Theory and Algorithms for Forecasting Non-Stationary Time Series

VITALY KUZNETSOV KUZNETSOV@

GOOGLE RESEARCH

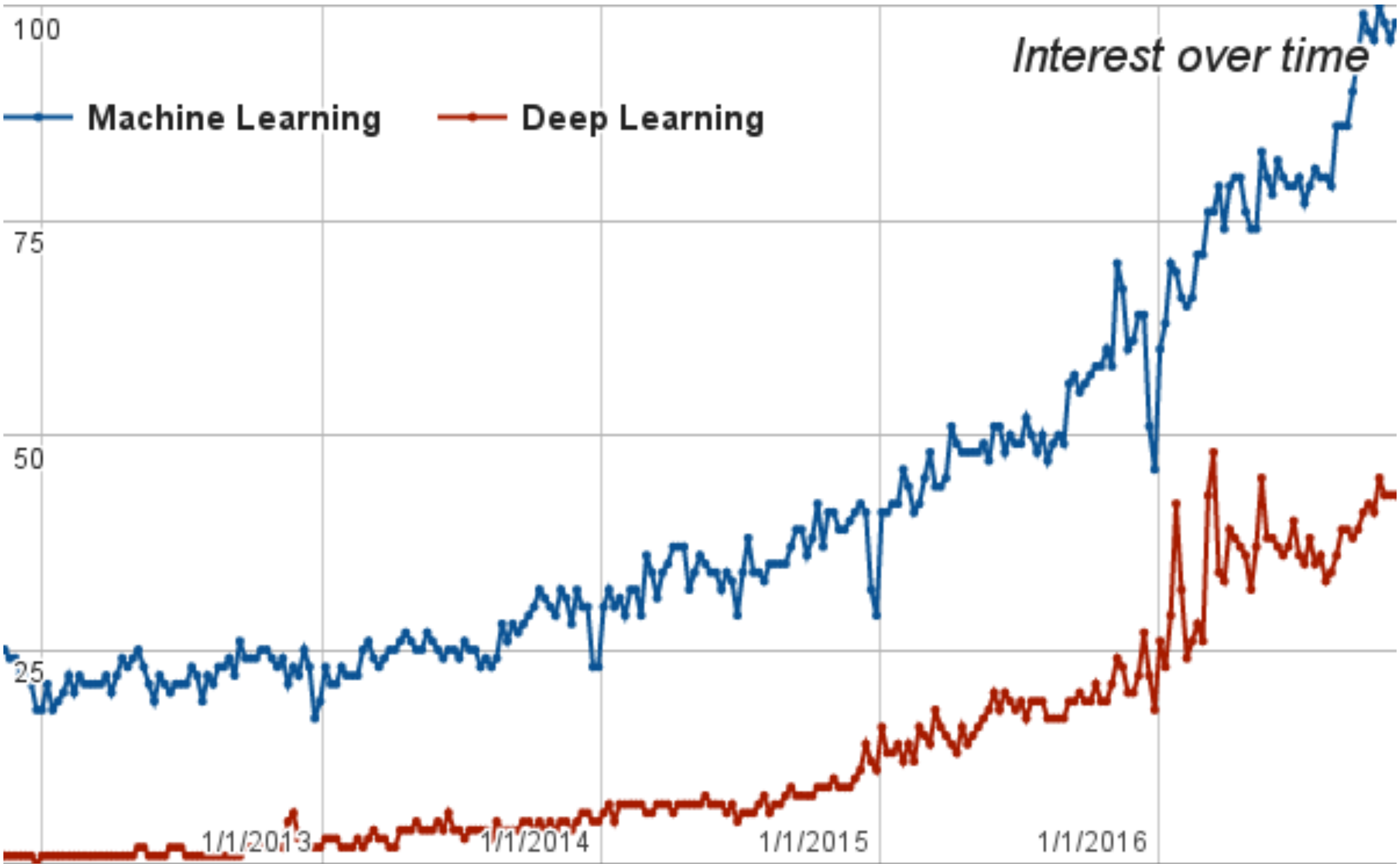
MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

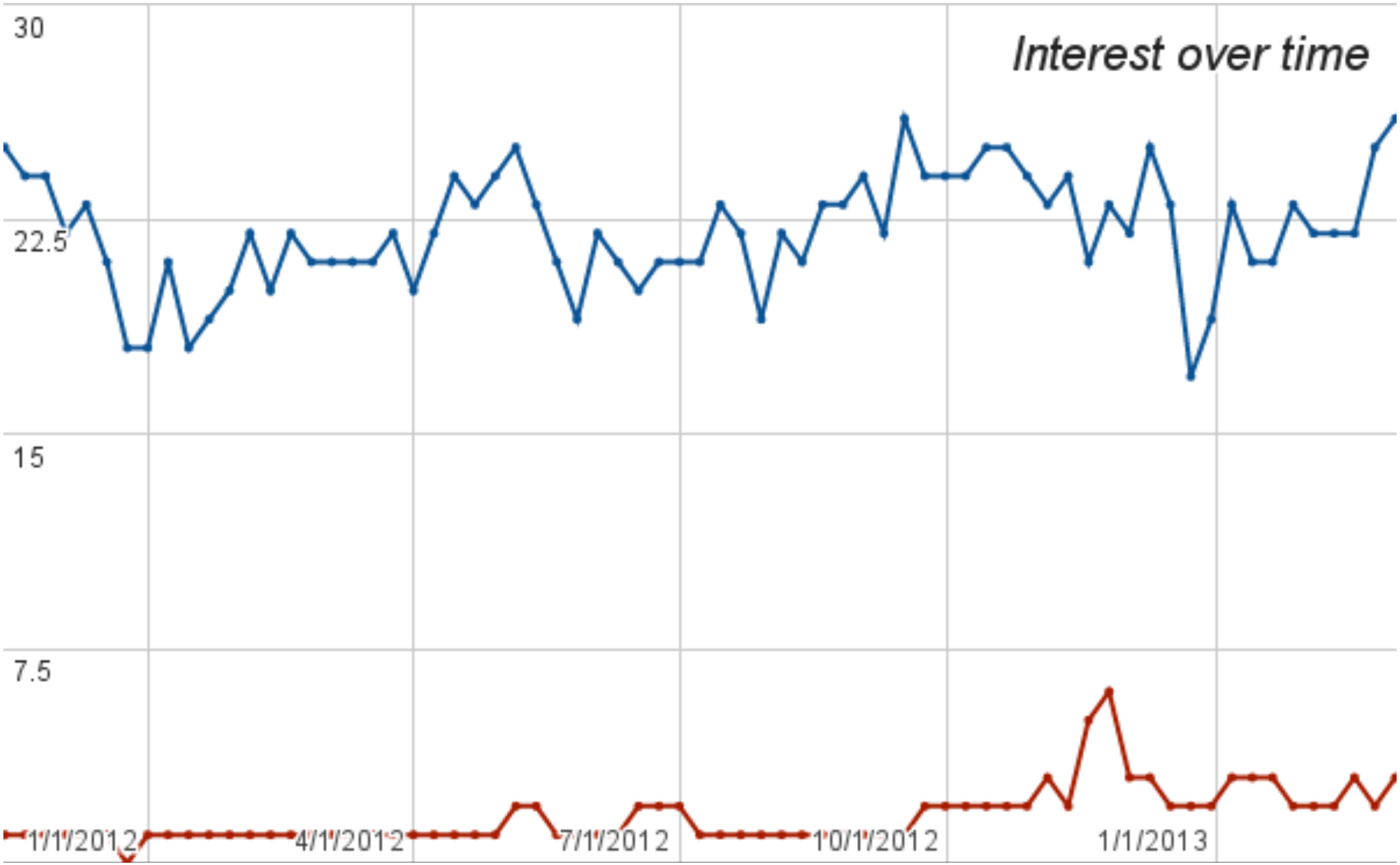
Motivation

- 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.

