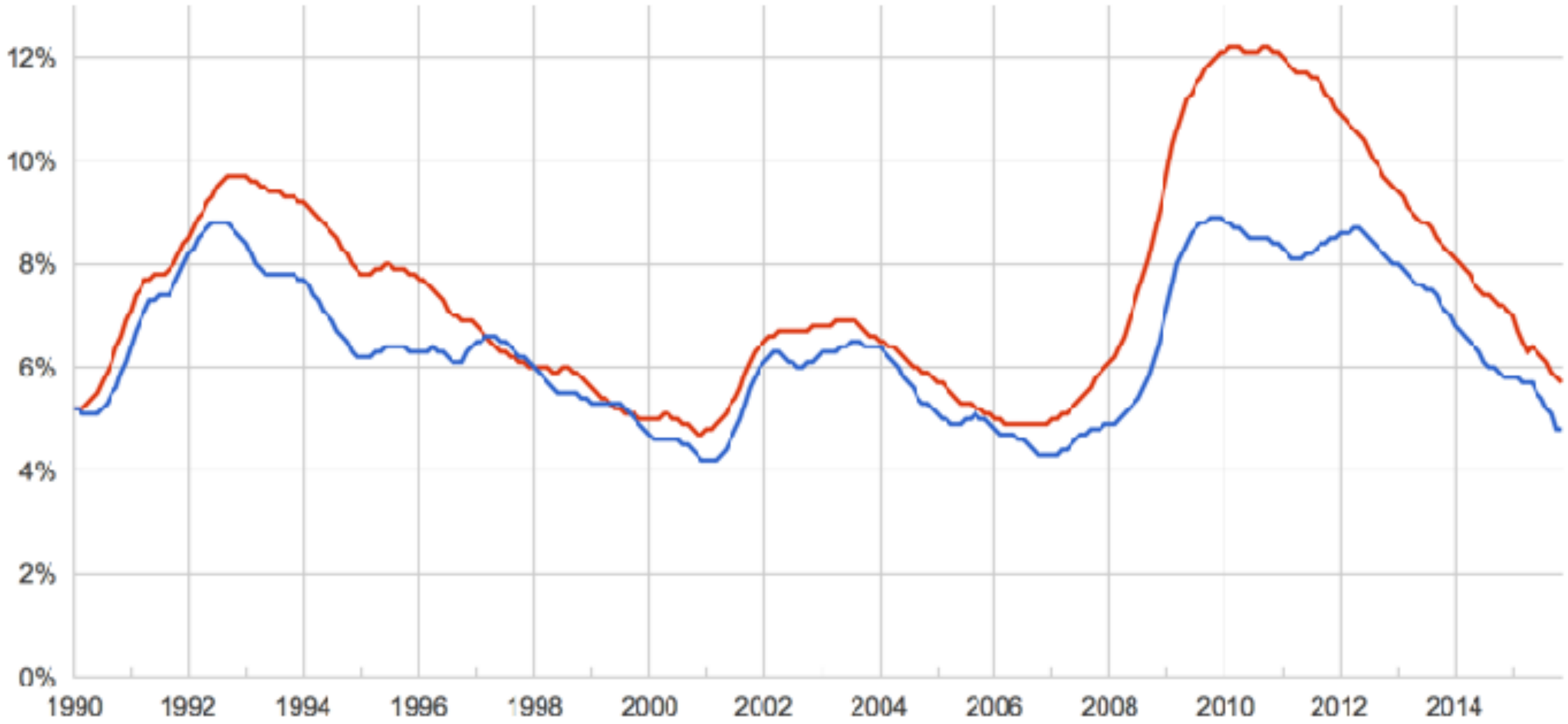
MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Motivation

