Time Series Prediction & Online Learning

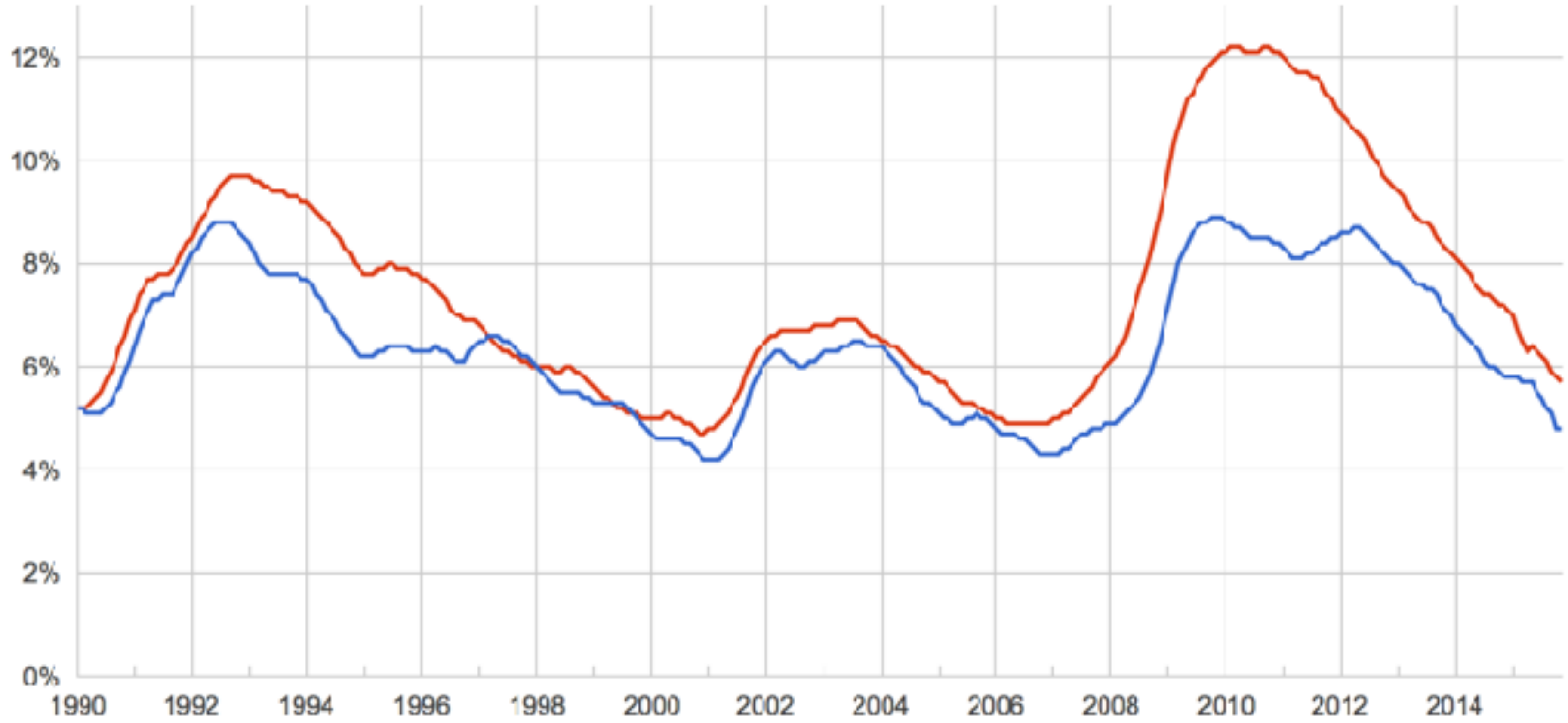
Joint work with Vitaly Kuznetsov (Google Research)

