Time Series Prediction & Online Learning

Joint work with Vitaly Kuznetsov (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.
On-Line Learning
On-Line Learning Setup

- Adversarial setting with hypothesis/action set $H$.
- For $t = 1$ to $T$ do
  - player receives $x_t \in X$.
  - 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

\[
\text{Reg}_T = \sum_{t=1}^{T} L(h_t(x_t), y_t) - \min_{h \in H^*} \sum_{t=1}^{T} L(h(x_t), y_t).
\]
Expert set $H^* = \{ \mathcal{E}_1, \ldots, \mathcal{E}_N \}$, $H = \text{conv}(H^*)$.

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

1. for $i \leftarrow 1$ to $N$ do
2. \hspace{1em} $w_{1,i} \leftarrow 1$
3. for $t \leftarrow 1$ to $T$ do
4. \hspace{1em} RECEIVE($x_t$)
5. \hspace{2em} $h_t \leftarrow \frac{\sum_{i=1}^{N} w_{t,i} \mathcal{E}_i}{\sum_{i=1}^{N} w_{t,i}}$
6. \hspace{1em} RECEIVE($y_t$)
7. INCUR-LOSS($L(h_t(x_t), y_t)$)
8. for $i \leftarrow 1$ to $N$ do
9. \hspace{2em} $w_{t+1,i} \leftarrow w_{t,i} e^{-\eta L(\mathcal{E}_i(x_t), y_t)}$ \hfill $\triangleright$ (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 $y_1, \ldots, y_T \in \mathcal{Y}$, the regret of EW at time $T$ satisfies

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

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

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

\[
\frac{\text{Reg}_T}{T} = O\left(\sqrt{\frac{\log N}{T}}\right).
\]
**EW - Proof**

- **Potential:** \( \Phi_t = \log \sum_{i=1}^{N} w_{t,i} \).

- **Upper bound:**

\[
\Phi_t - \Phi_{t-1} = \log \left( \frac{\sum_{i=1}^{N} w_{t-1,i} e^{-\eta L(\mathcal{E}_i(x_t), y_t)}}{\sum_{i=1}^{N} w_{t-1,i}} \right) \\
= \log \left( \mathbb{E}_{w_{t-1}} \left[ e^{-\eta L(\mathcal{E}_i(x_t), y_t)} \right] \right) \\
= \log \left( \mathbb{E}_{w_{t-1}} \left[ \exp \left( -\eta \left( L(\mathcal{E}_i(x_t), y_t) - \mathbb{E}_{w_{t-1}} \left[ L(\mathcal{E}_i(x_t), y_t) \right] \right) - \eta \mathbb{E}_{w_{t-1}} \left[ L(\mathcal{E}_i(x_t), y_t) \right] \right) \right] \right) \\
\leq -\eta \mathbb{E}_{w_{t-1}} \left[ L(\mathcal{E}_i(x_t), y_t) \right] + \frac{\eta^2}{8} \quad \text{(Hoeffding’s ineq.)} \\
\leq -\eta L \left( \mathbb{E}_{w_{t-1}} \left[ \mathcal{E}_i(x_t) \right], y_t \right) + \frac{\eta^2}{8} \quad \text{(convexity of first arg. of } L) \\
= -\eta L(h_t(x_t), y_t) + \frac{\eta^2}{8}.
\]
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:**

\[
\Phi_T - \Phi_0 = \log \sum_{i=1}^{N} e^{-\eta \sum_{t=1}^{T} L(E_i(x_t), y_t)} - \log N
\geq \log \max_{i=1}^{N} e^{-\eta \sum_{t=1}^{T} L(E_i(x_t), y_t)} - \log N
= -\eta \min_{i=1}^{N} \sum_{t=1}^{T} L(E_i(x_t), y_t) - \log N.
\]

- **Comparison:**

\[
\sum_{t=1}^{T} L(h_t(x_t), y_t) - \min_{i=1}^{N} \sum_{t=1}^{T} L(E_i(x_t), y_t) \leq \frac{\log N}{\eta} + \frac{\eta T}{8}.
\]
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

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

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

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

- **IID case is standard**: (Littlestone, 1989), (Cesa-Bianchi et al., 2004).

- **general stochastic process?**
Standard Assumptions

- **Stationarity:**

  \[(Z_t, \ldots, Z_{t+m})\] \(\sim\) \[(Z_{t+k}, \ldots, Z_{t+m+k})\]

- **Mixing:**

  Dependence between events decaying with \(k\).
Problem

Stationarity and mixing assumptions:


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

key difference

\[ \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \quad \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \]

Average difference:

\[
\frac{1}{T} \sum_{t=1}^{T} \left[ \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) \right].
\]
On-line Discrepancy

**Definition:**

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

- \( \mathcal{H}_A \) : sequences that \( A \) can return.
- \( q = (q_1, \ldots, q_T) \) : 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 \( \text{(Kuznetsov and MM, 2015)} \).
Discrepancy Estimation

- Batch discrepancy estimation method (Kuznetsov and MM, 2015).