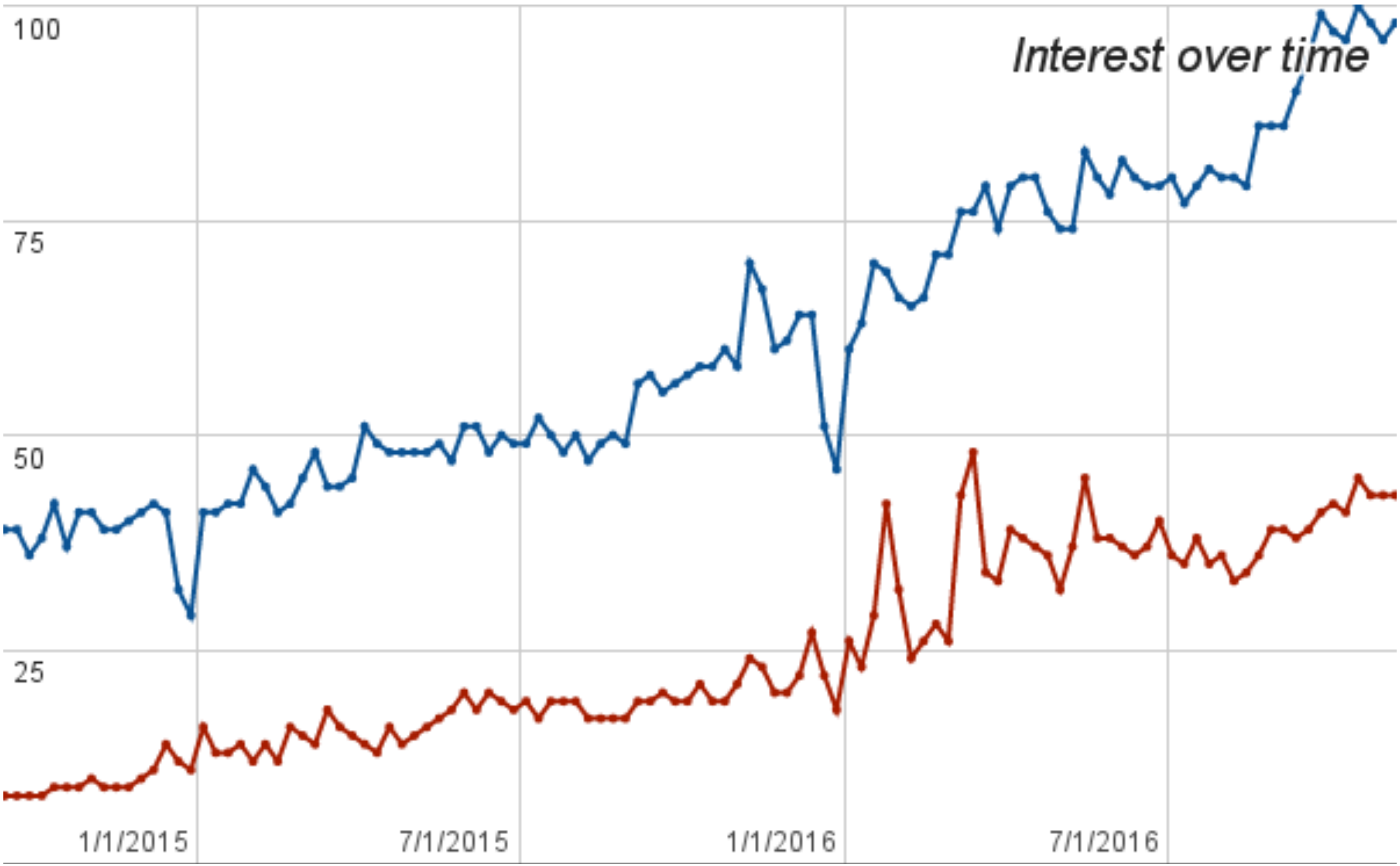
Google Trends



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.

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 pth order Markov assumption:

$$\forall t, Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \epsilon_t$$

where

- ϵ_ts are zero mean uncorrelated random variables with variance σ.
- $a_1, \ldots, a_p$ 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 qth order Markov assumption:

$$\forall t, Y_t = \epsilon_t + \sum_{j=1}^{q} b_j \epsilon_{t-j}$$

where

- $b_1, \ldots, b_q$ are moving average coefficients.

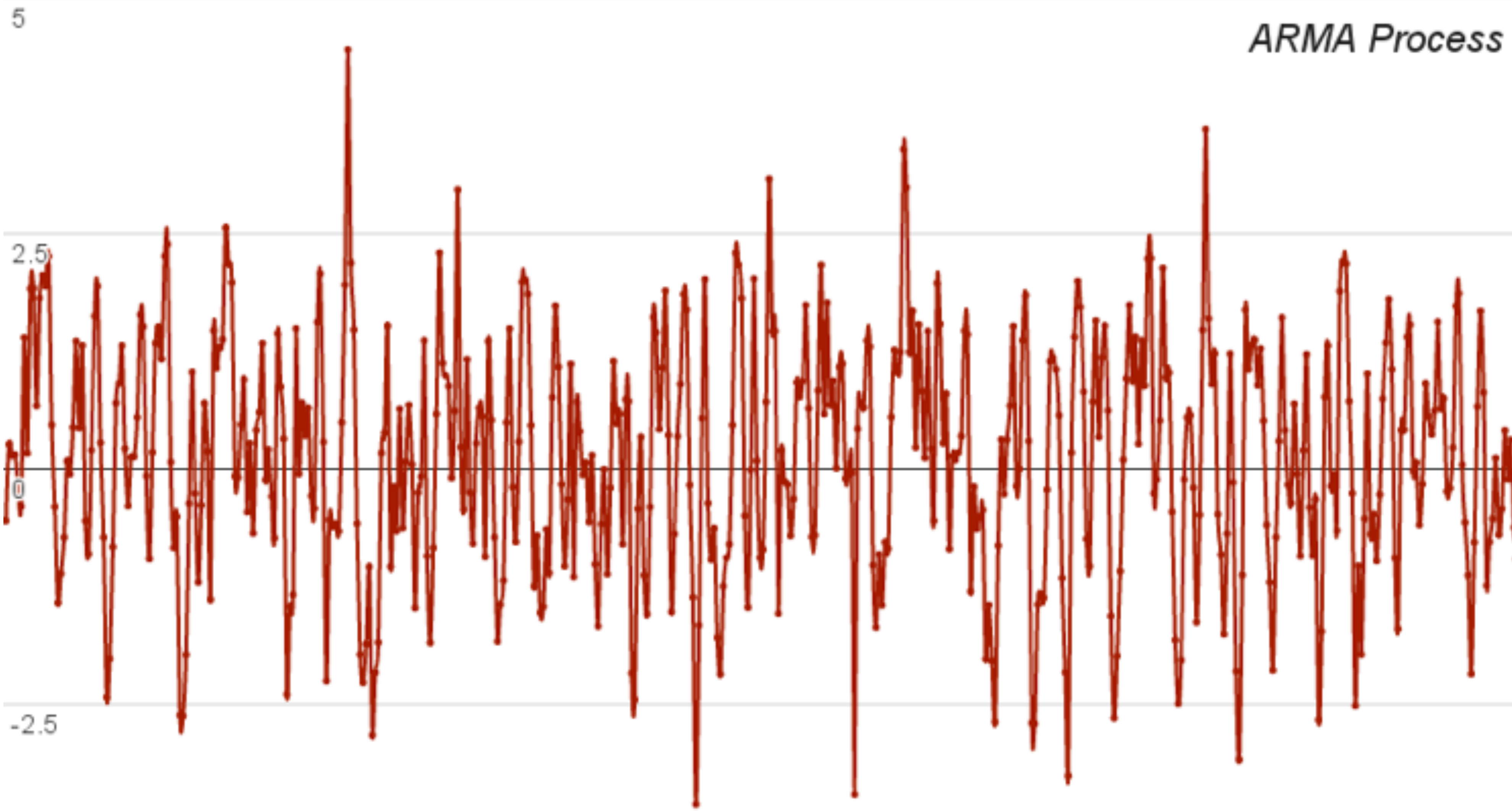
ARMA model

(Whittle, 1951; Box & Jenkins, 1971)

- Definition: ARMA(p, q) model is a generative linear model that combines AR(p) and MA(q) models:

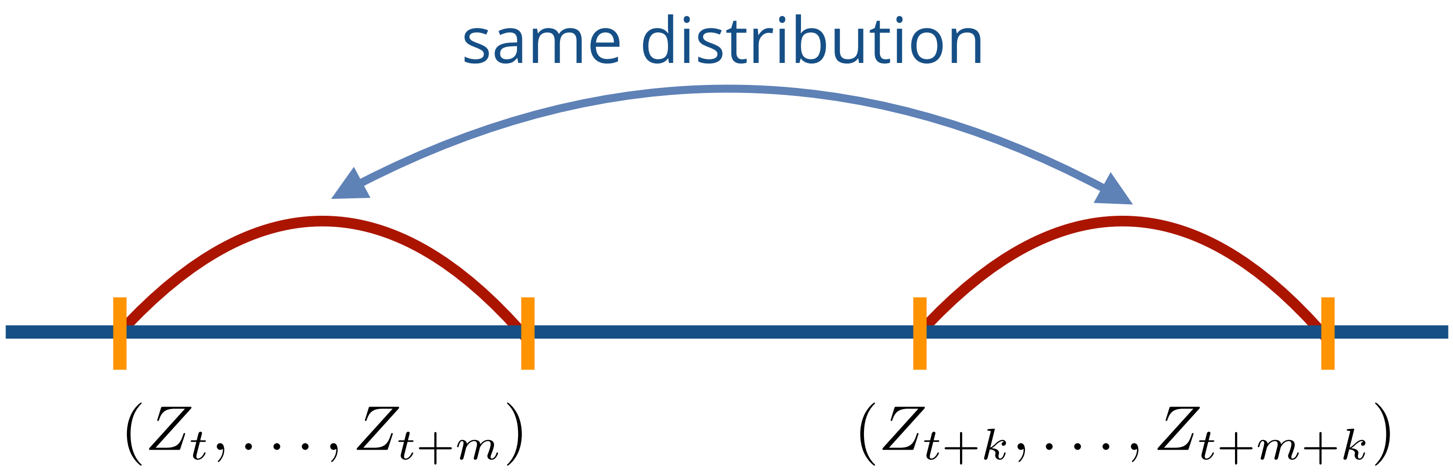
$$\forall t, Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \epsilon_t + \sum_{j=1}^{q} b_j \epsilon_{t-j}.$$

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 $\mathfrak{L}$ is defined by $\mathfrak{L}Y_t = Y_{t-1}$.
- ARMA model in terms of the lag operator:

$$\left(1 - \sum_{i=1}^{p} a_i \mathfrak{L}^i\right) Y_t = \left(1 + \sum_{j=1}^{q} b_j \mathfrak{L}^j\right) \epsilon_t$$

- 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:

$$P(z) = 1 - \sum_{i=1}^{p} a_i z^i = c(\psi_1 - z) \cdots (\psi_p - z)$$
$$= c'(1 - \psi_1^{-1} z) \cdots (1 - \psi_p^{-1} z)$$

where $|\psi_i| > 1$ for all $i = 1, \ldots, p$ and c, c' are constants.