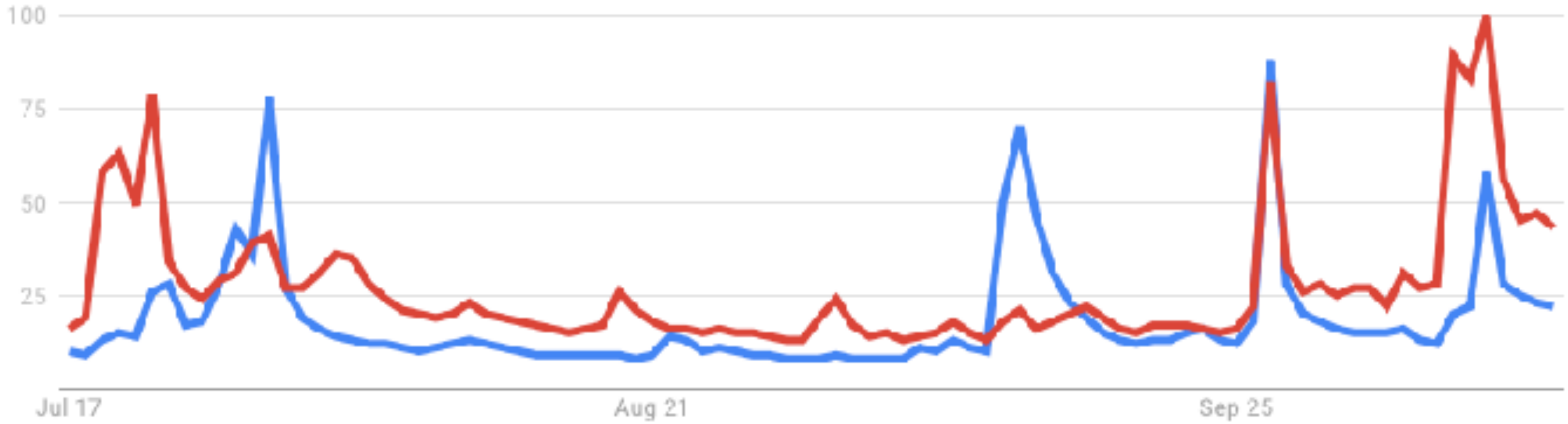
- Time series prediction:
  - stock values.
  - economic variables.
  - weather: e.g., local/global temperature.
  - earthquakes.
  - energy demand.
  - signal processing.
  - sales forecasting.
  - election forecast.

Time Series



NY Unemployment Rate

Time Series



US Presidential Election 2016

Online Learning

- Advantages:
  - active research area (Cesa-Bianchi and Lugosi, 2006).
  - no distributional assumption.
  - algorithms with tight regret guarantees.
  - flexibility: e.g., non-static competitor classes (Herbster and Warmuth, 1998, 2001; Koolen et al., 2015; Rakhlin and Sridharan 2015).

Online Learning

- Drawbacks:
  - real-world time series data is not adversarial.
  - the stochastic process must be taken into account.
  - the quantity of interest is the conditional expected loss, not the regret.

→ can we leverage online algorithms for time series forecasting?

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

$$\mathrm{Reg}_T = \sum_{t=1}^{T} L(h_t, Z_t) - \inf_{\mathbf{h}^* \in \mathbf{H}^*} \sum_{t=1}^{T} L(\mathbf{h}^*, Z_t).$$

On-Line-to-Batch Problem

- Problem: use $\mathbf{h} = (h_1, \ldots, h_T)$ to derive a hypothesis $h \in H$ with small path-dependent expected loss,

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathop{\mathbb{E}}_{Z_{T+1}} \big[L(h, Z_{T+1}) | \mathbf{Z}_1^T \big].$$

  - IID case is standard: (Littlestone, 1989), (Cesa-Bianchi et al., 2004).
  - how do we handle general stochastic processes?

Previous Work

(Kuznetsov and MM, 2015)

- Theory and algorithms for time series prediction:
  - general non-stationary non-mixing stochastic processes.
  - generalization bounds based on a notion of discrepancy.
  - convex optimization algorithms.
  - algorithms perform well in experiments.

→ But, how do we tackle some difficult time series problems such as ensemble learning or model selection?