- Alternative method:
  - assume that the loss is $\mu$-Lipschitz.
  - assume that there exists an accurate hypothesis $h^*$:

$$\eta = \inf_{h^*} \mathbb{E} \left[ L(Z_{T+1}, h^*(X_{T+1})) | Z_1^T \right] \ll 1.$$
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 $\alpha > 0$:

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

where

$$
\hat{\text{disc}}_H (\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in H_A} \left| \sum_{t=1}^{T} q_t \left[ L (h_t (X_{T+1}), h(X_{T+1})) - L (h_t, Z_t) \right] \right|.
$$
Proof Sketch

\[ \text{disc}(q) = \sup_{h \in H_A} \left| \sum_{t=1}^{T} q_t \left[ \mathcal{L}_{T+1}(h_t, Z_1^{T}) - \mathcal{L}_{t}(h_t, Z_1^{t-1}) \right] \right| \]

\[ \leq \sup_{h \in H_A} \left| \sum_{t=1}^{T} q_t \left[ \mathcal{L}_{T+1}(h_t, Z_1^{T}) - \mathbb{E} \left[ L(h_t(X_{T+1}), h^*(X_{T+1})) \mid Z_1^{T} \right] \right] \right| \]

\[ + \sup_{h \in H_A} \left| \sum_{t=1}^{T} q_t \left[ \mathbb{E} \left[ L(h_t(X_{T+1}), h^*(X_{T+1})) \mid Z_1^{T} \right] \right] - \mathcal{L}_{t}(h_t, Z_1^{t-1}) \right| . \]

\[ \leq \mu \sup_{h \in H_A} \sum_{t=1}^{T} q_t \mathbb{E} \left[ L(h^*(X_{T+1}), Y_{T+1}) \mid Z_1^{T} \right] \]

\[ = \mu \sup_{h \in H_A} \mathbb{E} \left[ L(h^*(X_{T+1}), Y_{T+1}) \mid Z_1^{T} \right] . \]
Lemma

Lemma: let $L$ be a convex loss bounded by $M$ and $h^T_1$ a hypothesis sequence adapted to $Z^T_1$. Fix $\mathbf{q} \in \Delta$. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$ for the hypothesis $h = \sum_{t=1}^{T} q_t h_t$:

$$
\mathcal{L}_{T+1}(h, Z^T_1) \leq \sum_{t=1}^{T} q_t L(h_t, Z_t) + \text{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.
$$
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,

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

$$A_t = q_t \left[ \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1}) - L(h_t, \mathbf{Z}_t) \right]$$ is a martingale difference, thus by Azuma’s inequality, whp,

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

Theorem: let $L$ be a convex loss bounded by $M$ and $H^*$ a set of hypothesis sequences adapted to $Z_1^T$. Fix $q \in \Delta$. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$ for the hypothesis $h = \sum_{t=1}^T q_t h_t$:

$$
L_{T+1}(h, Z_1^T) \\
\leq \inf_{h^* \in H} \sum_{t=1}^T L_{T+1}(h^*, Z_1^T) + 2\text{disc}(q) + \frac{\text{Reg}_T}{T} \\
+ M\|q - u\|_1 + 2M\|q\|_2 \sqrt{2 \log \frac{2}{\delta}}.
$$
Theorem extends to non-convex losses when $h$ is selected as follows:

$$h = \arg\min_{h_t} \left\{ \sum_{s=t}^{T} q_s L(h_t, Z_s) + \text{disc}(q_t^T) + M \|q_t^T\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}} \right\}.$$ 

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:

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

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

- Hypotheses no longer adapted, but output by a uniformly stable algorithm.

- Stable hypotheses:
  - $\mathcal{H} = \{h \in H : \text{there exists } A \in \mathcal{A} \text{ such that } h = A(Z^T_1)\}$.
  - $\beta_t = \beta_{h_t}$: 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 $q \in \Delta$ to minimize discrepancy

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

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

- select model according to

$$h = \arg\min_{h_t} \left\{ \sum_{s=t}^{T} q_s L(h_t, Z_s) + \text{disc}(q_t^T) + M \|q_t^T\|_2 \sqrt{2 \log \frac{2(T+1)}{\delta}} \right\}.$$
Learning Ensembles

Problem: given a hypothesis set $H$ and a sample $Z_1^T$, find accurate convex combination $h = \sum_{t=1}^{T} q_t h_t$ with $h \in H_A$ and $q \in \Delta$.

- in most general case, hypotheses may have been pre-trained on $Z_1^T$. 

Learning Ensembles

Algorithm:

• run regret minimization on $Z_1^T$ to return $h$.

• minimize learning bound. For $\Lambda_2 \geq 0$,

$$
\min_q \ \hat{\text{disc}}_H(q) + \sum_{t=1}^{T} q_t L(h_t, Z_t)
$$

subject to $\|q - u\|_2 \leq \Lambda_2$.

• 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.