Proof

- Therefore, the ARMA(p,q) process

$$\left(1 - \sum_{i=1}^{p} a_i \mathfrak{L}^i\right) Y_t = \left(1 + \sum_{j=1}^{q} b_j \mathfrak{L}^j\right) \epsilon_t$$

admits MA(∞) representation:

$$Y_t = \left(1 - \psi_1^{-1} \mathfrak{L}\right)^{-1} \cdots \left(1 - \psi_p^{-1} \mathfrak{L}\right)^{-1} \left(1 + \sum_{j=1}^{q} b_j \mathfrak{L}^j\right) \epsilon_t$$

where

$$\left(1 - \psi_i^{-1} \mathfrak{L}\right)^{-1} = \sum_{k=0}^{\infty} \left(- \psi_i^{-1} \mathfrak{L}\right)^k$$

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:

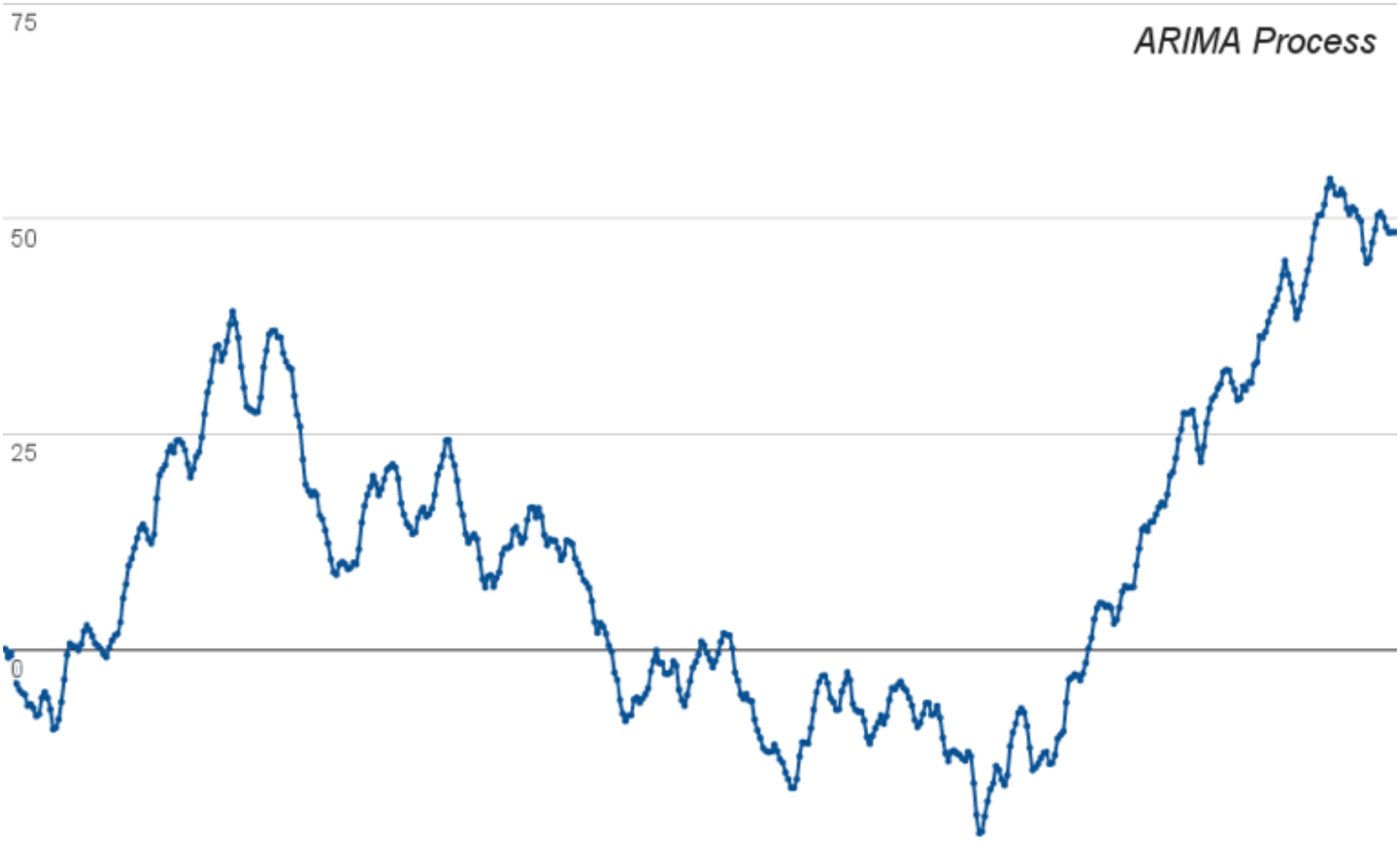
$$P(z) = R(z)(1 - z)^D$$

where $R(z)$ has no unit roots.

- Definition: ARIMA(p, D, q) model is an ARMA(p,q) model for $(1 - \mathfrak{L})^D Y_t$:

$$\left(1 - \sum_{i=1}^{p} a_i \mathfrak{L}^i\right)\left(1 - \mathfrak{L}\right)^D Y_t = \left(1 + \sum_{j=1}^{q} b_j \mathfrak{L}^j\right)\epsilon_t.$$

ARIMA



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)

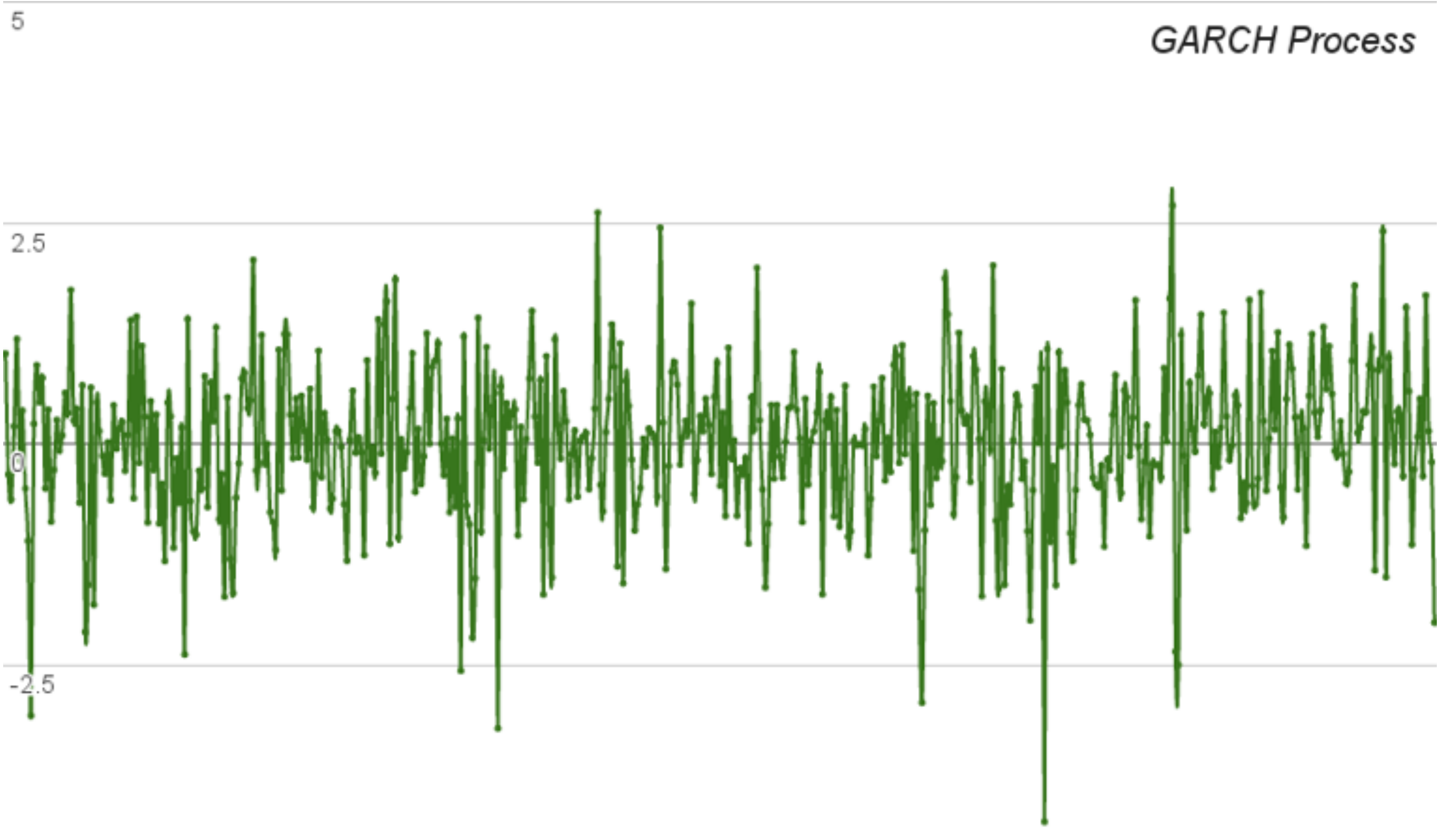
- Definition: the generalized autoregressive conditional heteroscedasticity GARCH(p,q) model is an ARMA(p,q) model for the variance σ_t of the noise term ϵ_t:

$$\forall t,\ \sigma_{t+1}^2 = \omega + \sum_{i=0}^{p-1} \alpha_i \sigma_{t-i}^2 + \sum_{j=0}^{q-1} \beta_j \epsilon_{t-j}^2$$

where

- ϵ_ts are zero mean Gaussian random variables with variance σ_t conditioned on $\{Y_{t-1}, Y_{t-2}, \ldots\}$.
- $\omega > 0$ is the mean parameter.