Questions

- Theoretical:
  - can we derive learning guarantees for a convex combination $\sum_{t=1}^{T} q_t h_t$?
  - can we derive guarantees for h selected in $(h_1, \ldots, h_T)$?
- Algorithmic:
  - on-line-to-batch conversion.
  - learning ensembles.
  - model selection.

Theory

Learning Guarantee

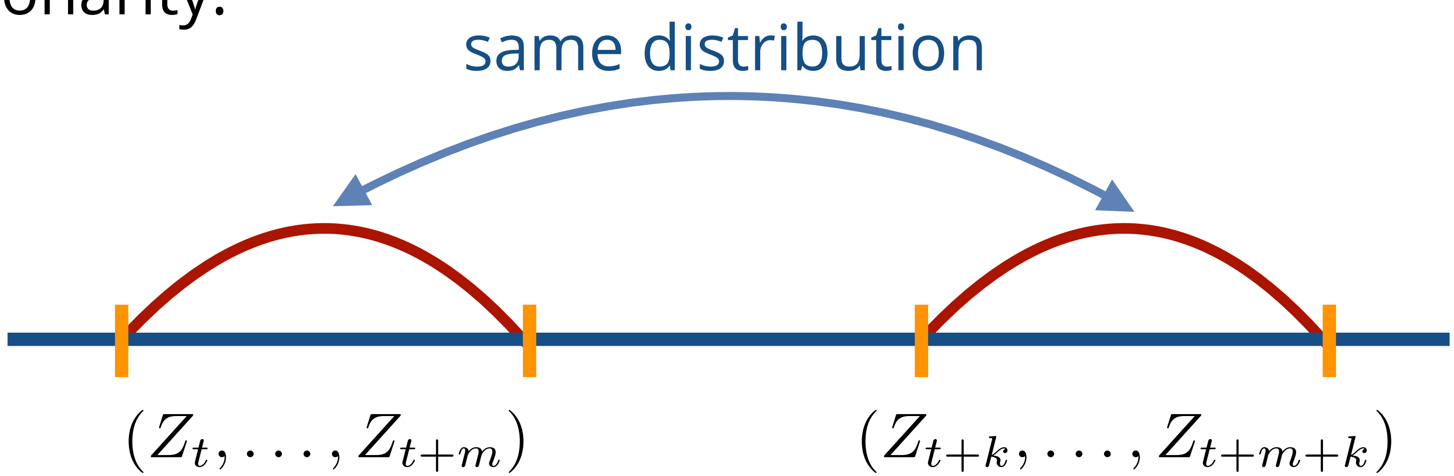
- Problem: given hypotheses $(h_1, \ldots, h_T)$ give bound on

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathop{\mathbb{E}}_{Z_{T+1}} \big[L(h, Z_{T+1}) | \mathbf{Z}_1^T\big],$$

where $h = \sum_{t=1}^T q_t h_t$.

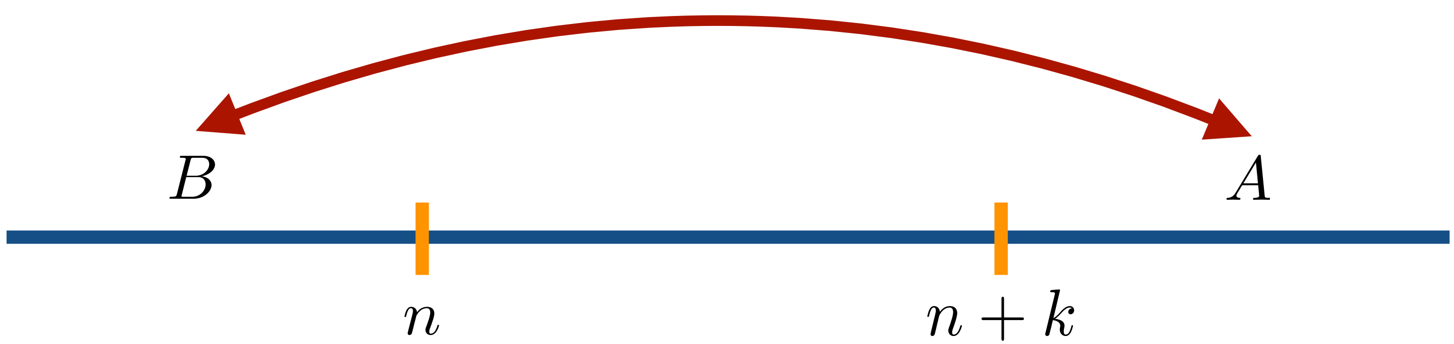
Standard Assumptions

- Stationarity:



- Mixing:

dependence between events decaying with k.