MEHRYAR MOHRI MOHRI@
COURANT INSTITUTE & GOOGLE RESEARCH

Motivation

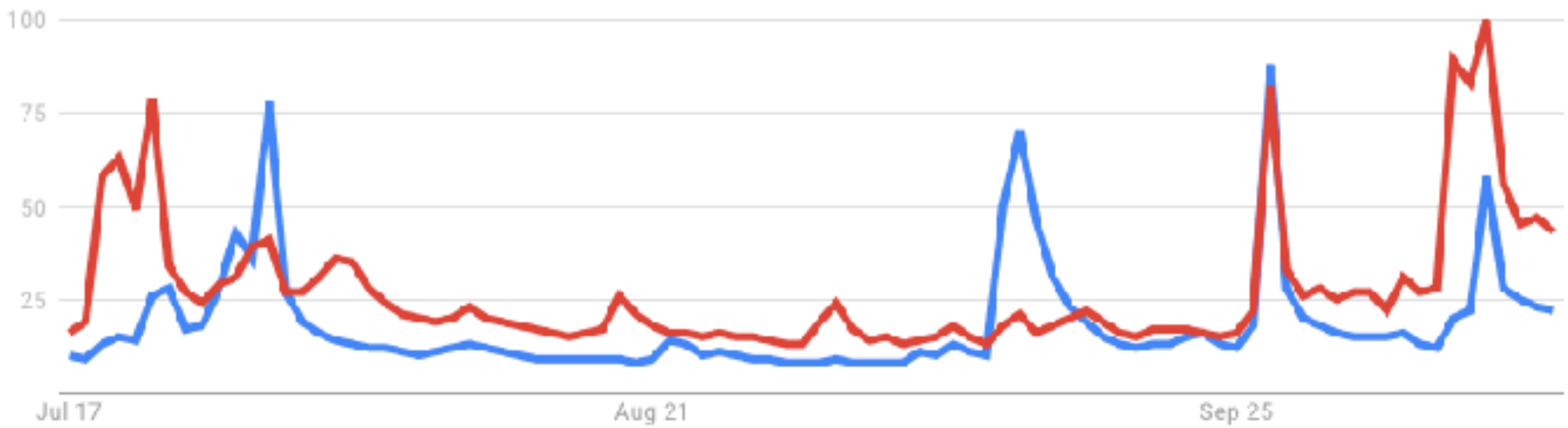
- Time series prediction:
  - stock values.
  - economic variables.
  - weather: e.g., local/global temperature.
  - earthquakes.
  - energy demand.
  - signal processing.
  - sales forecasting.
  - election forecast.

Time Series



NY Unemployment Rate

Time Series



US Presidential Election 2016

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

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

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

Example: Exp. Weights (EW)

- Expert set $H^* = \{\mathcal{E}_1, \ldots, \mathcal{E}_N\}$, $H = \text{conv}(H^*)$.

$\text{EW}(\{\mathcal{E}_1, \ldots, \mathcal{E}_N\})$

for $i \leftarrow 1$ **to** N **do**

    $w_{1,i} \leftarrow 1$

for $t \leftarrow 1$ **to** T **do**

    $\textsc{Receive}(x_t)$

    $h_t \leftarrow \frac{\sum_{i=1}^N w_{t,i} \mathcal{E}_i}{\sum_{i=1}^N w_{t,i}}$

    $\textsc{Receive}(y_t)$

    $\textsc{Incur-Loss}(L(h_t(x_t), y_t))$

    for $i \leftarrow 1$ **to** N **do**

        $w_{t+1,i} \leftarrow w_{t,i}\, e^{-\eta L(\mathcal{E}_i(x_t), y_t)}$ $\triangleright$ (parameter $\eta > 0$)

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 $y_1, \ldots, y_T \in \mathcal{Y}$, the regret of EW at time T satisfies

$$\text{Reg}_T \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.$$

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

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

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

EW - Proof

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

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \log \frac{\sum_{i=1}^N w_{t-1,i}\, e^{-\eta L(\mathcal{E}_i(x_t), y_t)}}{\sum_{i=1}^N w_{t-1,i}} \\
&= \log \Big(\mathop{\mathbb{E}}_{w_{t-1}} [e^{-\eta L(\mathcal{E}_i(x_t), y_t)}] \Big) \\
&= \log \bigg(\mathop{\mathbb{E}}_{w_{t-1}} \bigg[\exp \bigg(-\eta \Big(L(\mathcal{E}_i(x_t), y_t) - \mathop{\mathbb{E}}_{w_{t-1}} [L(\mathcal{E}_i(x_t), y_t)] \Big) - \eta \mathop{\mathbb{E}}_{w_{t-1}} [L(\mathcal{E}_i(x_t), y_t)] \bigg) \bigg] \bigg) \\
&\leq -\eta \mathop{\mathbb{E}}_{w_{t-1}} [L(\mathcal{E}_i(x_t), y_t)] + \frac{\eta^2}{8} \qquad \text{(Hoeffding's ineq.)} \\
&\leq -\eta L\Big(\mathop{\mathbb{E}}_{w_{t-1}} [\mathcal{E}_i(x_t)], y_t \Big) + \frac{\eta^2}{8} \qquad \text{(convexity of first arg. of } L) \\
&= -\eta L(h_t(x_t), y_t) + \frac{\eta^2}{8}.
\end{aligned}$$

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 stochastic and on-line results? Can we tackle notoriously difficult time series problems?
  - on-line-to-batch conversion.
  - model selection.
  - learning ensembles.