GARCH Process



State-Space Models

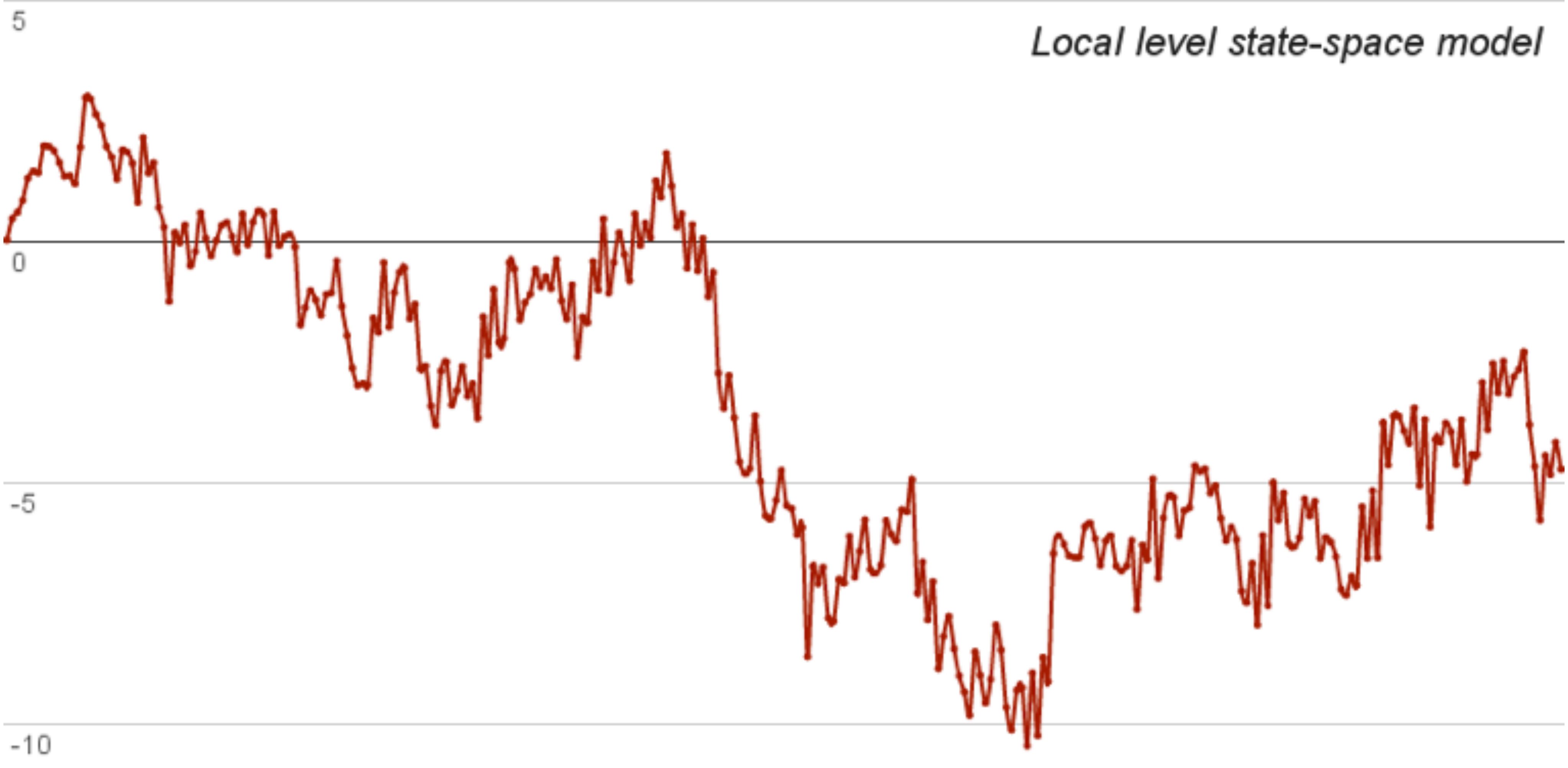
- Continuous state space version of Hidden Markov Models:

$$\mathbf{X}_{t+1} = \mathbf{B}\mathbf{X}_t + \mathbf{U}_t,$$
$$Y_t = \mathbf{A}\mathbf{X}_t + \epsilon_t$$

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



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 $Y_t \sim$ ARMA(p,q) is weakly stationary and invertible. Let $\widehat{\mathbf{a}}_T$ denote the least square estimate of $\mathbf{a} = (a_1, \ldots, a_p)$ and assume that p is known. Then, $\|\widehat{\mathbf{a}}_T - \mathbf{a}\|$ 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,

$$(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}.$$

- Loss function: $L\colon H \times \mathcal{Z} \to [0,1]$, where H is a hypothesis set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$.

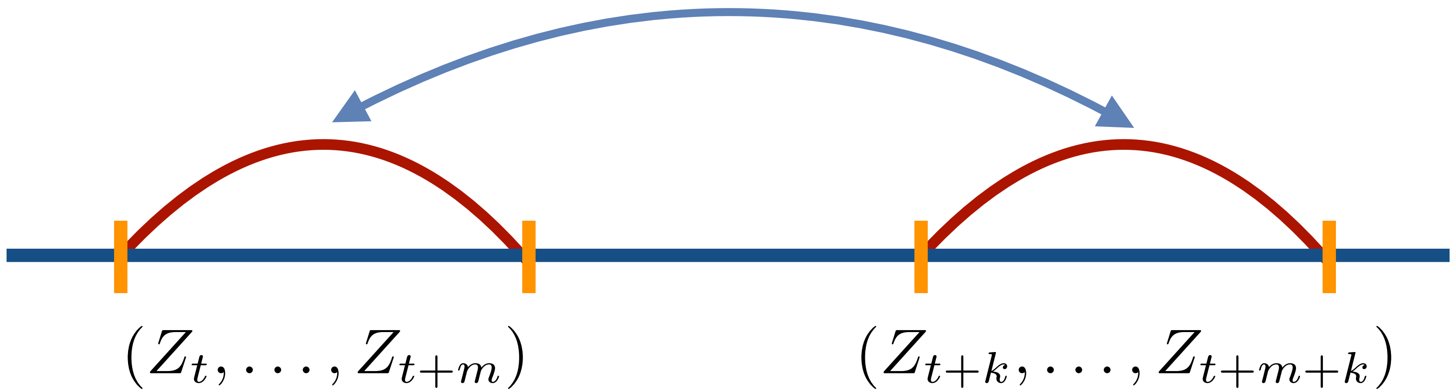
- Problem: find $h \in H$ with small path-dependent expected loss,

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

Standard Assumptions

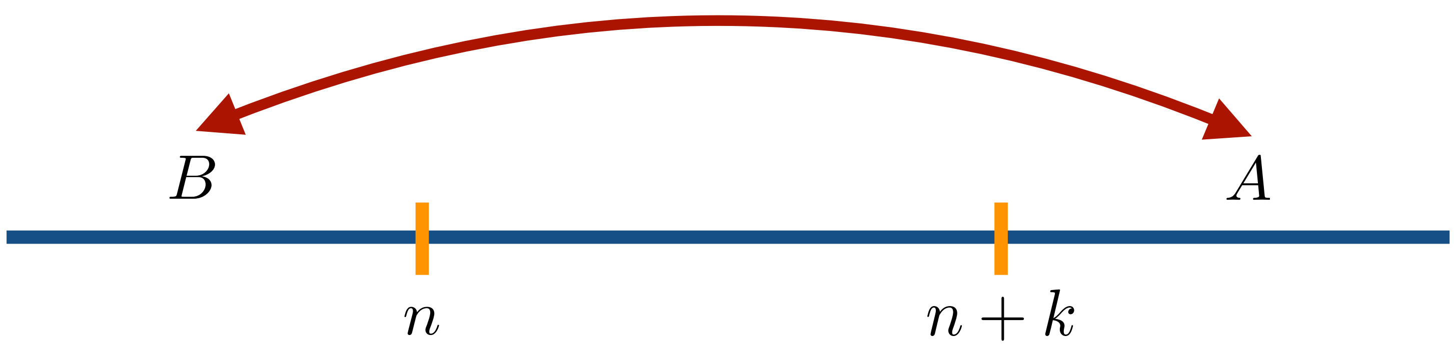
- Stationarity:

same distribution



- Mixing:

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

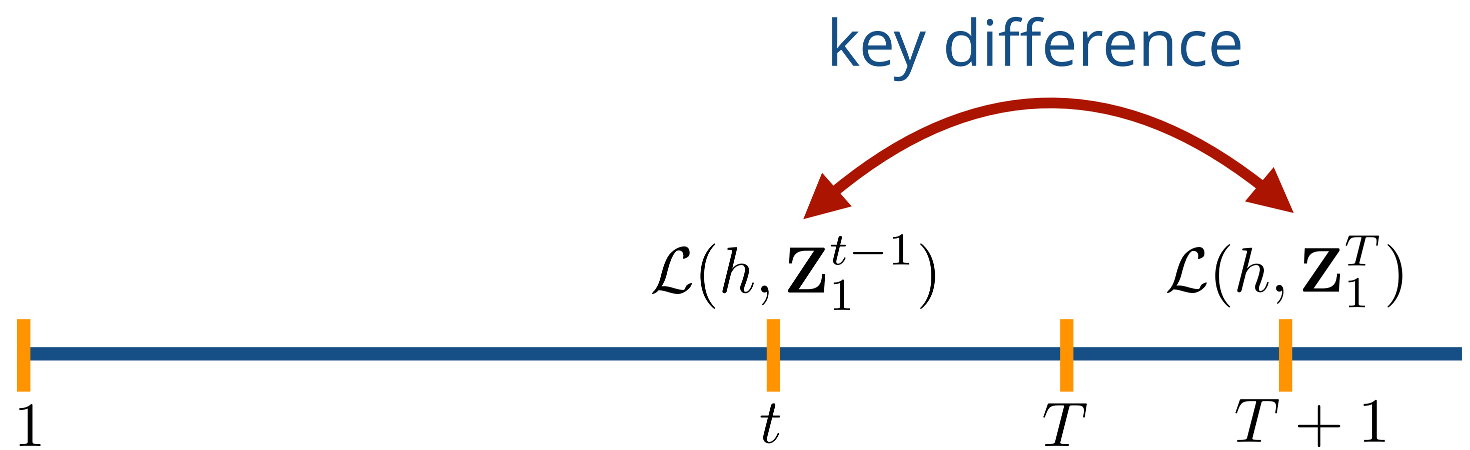
- Stationarity and mixing assumptions:
  - often do not hold (think trend or periodic signals).
  - not testable.
  - estimating mixing parameters can be hard, even if general functional form known.
  - hypothesis set and loss function ignored.

