

Generalization Bounds for Non-stationary Mixing Processes by Kuznetsov, Mohri

Paper Summary Presentation

Patricio¹

¹Department of Economics, Stanford University

¹Department of Statistics, Stanford University

Winter 2025

Overview

- 1 Introduction
- 2 Rademacher Complexity Primer
- 3 Generalization Error for Time Series
- 4 Technical Foundations
- 5 Path-Dependent Generalization
- 6 Asymptotically Stationary Processes
- 7 Fast Learning Rates
- 8 Unbounded Losses

Motivation: Beyond i.i.d. Data

Traditional Statistical Learning Assumption:

- Data drawn independently and identically distributed (i.i.d.)
- Well-understood generalization theory

Reality of Time Series Forecasting:

- Distribution may **change over time** (non-stationarity) :(
- Previous theory mostly limited to **stationary** mixing processes

Their Contribution

- Quite a few things...

Setting: Non-stationary Stochastic Processes

- Doubly infinite sequence of $\mathcal{Z}$ -valued random variables $\{Z_t\}_{t=-\infty}^{\infty}$
- Notation: Z_a^b denotes vector $(Z_a, Z_{a+1}, \dots, Z_b)$
- P_a^b denotes the distribution of Z_a^b

Mixing Coefficients: Measure how dependency weakens over time

$$\beta(a) = \sup_t \|P_{-\infty}^t \otimes P_{t+a}^{\infty} - P_{-\infty}^t \wedge P_{t+a}^{\infty}\|_{TV} \quad (1)$$

$$\phi(a) = \sup_t \sup_{B \in \mathcal{F}_t} \|P_{t+a}^{\infty}(\cdot|B) - P_{t+a}^{\infty}\|_{TV} \quad (2)$$

where $\mathcal{F}_t$ is the σ -algebra generated by $Z_{-\infty}^t$

A process is β -mixing or ϕ -mixing if $\beta(a) \rightarrow 0$ or $\phi(a) \rightarrow 0$ as $a \rightarrow \infty$

Rademacher Complexity: A Brief Review

Definition (Rademacher Complexity)

For a class $\mathcal{F}$ of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, the *empirical Rademacher complexity* of $\mathcal{F}$ at $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ is

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Rademacher random variables. The *Rademacher complexity* of $\mathcal{F}$ is

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{X}} \mathcal{R}(\mathcal{F}, \mathbf{X})$$

where $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$.

Intuition: $\mathcal{R}(\mathcal{F}, \mathbf{x})$ measures how well $\mathcal{F}$ can express a random sequence of labels at $\mathbf{x}$

Rademacher Complexity and Generalization

Theorem

For a class $\mathcal{F}$ of bounded functions $f : \mathcal{X} \rightarrow [-b, b]$, for any $n \geq 1, \delta \geq 0$,

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n f - P f| \leq 2 \cdot \mathcal{R}_n(\mathcal{F}) + \delta$$

with probability at least $1 - \exp\left(-\frac{\delta^2 n}{2b^2}\right)$.

Challenge: Classical results require i.i.d. data

Their work extends to:

- Block Rademacher complexity for temporal data
- Sequential Rademacher complexity for path-dependent errors

Two Types of Generalization Error

For time series prediction, generalization is not uniquely defined:

1. Path-Dependent Error: How well we predict s steps ahead on a specific trajectory

$$\mathcal{L}_{T+s}(h) = \mathbb{E}_{Z_{T+s}}[L(h(X_{T+s}), Y_{T+s}) | Z_1^T] \quad (3)$$

2. Averaged Error: How well we predict s steps ahead on an average trajectory

$$\bar{\mathcal{L}}_{T+s}(h) = \mathbb{E}_{Z_1^T}[\mathcal{L}_{T+s}(h)] = \mathbb{E}_{Z_{T+s}}[\ell(h, Z_{T+s})] \quad (4)$$

where $\ell(h, z) = L(h(x), y)$ for $z = (x, y)$

The Discrepancy Measure: Key Innovation

Discrepancy: Measures how statistical properties change over time

$$d(t_1, t_2) = \sup_{h \in \mathcal{H}} |\mathcal{L}_{t_1}(h) - \mathcal{L}_{t_2}(h)| \quad (5)$$

$$\bar{d}(t_1, t_2) = \sup_{h \in \mathcal{H}} |\bar{\mathcal{L}}_{t_1}(h) - \bar{\mathcal{L}}_{t_2}(h)| \quad (6)$$

Key Properties:

- Emerges naturally in analysis, not artificially introduced
- Integrates hypothesis class $\mathcal{H}$ and loss function L
- For strictly stationary processes: $\bar{d}(t_1, t_2) = 0$ for all t_1, t_2
- Special case: For weakly stationary processes with squared loss and linear hypotheses, $\bar{d}(t_1, t_2) = 0$
- Can be estimated from data under mild conditions