On-line-to-Batch Conversion

On-Line-to-Batch (OTB)

- Input: sequence of hypotheses $\mathbf{h} = (h_1, \ldots, h_T)$ returned after T rounds by an on-line algorithm $\mathcal{A}$ minimizing general regret

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

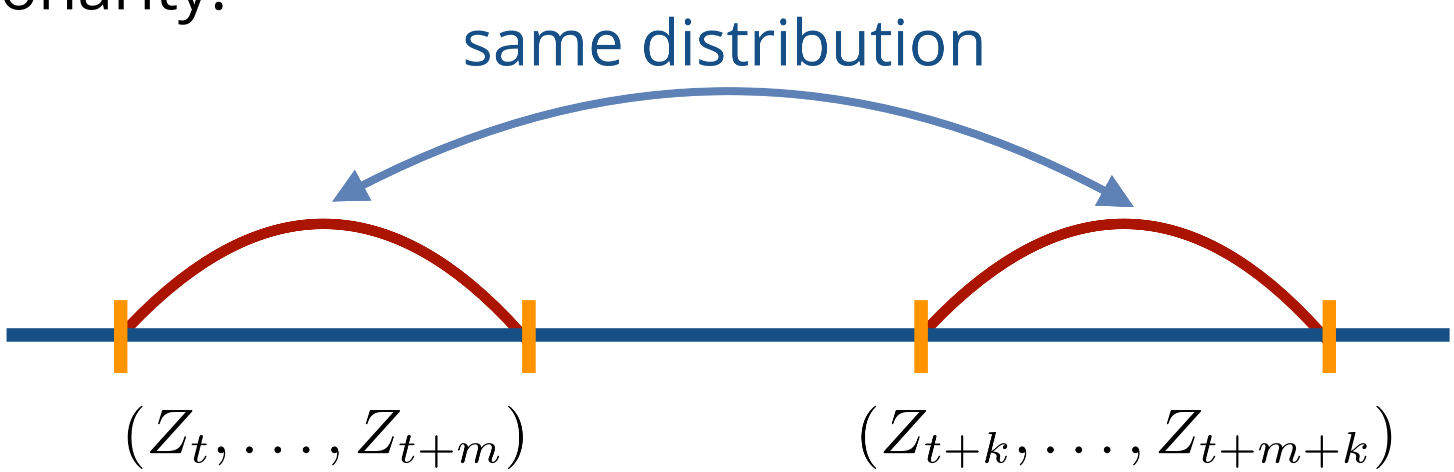
- Problem: use $\mathbf{h} = (h_1, \ldots, h_T)$ to derive a hypothesis $h \in H$ with small path-dependent expected loss,

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

  - IID case is standard: (Littlestone, 1989), (Cesa-Bianchi et al., 2004).
  - general stochastic process?

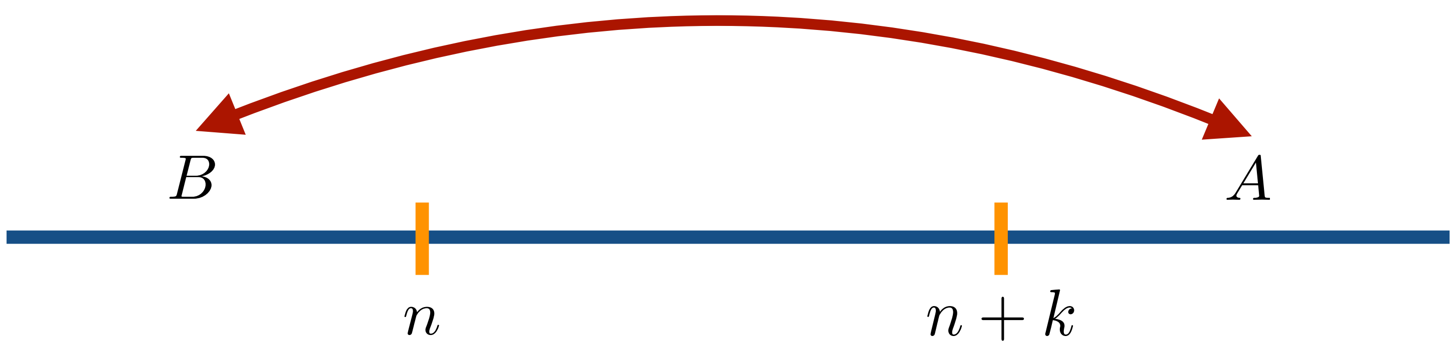
Standard Assumptions

- Stationarity:



- Mixing:

dependence between events decaying with k.