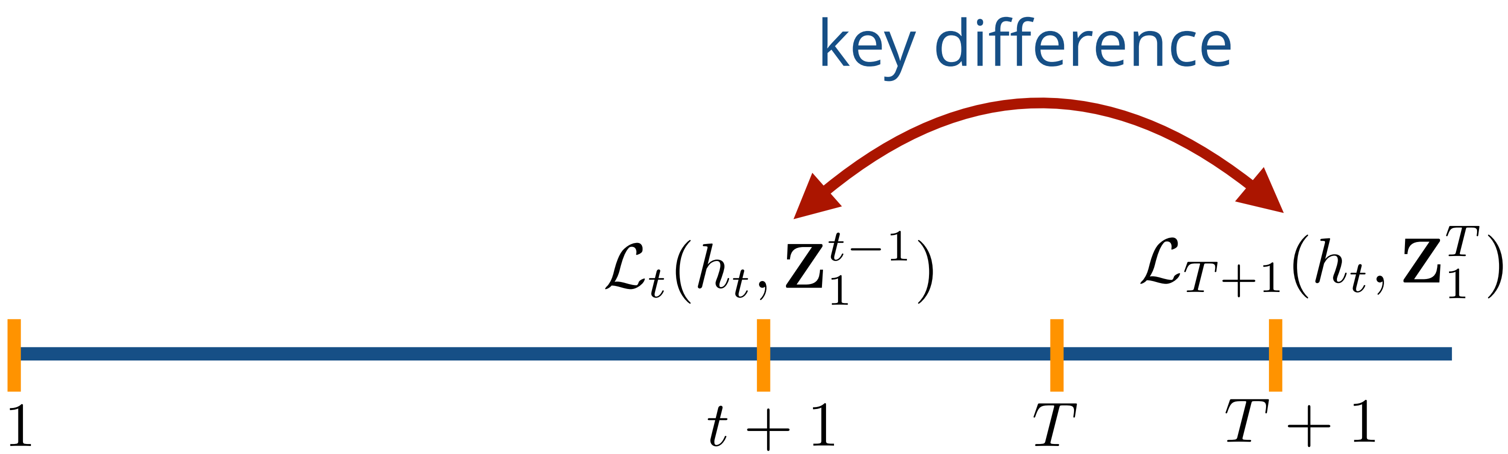


Problem

- Stationarity and mixing assumptions:
  - widely adopted: (Alquier and Wintenberger, 2010, 2014), (Agarwal and Duchi, 2013), (Lozano et al., 1997), (Vidyasagar, 1997), (Yu, 1994), (Meir, 2000), (MM and Rostamizadeh, 2000), (Kuznetsov and MM, 2014).
- But,
  - they **often do not hold** (think trend or periodic signals).
  - they are not testable.
  - estimating mixing parameters can be hard, even if general functional form known.
  - hypothesis set and loss function ignored.

→ we need a new tool for the analysis.

Relevant Quantity



→ Average difference: $\frac{1}{T}\sum_{t=1}^{T}\Big[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^{T}) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1})\Big].$

On-line Discrepancy

- Definition:

$$\mathrm{disc}(\mathbf{q}) = \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \Big] \right|.$$

  - $\mathbf{H}_{\mathcal{A}}$: sequences that $\mathcal{A}$ can return.
  - $\mathbf{q} = (q_1, \ldots, q_T)$: 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) .