Interpretation: Measures the "degree" of non-stationarity for a specific learning problem

Independent Blocks Technique

To handle dependencies, they divide the sample Z_1^T into $2m$ blocks:

- Each block has size a_i where $T = \sum_{i=1}^{2m} a_i$
- $Z^{(i)} = Z_{l(i)}^{u(i)}$ where $l(i) = 1 + \sum_{j=1}^{i-1} a_j$ and $u(i) = \sum_{j=1}^i a_j$
- Odd blocks: $Z^o = (Z^{(1)}, Z^{(3)}, \dots, Z^{(2m-1)})$
- Even blocks: $Z^e = (Z^{(2)}, Z^{(4)}, \dots, Z^{(2m)})$

Key Result (Proposition 1): For bounded function g with $-M_1 \leq g \leq M_2$:

$$|E[g(\tilde{Z}^o)] - E[g(Z^o)]| \leq (M_1 + M_2) \sum_{i=1}^{m-1} \beta(a_{2i}) \quad (7)$$

where $\tilde{Z}^o$ consists of independent blocks with same distributions as Z^o

Sub-sample Selection Technique

Novel approach for asymptotically stationary processes:

- Given sample Z_1^T with $T = ma$ for some $m, a \geq 1$
- Define sub-samples $Z^{(j)} = (Z_{1+j}, \dots, Z_{m-1+j})$ for $j = 0, \dots, a-1$

Key Result (Proposition 2): For bounded function g with $-M_1 \leq g \leq M_2$:

$$|E[g(\tilde{Z}_\Pi)] - E[g(Z^{(j)})]| \leq (M_1 + M_2)m\beta(a) \quad (8)$$

where:

- $\beta(a) = \sup_t E[\|P_{t+a}(\cdot|Z_1^t) - \Pi\|_{TV}]$
- $\tilde{Z}_\Pi$ is an i.i.d. sample of size m from distribution Π

Particularly useful when the process converges to a stationary distribution Π

Bounding the Averaged Error (1/2)

Approach: Define $\Phi(Z_1^T) = \sup_{h \in \mathcal{H}} (\bar{\mathcal{L}}_{T+s}(h) - \frac{1}{T} \sum_{t=1}^T \ell(h, Z_t))$

Block Rademacher Complexity:

$$\mathcal{R}(Z^o) = \frac{1}{|I_1|} E \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i l(h, Z^{(2i-1)}) \right] \quad (9)$$

where:

- σ_i are Rademacher random variables
- $l(h, Z^{(2i-1)}) = \sum_t \ell(h, Z_t)$ over the i th odd block
- This adapts standard Rademacher complexity to block-structured data

Key insight: This measures how well hypothesis class can fit random noise in dependent data

Bounding the Averaged Error (2/2)

Main Result (Theorem 1): With probability $1 - \delta$, for all $h \in \mathcal{H}$:

$$\bar{\mathcal{L}}_{T+s}(h) \leq \frac{1}{T} \sum_{t=1}^T \ell(h, Z_t) + 2 \max(\mathcal{R}(Z^o), \mathcal{R}(Z^e)) \quad (10)$$

$$+ \max(\Delta^1, \Delta^2) + M \max(\|a^e\|_2, \|a^o\|_2) \sqrt{\frac{\log \frac{2}{\delta'}}{2T^2}} \quad (11)$$

where:

- $\Delta^j = \frac{1}{|I_j|} \sum_{t \in I_j} \bar{d}(t, T+s)$ is average discrepancy
- $\delta' = \delta - \sum_{i=2}^{m-1} \beta(a_i)$
- a^e and a^o are vectors of block sizes

Interpretation: Bound has three components:

- 1 Empirical error term: $\frac{1}{T} \sum_{t=1}^T \ell(h, Z_t)$
- 2 Complexity terms: Block Rademacher complexities
- 3 Non-stationarity effect: Discrepancy $\max(\Delta^1, \Delta^2)$

Implications of the Bound

Impact of Non-stationarity:

- Presence of $\max(\Delta^1, \Delta^2)$ shows dependency on "degree" of non-stationarity
- May not converge to zero for a general non-stationary process
- Effective sample size reduced from T to $(T/(\|a^e\|_2 + \|a^o\|_2))^2$

Special Cases:

- Stationary β -mixing processes: Recovers bounds of Mohri and Rostamizadeh (2009)
- Independent but non-identical: Recovers bounds of Mohri and Muñoz (2012)
- I.i.d. data: Reduces to standard Rademacher complexity bounds