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: $\left| \frac{1}{T} \sum_{t=1}^{T} \left[\mathcal{L}(h, \mathbf{Z}_1^T) - \mathcal{L}(h, \mathbf{Z}_1^{t-1}) \right] \right|.$

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 $(q_1, \ldots, q_T) = \mathbf{q}$.

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

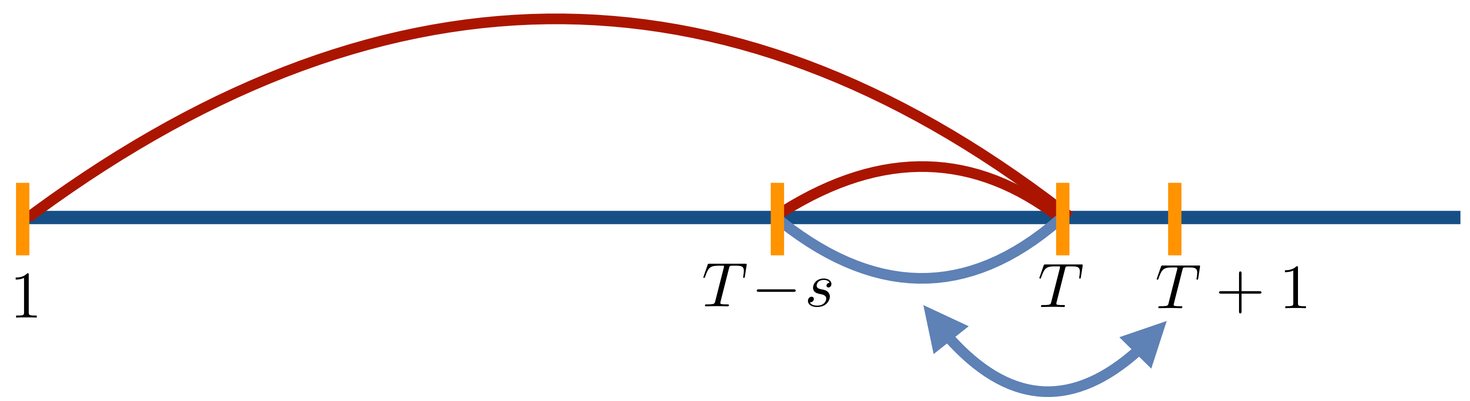
  - strictly extends discrepancy definition in drifting (MM and Muñoz Medina, 2012) or domain adaptation (Mansour, MM, Rostamizadeh 2009; Cortes and MM 2011, 2014); or for binary loss (Devroye et al., 1996; Ben-David et al., 2007).
  - admits upper bounds in terms of relative entropy, or in terms of ϕ-mixing coefficients of asymptotic stationarity for an asymptotically stationary process.

Estimation

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

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

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



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 $\mathcal{G} = \{z \mapsto L(h, z) \colon h \in H\}$.

Bound with Emp. Discrepancy

- Corollary: for any $\delta > 0$, with probability at least $1-\delta$, for all $h \in H$ and $\alpha > 0$,

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + 4\alpha + \Big[\|\mathbf{q}\|_2 + \|\mathbf{q} - \mathbf{u}_s\|_2\Big] \sqrt{2 \log \frac{2\,\mathbb{E}_{\mathbf{z}}\left[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})\right]}{\delta}},$$

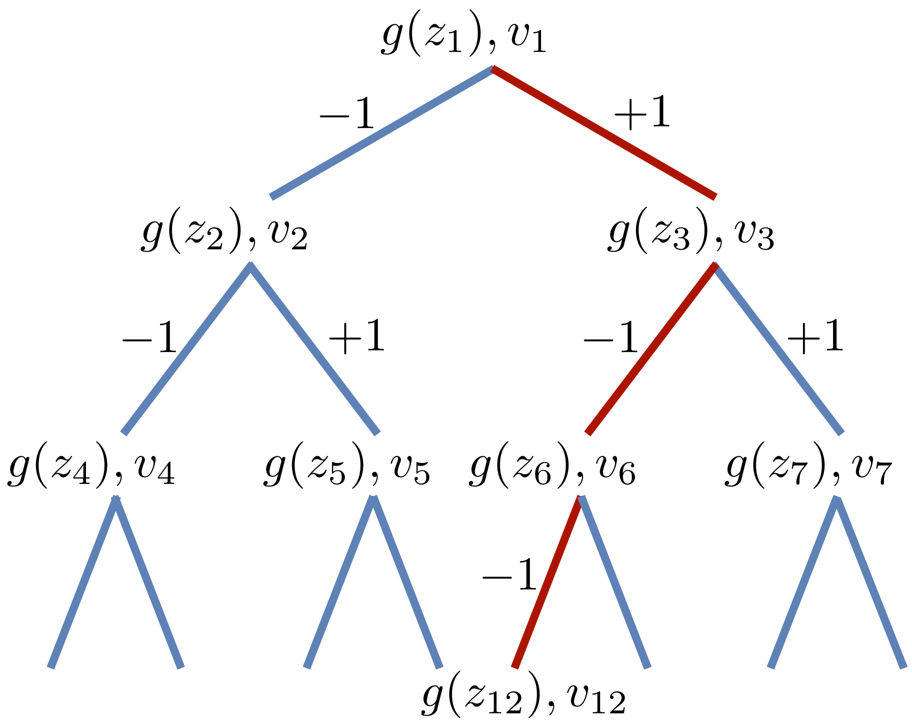
where
$$\begin{cases} \widehat{\Delta}(\mathbf{q}) = \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^{T} L(h, Z_t) - \sum_{t=1}^{T} q_t L(h, Z_t) \right) \\ \mathbf{u}_s \text{ unif. dist. over } [T-s, T] \\ \mathcal{G} = \{z \mapsto L(h, z) \colon h \in H\}. \end{cases}$$

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

$$\forall g \in \mathcal{G}, \forall \boldsymbol{\sigma} \in \{\pm 1\}^T, \exists \mathbf{v} \in \mathcal{V} \colon \sum_{t=1}^{T} |v_t(\boldsymbol{\sigma}) - g(z_t(\boldsymbol{\sigma}))| \leq \frac{\alpha}{\|\mathbf{q}\|_\infty}.$$



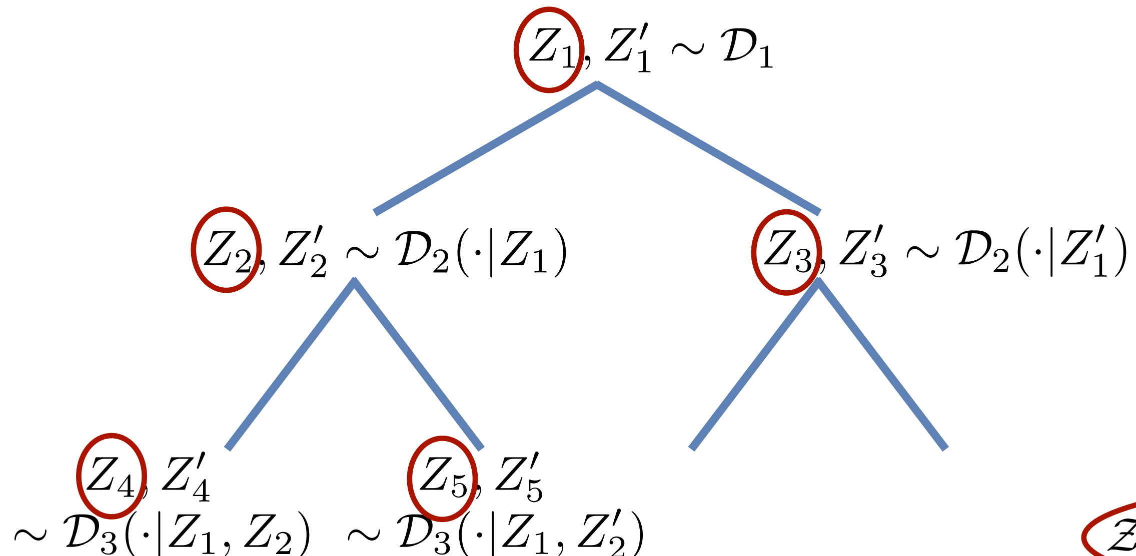
$$\left\| \begin{bmatrix} v_1 - g(z_1) \\ v_3 - g(z_3) \\ v_6 - g(z_6) \\ v_{12} - g(z_{12}) \end{bmatrix} \right\|_1 \leq \frac{\alpha}{\|\mathbf{q}\|_\infty}.$$

Sequential Covering Numbers

- Definitions:
  - sequential covering number:

    $$\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z}) = \min\{|\mathcal{V}| : \mathcal{V}\ l_1\text{-norm } \mathbf{q}\text{-weighted } \alpha\text{-cover of } \mathcal{G}\}.$$

  - expected sequential covering number: $\underset{\mathbf{z} \sim \mathcal{Z}_T}{\mathbb{E}}[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})]$.