Learning Guarantee

- Theorem: 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 $h = \sum_{t=1}^T q_t h_t$:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t L(h_t, Z_t) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

Proof

- By convexity of the loss:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T).$$

- By definition of the on-line discrepancy,

$$\sum_{t=1}^{T} q_t \Big[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \Big] \leq \operatorname{disc}(\mathbf{q}).$$

- $A_t = q_t \Big[\mathcal{L}_t(h_t, Z_1^{t-1}) - L(h_t, Z_t) \Big]$ is a martingale difference, thus by Azuma's inequality, whp,

$$\sum_{t=1}^{T} q_t \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \leq \sum_{t=1}^{T} q_t L(h_t, Z_t) + \|\mathbf{q}\|_2 \sqrt{2 \log \tfrac{1}{\delta}}.$$

Learning Guarantee

- Theorem: let L be a convex loss bounded by M and $\mathbf{H}^*$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 $h = \sum_{t=1}^T q_t h_t$:

$$
\begin{aligned}
&\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \\
&\quad \leq \inf_{\mathbf{h}^* \in H} \sum_{t=1}^T \mathcal{L}_{T+1}(h^*, \mathbf{Z}_1^T) + 2\mathrm{disc}(\mathbf{q}) + \frac{\mathrm{Reg}_T}{T} \\
&\quad + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{2}{\delta}}.
\end{aligned}
$$

Notes

- Learning guarantees with same flavor as those of (Kuznetsov and MM, 2015) but simpler proofs, no complexity measure.
- Bounds admit as special case the learning guarantees for
  - the i.i.d. scenario (Littlestone, 1989), (Cesa-Bianchi et al., 2004).
  - the drifting scenario (MM and Muñoz Medina, 2012).

Extension: Non-Convex Loss

- Theorems extend to non-convex losses when h is selected as follows (Cesa-Bianchi et al., 2004):

$$h = \operatorname*{argmin}_{h_t} \left\{ \sum_{s=t}^{T} q_s L(h_t, Z_s) + \operatorname{disc}(\mathbf{q}_t^T) + M \|\mathbf{q}_t^T\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}} \right\}.$$

Extension: General Regret

- General regret definition:

$$\text{Reg}_T = \sum_{t=1}^{T} L(h_t, Z_t) - \inf_{\mathbf{h}^* \in \mathbf{H}^*} \Big\{ \sum_{t=1}^{T} L_t(\mathbf{h}^*, Z_t) + \mathcal{R}(\mathbf{h}^*) \Big\}.$$

  - standard regret: $\mathcal{R} = 0$, $\mathbf{H}^*$ constant sequences.
  - tracking: $\mathbf{H}^* \subseteq H^T$.
  - $\mathcal{R}$ can be a kernel-based regularization (Herbster and Warmuth, 2001).

Extension: Non-Adapted Seqs

- Hypotheses no longer adapted, but output by a uniformly stable algorithm.
- Stable hypotheses:
  - $\mathcal{H} = \{h \in H \colon \text{ there exists } \mathcal{A} \in \mathfrak{A} \text{ such that } h = \mathcal{A}(\mathbf{Z}_1^T)\}.$
  - $\beta_t = \beta_{h_t}$: stability coefficient of algorithm returning h_t .
- Similar learning bounds with additional term $\sum_{t=1}^T q_t \beta_t$.
  - admit as special cases results of (Agarwal and Duchi, 2013) for asymptotically stationary mixing processes.

Discrepancy Estimation

- Batch estimation method (Kuznetsov and MM, 2015).
- On-line estimation method:
  - assume that the loss is μ-Lipschitz.
  - assume that there exists an accurate hypothesis h^*:

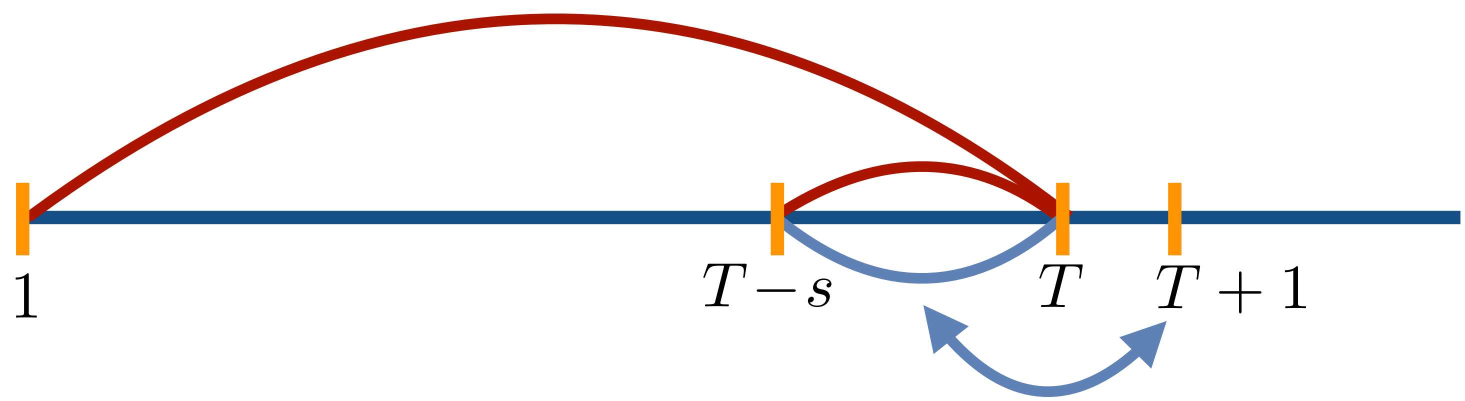
$$\inf_{h^*} \mathbb{E}\left[L(Z_{T+1}, h^*(X_{T+1})) | \mathbf{Z}_1^T\right] \ll 1.$$

Batch Estimation

- Decomposition: $\Delta(\mathbf{q}) \leq \Delta_0(\mathbf{q}) + \Delta_s$.

$$\Delta(\mathbf{q}) \leq \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^{T} \mathcal{L}(h, \mathbf{Z}_1^{t-1}) - \sum_{t=1}^{T} q_t \, \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right)$$

$$+ \sup_{h \in H} \left(\mathcal{L}(h, \mathbf{Z}_1^{T}) - \frac{1}{s} \sum_{t=T-s+1}^{T} \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right).$$



Online Estimation

- Lemma: assume that L is μ-Lipschitz. 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 $\alpha > 0$:

$$\text{disc}(\mathbf{q}) \leq \widehat{\text{disc}}_{\mathbf{H}_\mathcal{A}}(\mathbf{q}) + \mu \inf_{h^*} \mathbb{E}\left[L(Z_{T+1}, h^*(X_{T+1})) | \mathbf{Z}_1^T\right]$$
$$+ 2\alpha + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{\mathbb{E}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where

$$\widehat{\text{disc}}_{\mathbf{H}_\mathcal{A}}(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H_\mathcal{A}} \left| \sum_{t=1}^{T} q_t \Big[L\big(h_t(X_{T+1}), h(X_{T+1})\big) - L\big(h_t, Z_t\big) \Big] \right|.$$

Proof Sketch

$$\mathrm{disc}(\mathbf{q}) = \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \Big] \right|$$

$$\leq \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathbb{E}\left[L(h_t(X_{T+1}), h^*(X_{T+1})) \middle| \mathbf{Z}_1^T \right] \right] \right|$$

$$+ \sup_{\mathbf{h} \in \mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \left[\mathbb{E}\left[L(h_t(X_{T+1}), h^*(X_{T+1})) \middle| \mathbf{Z}_1^T \right] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right] \right|.$$

$$\leq \mu \sup_{\mathbf{h} \in H_{\mathcal{A}}} \sum_{t=1}^{T} q_t \, \mathbb{E}\left[L(h^*(X_{T+1}), Y_{T+1}) \middle| \mathbf{Z}_1^T \right]$$

$$= \mu \sup_{\mathbf{h} \in H_{\mathcal{A}}} \mathbb{E}\left[L(h^*(X_{T+1}), Y_{T+1}) \middle| \mathbf{Z}_1^T \right].$$