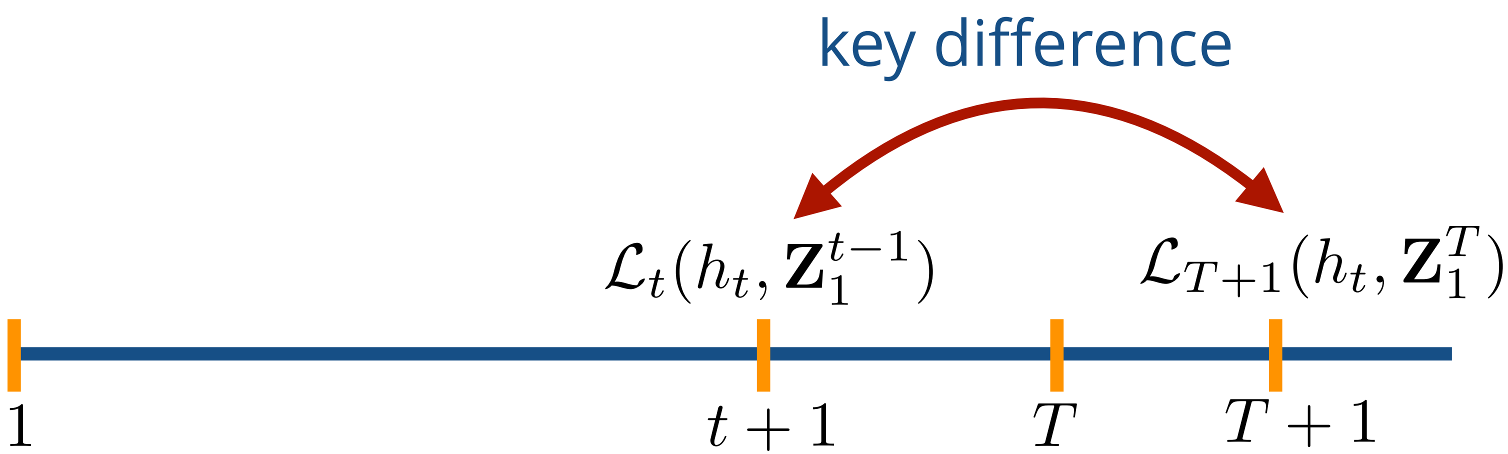


Problem

- 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:

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

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

Discrepancy Estimation

- Batch discrepancy estimation method (Kuznetsov and MM, 2015).
- 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 $\alpha > 0$:

$$\mathrm{disc}(\mathbf{q}) \leq \widehat{\mathrm{disc}}_{H^T}(\mathbf{q}) + \mu\eta + 2\alpha + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{\mathbb{E}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]}{\delta}},$$

where

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

Proof Sketch

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

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

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

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

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

$$\widehat{\text{disc}}_H(\mathbf{q})$$

Lemma

- 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 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,

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

- $A_t = q_t \Big[\mathcal{L}_t(h_t, Z_1^{t-1}) - L(h_t, Z_t)\Big]$ 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

- Theorem extends to non-convex losses when h is selected as follows:

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

- Learning guarantees with same flavor as those of (Kuznetsov and MM, 2015) but simpler proofs, no complexity measure.
- They 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

- General regret definition:

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

  - 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).

Stable Hypothesis Sequences

- 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.

Applications

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 $\mathbf{Z}_1^T$.

Model Selection

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

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

  - 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

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

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 $\mathbf{h} \in H_{\mathcal{A}}$ and $\mathbf{q} \in \Delta$.
    - in most general case, hypotheses may have been pre-trained on $\mathbf{Z}_1^T$.

Learning Ensembles

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

$$\begin{aligned} \min_{\mathbf{q}} \quad & \widehat{\operatorname{disc}}_H(\mathbf{q}) + \sum_{t=1}^{T} q_t L(h_t, Z_t) \\ \text{subject to} \quad & \|\mathbf{q} - \mathbf{u}\|_2 \leq \Lambda_2. \end{aligned}$$

  - for convex loss and convex H, can be cast as a DC-programming problem, and solved using the DC-algorithm (Tao and An, 1998).
  - 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.