$\mathcal{Z}_T$: distribution based on Z_ts.

Proof

- **Key quantities:** $\Phi(\mathbf{Z}_1^T) = \sup_{h \in H} \Big(\mathcal{L}(h, Z_T) - \sum_{t=1}^{T} q_t L(h, Z_t) \Big)$

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

- Chernoff technique: for any $t > 0$,

$$\mathbb{P}\big[\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)\big]$$

$$\leq \mathbb{P}\left[\sup_{h \in H} \sum_{t=1}^{T} q_t \big[\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)\big] > \epsilon \right] \qquad \text{(sub-add. of sup)}$$

$$= \mathbb{P}\left[\exp\Big(t \sup_{h \in H} \sum_{t=1}^{T} q_t \big[\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)\big] \Big) > e^{t\epsilon} \right] \qquad (t > 0)$$

$$\leq e^{-t\epsilon} \, \mathbb{E}\left[\exp\Big(t \sup_{h \in H} \sum_{t=1}^{T} q_t \big[\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)\big] \Big) \right]. \qquad \text{(Markov's ineq.)}$$

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 $\mathbf{Z}_1^{t-1}$.

$$
\begin{aligned}
&\mathbb{P}\left[\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)\right] \\
&\leq e^{-t\epsilon}\,\mathbb{E}\left[\exp\left(t \sup_{h\in H} \sum_{t=1}^T q_t\left[\mathcal{L}(h, \mathbf{Z}_1^{t-1}) - L(h, Z_t)\right]\right)\right] \\
&= e^{-t\epsilon}\,\mathbb{E}\left[\exp\left(t \sup_{h\in H} \sum_{t=1}^T q_t\left[\,\mathbb{E}[L(h, Z'_t)|\mathbf{Z}_1^{t-1}] - L(h, Z_t)\right]\right)\right] && \text{(tangent seq.)} \\
&= e^{-t\epsilon}\,\mathbb{E}\left[\exp\left(t \sup_{h\in H} \mathbb{E}\left[\sum_{t=1}^T q_t\left[L(h, Z'_t) - L(h, Z_t)\right] \,\Big|\, \mathbf{Z}_1^T\right]\right)\right] && \text{(lin. of expectation)} \\
&\leq e^{-t\epsilon}\,\mathbb{E}\left[\exp\left(t \sup_{h\in H} \sum_{t=1}^T q_t\left[L(h, Z'_t) - L(h, Z_t)\right]\right)\right]. && \text{(Jensen's ineq.)}
\end{aligned}
$$

Symmetrization

$$
\begin{aligned}
&\mathbb{P}\left[\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)\right] \\
&\leq e^{-t\epsilon}\,\mathbb{E}\left[\exp\left(t\sup_{h\in H}\sum_{t=1}^{T} q_t\left[L(h,Z'_t) - L(h,Z_t)\right]\right)\right] \\
&= e^{-t\epsilon}\underset{(\mathbf{z},\mathbf{z}')}{\mathbb{E}}\;\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\exp\left(t\sup_{h\in H}\sum_{t=1}^{T} q_t\sigma_t\left[L(h,z'_t(\boldsymbol{\sigma})) - L(h,z_t(\boldsymbol{\sigma}))\right]\right)\right] && \text{(tangent seq. prop.)} \\
&= e^{-t\epsilon}\underset{(\mathbf{z},\mathbf{z}')}{\mathbb{E}}\;\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\exp\left(t\sup_{h\in H}\sum_{t=1}^{T} q_t\sigma_t L(h,z'_t(\boldsymbol{\sigma})) + t\sup_{h\in H}\sum_{t=1}^{T} q_t\sigma_t L(h,z_t(\boldsymbol{\sigma}))\right)\right] && \text{(sub-add. of sup)} \\
&\leq e^{-t\epsilon}\underset{(\mathbf{z},\mathbf{z}')}{\mathbb{E}}\;\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\frac{1}{2}\exp\left(2t\sup_{h\in H}\sum_{t=1}^{T} q_t\sigma_t L(h,z'_t(\boldsymbol{\sigma}))\right)\right. \\
&\qquad\qquad\left. + \frac{1}{2}\exp\left(2t\sup_{h\in H}\sum_{t=1}^{T} q_t\sigma_t L(h,z_t(\boldsymbol{\sigma}))\right)\right] && \text{(convexity. of exp)} \\
&= e^{-t\epsilon}\underset{\mathbf{z}}{\mathbb{E}}\;\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\exp\left(2t\sup_{h\in H}\sum_{t=1}^{T} q_t\sigma_t L(h,z_t(\boldsymbol{\sigma}))\right)\right].
\end{aligned}
$$

Covering Number

$$
\begin{aligned}
&\mathbb{P}\left[\Phi(\mathbf{Z}_1^T - \Delta(\mathbf{q}) > \epsilon)\right] \\
&\leq e^{-t\epsilon} \mathop{\mathbb{E}}_{\mathbf{z}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\exp\left(2t \sup_{h \in H} \sum_{t=1}^{T} q_t \sigma_t L(h, z_t(\boldsymbol{\sigma}))\right)\right] \\
&\leq e^{-t\epsilon} \mathop{\mathbb{E}}_{\mathbf{z}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\exp\left(2t \left[\max_{\mathbf{v} \in \mathcal{V}} \sum_{t=1}^{T} q_t \sigma_t v_t(\boldsymbol{\sigma}) + \alpha\right]\right)\right] && (\alpha\text{-covering}) \\
&\leq e^{-t(\epsilon - 2\alpha)} \mathop{\mathbb{E}}_{\mathbf{z}} \left[\sum_{\mathbf{v} \in \mathcal{V}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\exp\left(2t \sum_{t=1}^{T} q_t \sigma_t v_t(\boldsymbol{\sigma})\right)\right]\right] && (\text{monotonicity of exp}) \\
&\leq e^{-t(\epsilon - 2\alpha)} \mathop{\mathbb{E}}_{\mathbf{z}} \left[\sum_{\mathbf{v} \in \mathcal{V}} \exp\left(\frac{t^2 \|\mathbf{q}\|^2}{2}\right)\right] && (\text{Hoeffding's ineq.}) \\
&\leq \mathop{\mathbb{E}}_{\mathbf{z}} \left[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})\right] \exp\left[-t(\epsilon - 2\alpha) + \frac{t^2 \|\mathbf{q}\|^2}{2}\right].
\end{aligned}
$$

Algorithms

Review

- Theorem: for any $\delta > 0$, with probability at least $1-\delta$, for all $h \in H$ and $\alpha > 0$,

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + 4\alpha$$
$$+ \Big[\|\mathbf{q}\|_2 + \|\mathbf{q} - \mathbf{u}_s\|_2\Big] \sqrt{2 \log \frac{2 \operatorname*{\mathbb{E}}_{\mathbf{z}} \big[\mathcal{N}_1(\alpha, \mathcal{G}, \mathbf{z})\big]}{\delta}}.$$

- This bound can be extended to hold uniformly over $\mathbf{q}$ at the price of the additional term:

$$\widetilde{O}\big(\|\mathbf{q} - \mathbf{u}\|_1 \sqrt{\log_2 \log_2 (1 - \|\mathbf{q} - \mathbf{u}\|)^{-1}}\big).$$

- Data-dependent learning guarantee.

Discrepancy-Risk Minimization

- Key Idea: directly optimize the upper bound on generalization over $\mathbf{q}$ 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,

$$H = \Big\{x \mapsto \mathbf{w} \cdot \Phi_K(x) \colon \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\Big\}.$$

- Complexity term can be bounded by

$$O\Big((\log^{3/2} T)\Lambda \sup_x K(x, x) \|\mathbf{q}\|_2\Big).$$

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

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

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

$$\left. + \sum_{t=1}^{T} \left(\frac{1}{T} - q_t \right) \frac{1}{s} \sum_{t=T-s+1}^{T} L(h, Z_t) \right\}$$

$$\leq \sum_{t=1}^{T} q_t \sup_{h \in H} \left(\frac{1}{s} \sum_{t=T-s+1}^{T} L(h, Z_t) - q_t L(h, Z_t) \right) + M \|\mathbf{u} - \mathbf{q}\|_1.$$

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 $0 < \|\mathbf{q}-\mathbf{u}\|_1 \leq 1$

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + \widetilde{O}\Big(\log^{3/2} T \sup_x K(x,x)\Lambda + \|\mathbf{q}-\mathbf{u}\|_1\Big).$$

- 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 $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + \widetilde{O}\Big(\log^{3/2} T \sup_x K(x,x)\Lambda + \|\mathbf{q} - \mathbf{u}\|_1\Big).$$

- 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 $0 < \|\mathbf{q}-\mathbf{u}\|_1 \leq 1$

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \boxed{\widehat{\Delta}(\mathbf{q})} + \Delta_s + \widetilde{O}\Big(\log^{3/2} T \sup_x K(x,x)\Lambda + \|\mathbf{q}-\mathbf{u}\|_1\Big).$$

- 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 $0 < \|\mathbf{q} - \mathbf{u}\|_1 \leq 1$

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + \widetilde{O}\Big(\log^{3/2} T \sup_x K(x, x) \Lambda + \|\mathbf{q} - \mathbf{u}\|_1 \Big).$$

- 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 $0 < \|\mathbf{q}-\mathbf{u}\|_1 \leq 1$

$$\mathcal{L}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h, Z_t) + \widehat{\Delta}(\mathbf{q}) + \Delta_s + \widetilde{O}\Big(\log^{3/2} T \sup_x K(x,x)\Lambda + \|\mathbf{q}-\mathbf{u}\|_1\Big).$$

- Corresponding optimization problem:

$$\min_{\mathbf{q}\in[0,1]^T, \mathbf{w}} \left\{ \sum_{t=1}^{T} q_t (\mathbf{w}\cdot\Psi_K(x_t) - y_t)^2 + \lambda_1 \sum_{t=1}^{T} q_t d_t + \lambda_2 \|\mathbf{w}\|_{\mathbb{H}} + \lambda_3 \|\mathbf{q}-\mathbf{u}\|_1 \right\}.$$

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 $\mathcal{D} = \{\mathbf{r}\colon r_t \geq 1\}$.
- convex optimization problem.