$\mathrm{disc}_{\mathbf{H}_{\mathcal{A}}}(\mathbf{q})$

Algorithms

On-Line-to-Batch (OTB)

- Problem: use $\mathbf{h} = (h_1, \ldots, h_T)$ to derive a hypothesis $h \in H$ with small path-dependent expected loss,

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathop{\mathbb{E}}_{Z_{T+1}} \big[L(h, Z_{T+1}) | \mathbf{Z}_1^T \big].$$

OTB Algorithm

- Idea: choose weights $\mathbf{q}$ to minimize bound for $h = \sum_{t=1}^T q_t h_t$:

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^T q_t L(h_t, Z_t) + \mathrm{disc}(\mathbf{q}) + M \|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

- Optimization problem:

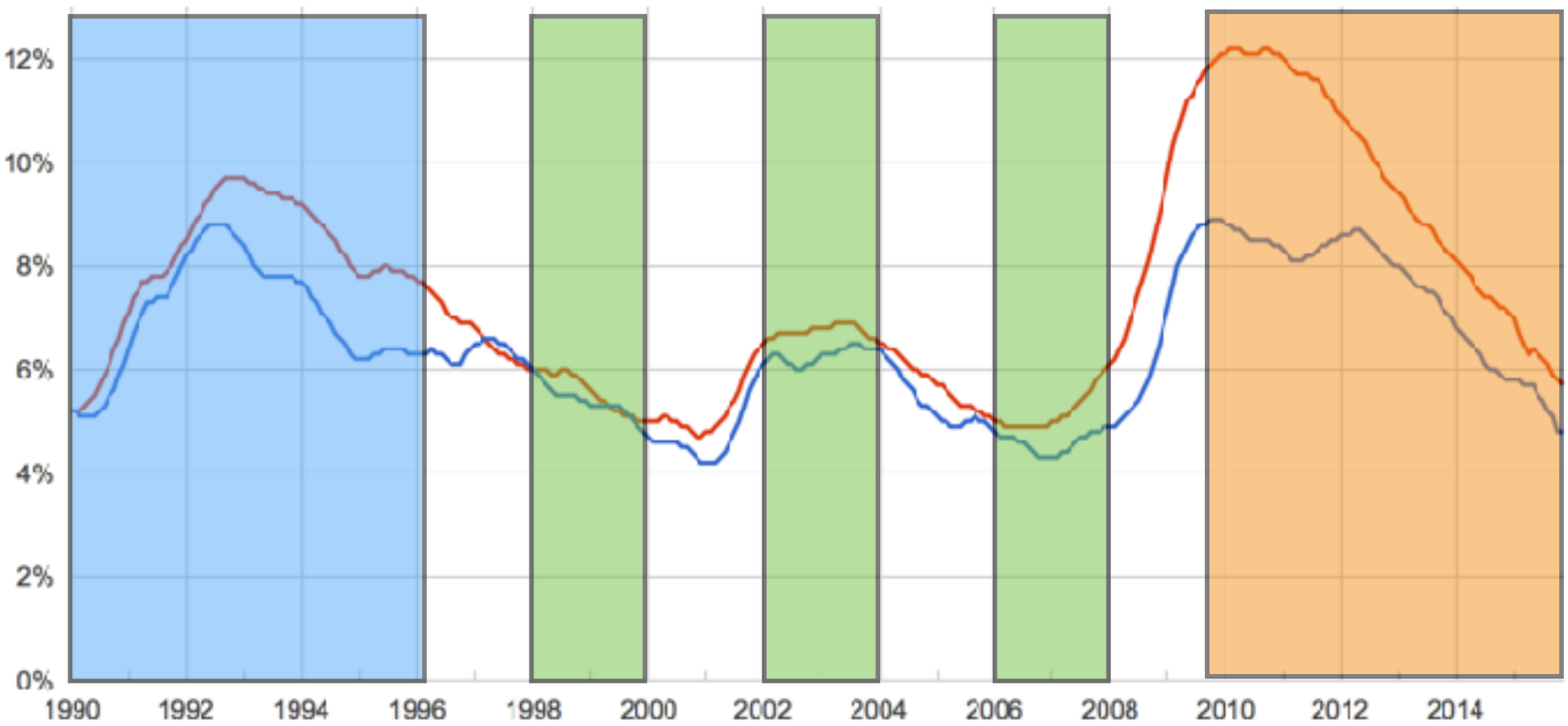
$$\min_{\mathbf{q} \in \Delta} \sum_{t=1}^T q_t L(h_t, Z_t) + \widehat{\mathrm{disc}}_{H_{\mathcal{A}}}(\mathbf{q}) + \lambda \|\mathbf{q}\|_2.$$