Two-Stage Algorithm

- Minimize empirical discrepancy $\widehat{\Delta}(\mathbf{q})$ 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

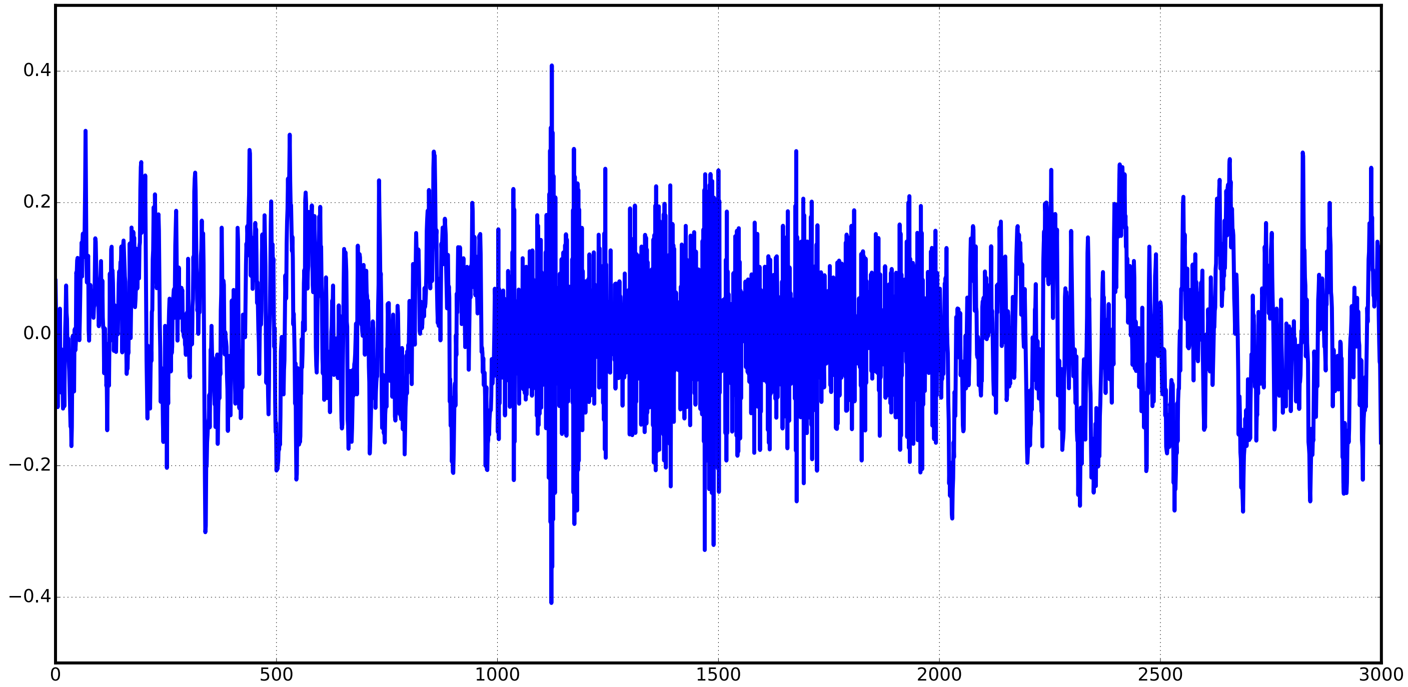
- Artificial data sets:

$$\textbf{ads1}: \quad Y_t = \alpha_t Y_{t-1} + \epsilon_t, \qquad \alpha_t = -0.9 \text{ if } t \in [1000, 2000] \text{ and } 0.9 \text{ otherwise},$$

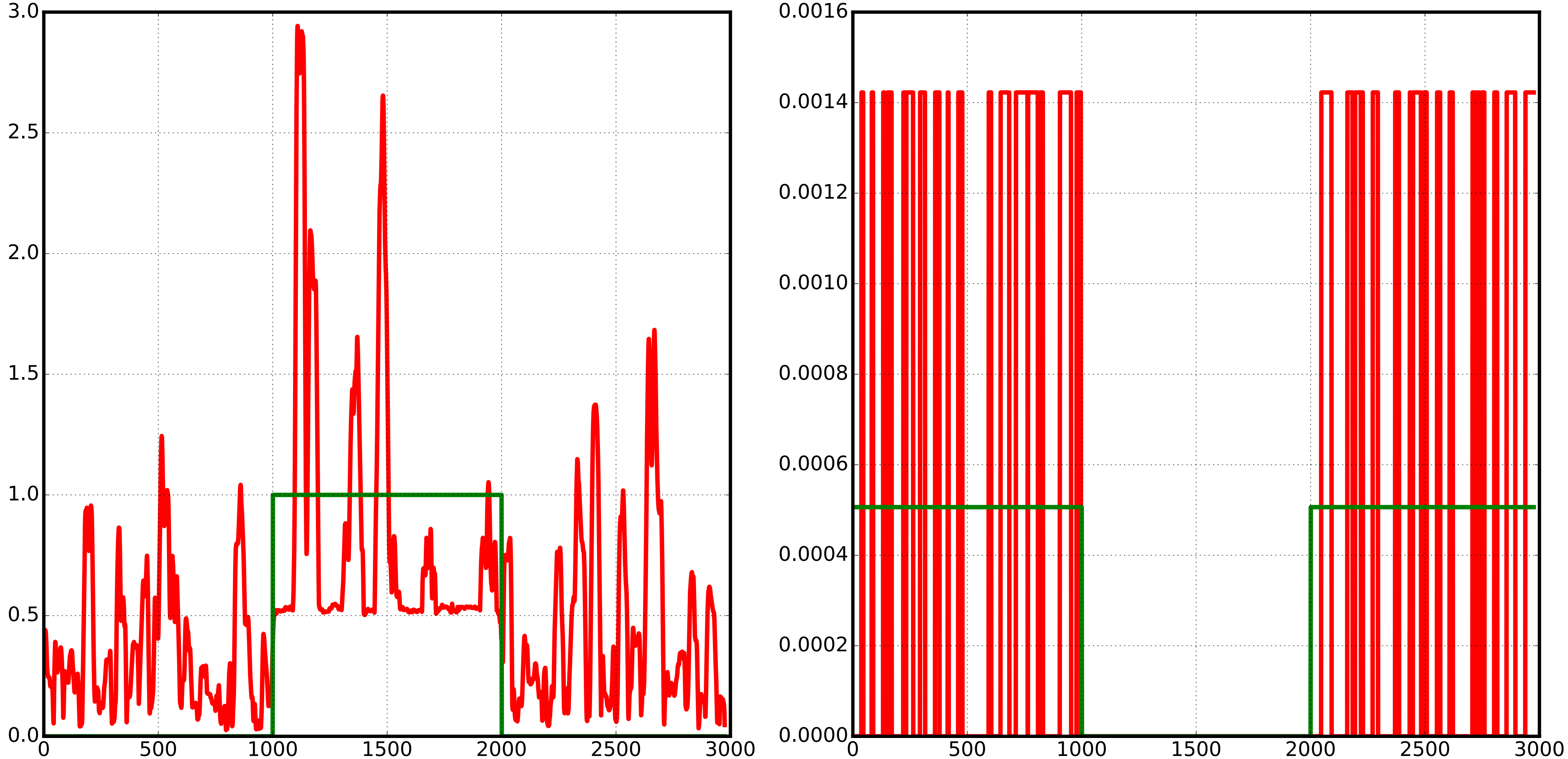
$$\textbf{ads2}: \quad Y_t = \alpha_t Y_{t-1} + \epsilon_t, \qquad \alpha_t = 1 - (t/1500),$$

$$\textbf{ads3}: \quad Y_t = \alpha_{i(t)} Y_{t-1} + \epsilon_t, \qquad \alpha_1 = -0.5 \text{ and } \alpha_2 = 0.9,$$

$$\textbf{ads4}: \quad Y_t = -0.5 Y_{t-1} + \epsilon_t,$$



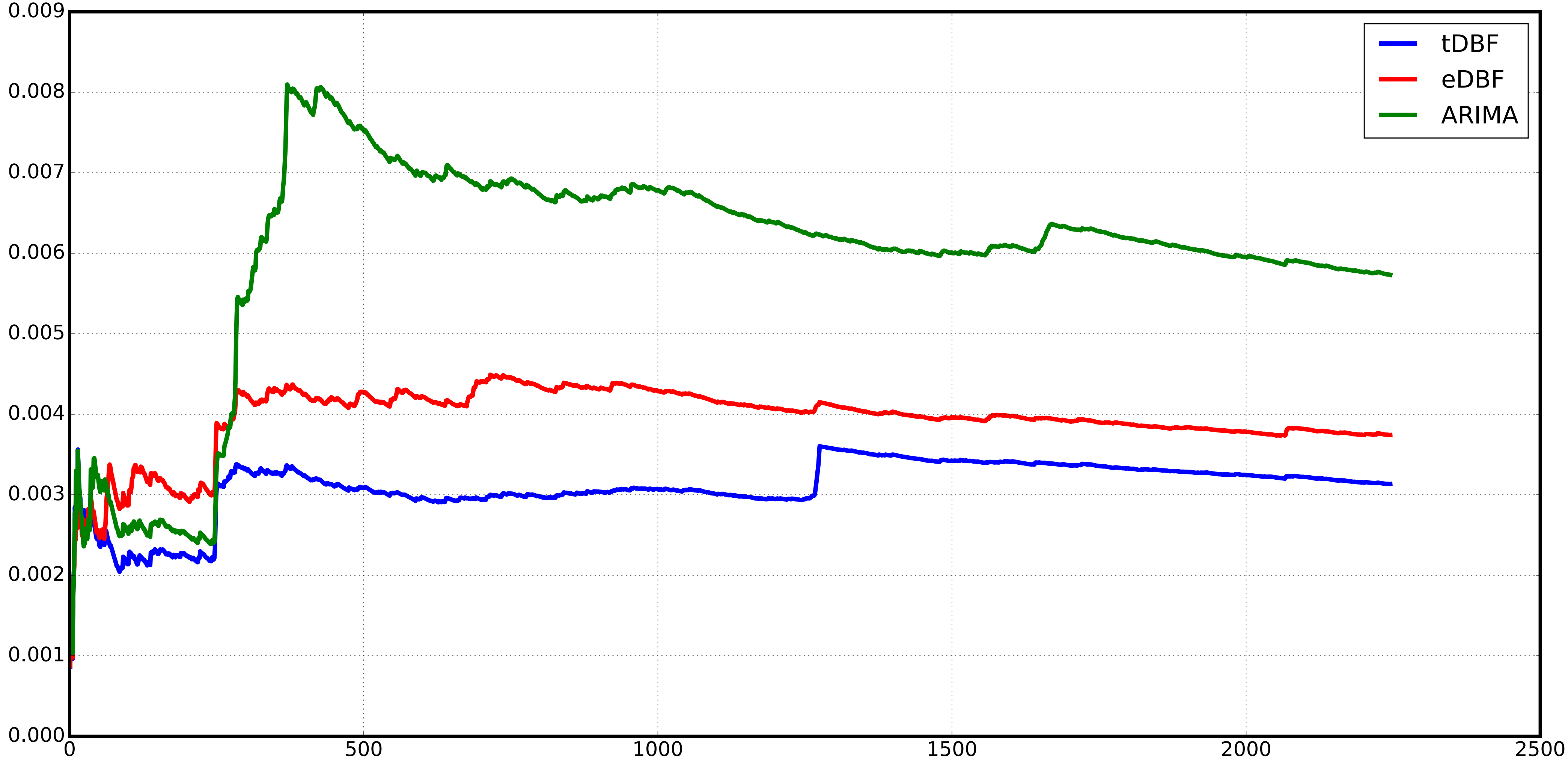
True vs. Empirical Discrepancies



Discrepancies

Weights

Running MSE



0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
tDBF
eDBF
ARIMA

Real-world Data

- Commodity prices, exchange rates, temperatures & climate.

Dataset	DBF	ARIMA
bitcoin	$\mathbf{4.400 \times 10^{-3}}$ $\mathbf{(26.500 \times 10^{-3})}$	4.900×10^{-3} (29.990×10^{-3})
coffee	$\mathbf{3.080 \times 10^{-3}}$ $\mathbf{(6.570 \times 10^{-3})}$	3.260×10^{-3} (6.390×10^{-3})
eur/jpy	$\mathbf{7.100 \times 10^{-5}}$ $\mathbf{(16.900 \times 10^{-5})}$	7.800×10^{-5} (24.200×10^{-5})
jpy/usd	$\mathbf{9.770 \times 10^{-1}}$ $\mathbf{(25.893 \times 10^{-1})}$	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	$\mathbf{7.640 \times 10^{-4}}$ $\mathbf{(46.65 \times 10^{-4})}$	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 $H^* = \{\mathcal{E}_1, \ldots, \mathcal{E}_N\}$, $H = \text{conv}(H^*)$.

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

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**
    $w_{t+1,i} \leftarrow w_{t,i}\, e^{-\eta L(\mathcal{E}_i(x_t), y_t)}$ ▷ (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

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

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

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

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

EW - Proof

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

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

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

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

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,

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

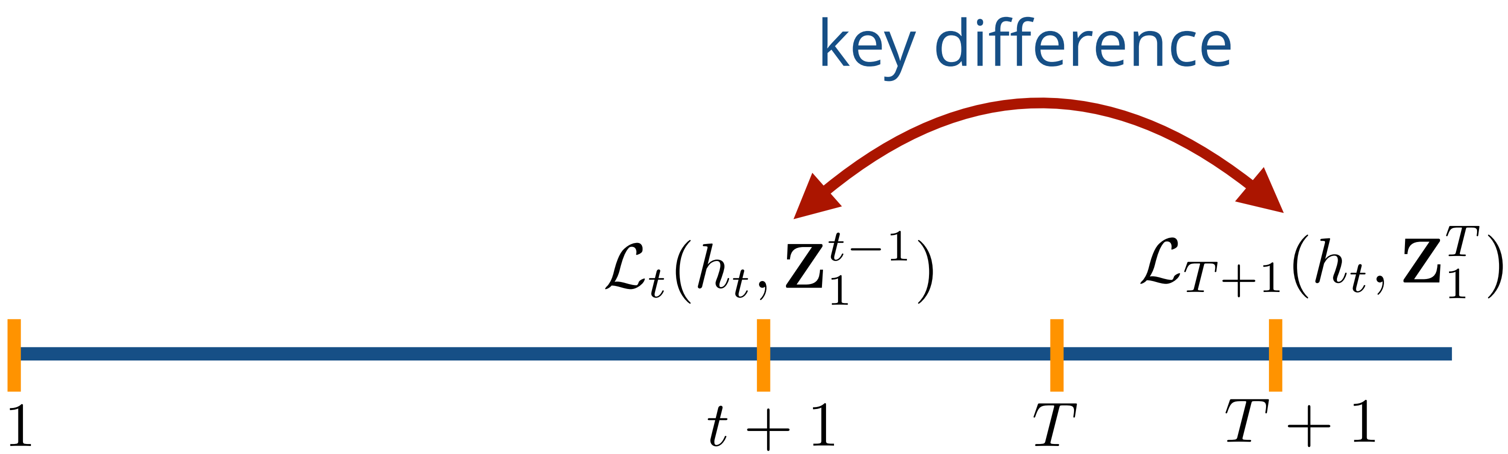
  - i.i.d. problem is standard: (Littlestone, 1989), (Cesa-Bianchi et al., 2004).
  - but, how do we design solutions for the general time-series scenario?

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

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) .

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 $\alpha > 0$:

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

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

$$
\begin{aligned}
\operatorname{disc}(\mathbf{q}) &= \sup_{\mathbf{h}\in\mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1})\Big] \right| \\
&\leq \sup_{\mathbf{h}\in\mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathbb{E}\Big[L(h_t(X_{T+1}), h^*(X_{T+1})) \Big| \mathbf{Z}_1^T\Big]\Big] \right| \\
&\quad + \sup_{\mathbf{h}\in\mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathbb{E}\Big[L(h_t(X_{T+1}), h^*(X_{T+1})) \Big| \mathbf{Z}_1^T\Big]\Big] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1})\Big] \right| \\
&\leq \mu \sup_{\mathbf{h}\in H_{\mathcal{A}}} \sum_{t=1}^{T} q_t \, \mathbb{E}\Big[L(h^*(X_{T+1}), Y_{T+1}) \Big| \mathbf{Z}_1^T\Big] \\
&\quad + \sup_{\mathbf{h}\in\mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathbb{E}\Big[L(h_t(X_{T+1}), h^*(X_{T+1})) \Big| \mathbf{Z}_1^T\Big]\Big] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1})\Big] \right| \\
&= \mu \sup_{\mathbf{h}\in H_{\mathcal{A}}} \mathbb{E}\Big[L(h^*(X_{T+1}), Y_{T+1}) \Big| \mathbf{Z}_1^T\Big] \\
&\quad + \sup_{\mathbf{h}\in\mathbf{H}_{\mathcal{A}}} \left| \sum_{t=1}^{T} q_t \Big[\mathbb{E}\Big[L(h_t(X_{T+1}), h^*(X_{T+1})) \Big| \mathbf{Z}_1^T\Big]\Big] - \mathcal{L}_t(h_t, \mathbf{Z}_1^{t-1})\Big] \right|.
\end{aligned}
$$

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

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

Proof

- By definition of the on-line discrepancy,

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

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

$$\sum_{t=1}^{T} q_t \mathcal{L}_t(h_t, Z_1^{t-1}) \leq \sum_{t=1}^{T} q_t L(h_t, Z_t) + \|\mathbf{q}\|_2 \sqrt{2 \log \tfrac{1}{\delta}}.$$

- 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 $\mathbf{H}^*$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 $h = \sum_{t=1}^T q_t h_t$:

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

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.

Time Series Workshop

Join us in Room 117
Friday, December 9th.

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.

β-Mixing

- Definition: a sequence of random variables $\mathbf{Z} = \{Z_t\}_{-\infty}^{+\infty}$ is β-mixing if

$$\beta(k) = \sup_n \mathop{\mathbb{E}}_{B \in \sigma_{-\infty}^n} \left[\sup_{A \in \sigma_{n+k}^\infty} \Big| \mathbb{P}[A \mid B] - \mathbb{P}[A] \Big| \right] \to 0.$$

dependence between events decaying with k.