- Solution: $h = \sum_{t=1}^T q_t h_t$.

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?
  - models may have been pre-trained on $\mathbf{Z}_1^T$.

Model Selection



Model Selection Algorithm

- Idea: use learning bound in terms of online discrepancy and regret per round for hypothesis

$$h = \operatorname*{argmin}_{h_t} \left\{ \sum_{s=t}^{T} q_s L(h_t, Z_s) + \operatorname{disc}(\mathbf{q}_t^T) + M \|\mathbf{q}_t^T\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}} \right\}.$$

Model Selection Algorithm

- Algorithm:
  - choose $\mathbf{q} \in \Delta$ to minimize discrepancy

$$\min_{\mathbf{q} \in \Delta} \widehat{\mathrm{disc}}_H(\mathbf{q}).$$

  - use on-line algorithm for prediction with expert advice to generate a sequence of hypotheses $\mathbf{h} \in \mathcal{H}^T$, with $\mathcal{H}$ the set of N models.
  - select model according to

$$h = \operatorname*{argmin}_{h_t} \left\{ \sum_{s=t}^{T} q_s L(h_t, Z_s) + \mathrm{disc}(\mathbf{q}_t^T) + M \|\mathbf{q}_t^T\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}} \right\}.$$

Learning Ensembles

- Problem: given a hypothesis set H and a sample $\mathbf{Z}_1^T$, find convex combination $h = \sum_{t=1}^T q_t h_t$ with $\mathbf{h} \in H_{\mathcal{A}}$ with small path-dependent expected loss.
  - in most general case, hypotheses may have been pre-trained on $\mathbf{Z}_1^T$.

Ensemble Learning Algorithm

- Algorithm:
  - run regret minimization on $\mathbf{Z}_1^T$ to return $\mathbf{h} = (h_1, \ldots, h_T)$.
  - minimize learning bound. For $\Lambda_2 \geq 0$,

$$\min_{\mathbf{q}} \quad \widehat{\mathrm{disc}}_{H_{\mathcal{A}}}(\mathbf{q}) + \sum_{t=1}^{T} q_t L(h_t, Z_t)$$
$$\text{subject to} \quad \|\mathbf{q} - \mathbf{u}\|_2 \leq \Lambda_2.$$

- convex optimization problem by convexity of

$$\widehat{\mathrm{disc}}_{H_{\mathcal{A}}}(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[L\big(h_t(X_{T+1}), h(X_{T+1})\big) - L\big(h_t, Z_t\big) \Big] \right|.$$

Ensemble Learning Algorithm

- If hypothesis set H is finite, then the supremum can be computed straightforwardly:

- If hypothesis set is not finite but is convex and the loss is convex, then the maximization can be cast as a DC-programming problem, and solved using the DC-algorithm (Tao and An, 1998):

$$\sup_{h\in H, \mathbf{h}\in H_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[L\big(h_t(X_{T+1}), h(X_{T+1})\big) - L\big(h_t, Z_t\big) \Big] \right|.$$

- for squared loss, global optimum.

Conclusion

- Time series prediction using on-line algorithms:
  - new learning bounds for non-stationary non-mixing processes.
  - on-line discrepancy measure that can be estimated.
  - general on-line-to-batch conversion.
  - application to model selection.
  - application to learning ensembles.
  - tools for tackling other time series problems.