Mixing is Necessary: For averaged error, mixing is a necessary condition for learning (Theorem 2)

Path-Dependent Generalization Bounds (1/2)

Approach: Uses McDiarmid's inequality for dependent variables

Sequential Rademacher Complexity:

$$\mathcal{R}_{T-s}^{\text{seq}}(\mathcal{H}_\ell) = \frac{1}{T} \sup_{z \in \mathcal{Z}^T} E \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^T \epsilon_t g(z_t(\epsilon)) \right] \quad (12)$$

Key features:

- Supremum is over all complete binary trees of depth T with values in $\mathcal{Z}$
- Extends standard Rademacher to sequential prediction problems
- Introduced by Rakhlin et al. (2010) for online learning
- Captures path-dependent nature of time series forecasting

Intuition: Measures complexity of hypothesis class on tree-structured data

Path-Dependent Generalization Bounds (2/2)

Main Result (Theorem 3): With probability $1 - \delta$, for all $h \in \mathcal{H}$:

$$\mathcal{L}_{T+s}(h) \leq \frac{1}{T} \sum_{t=1}^T \ell(h, Z_t) + E[\Delta] + 2\mathcal{R}_{T-s}^{\text{seq}}(\mathcal{H}_\ell) \quad (13)$$

$$+ M\|d\|_2 \sqrt{2 \log \frac{1}{\delta'}} + M \frac{s-1}{T} \quad (14)$$

where:

- $d = (d_1, \dots, d_T)$ with $d_t = \frac{j_t+1}{T} + \gamma_t \phi(j_t + 2) + \phi(s)$
- $\Delta = \frac{1}{T} \sum_{t=1}^{T-s} d(t+s, T+s)$ is average discrepancy
- $\gamma_t = 1$ if $j_t + t < T$ and 0 otherwise

Insight: Unlike averaged error case, this bound depends on ϕ -mixing coefficients rather than β -mixing, requiring stronger mixing conditions

Bounds for Asymptotically Stationary Processes

Definition: A process is asymptotically stationary if:

$$\beta(a) = \sup_t E[\|P_{t+a}(\cdot|Z_{-\infty}^t) - \Pi\|_{TV}] \rightarrow 0 \text{ as } a \rightarrow \infty \quad (15)$$

Examples:

- Homogeneous aperiodic irreducible Markov chains
- Many practical time series models

Long-term loss: $\mathcal{L}_{\Pi}(h) = E_{\Pi}[\ell(h, Z)]$

Key inequality: $\bar{\mathcal{L}}_T(h) \leq \mathcal{L}_{\Pi}(h) + M\beta(T)$

Main Result (Theorem 4): With probability $1 - \delta$, for all $h \in \mathcal{H}$:

$$\mathcal{L}_{\Pi}(h) \leq \frac{1}{T} \sum_{t=1}^T \ell(h, Z_t) + 2\mathcal{R}_m(\mathcal{H}, \Pi) + M\sqrt{\frac{\log \frac{a}{\delta'}}{2m}} \quad (16)$$

where $\delta' = \delta - T\beta(a)$ and $\mathcal{R}_m(\mathcal{H}, \Pi)$ is Rademacher complexity with respect to Π

Fast Learning Rates for Non-i.i.d. Data

Approach: Extends local Rademacher complexity analysis to non-i.i.d. setting

Assumption: Rademacher complexity $\mathcal{R}_m(\{g \in \mathcal{H}_\ell : E[g^2] \leq r\})$ is bounded by sub-root function $\psi(r)$ with fixed point r^*

Main Result (Theorem 5): For $T = am$, with probability $1 - \delta$, for all $h \in \mathcal{H}$:

$$\mathcal{L}_\Pi(h) \leq \left(\frac{K}{K-1} \right) \frac{1}{T} \sum_{t=1}^T \ell(h, Z_t) + C_1 r^* + \frac{C_2 \log \frac{a}{\delta'}}{m} \quad (17)$$

where $K > 1$, $\delta' = \delta - T\beta(a)$, $C_1 = 704K/M$, and $C_2 = 26MK + 11M$

Dependency on mixing rate:

- Algebraic mixing ($\beta(a) \leq Ca^{-d}$): Need $d > 2$ for fast rates
- Exponential mixing ($\beta(a) \leq Ce^{-da}$): Leads to rate $T^{-1}(\log T)^2$

Generalization with Unbounded Losses

Innovation: First generalization bounds for unbounded losses with non-i.i.d. data

Approach: Extends relative deviation bounds to asymptotically stationary processes

Define:

$$\Phi_{\tau,\alpha}(Z_1^T) = \sup_h \left(\frac{\mathcal{L}_{\Pi}(h) - \frac{1}{T} \sum_{t=1}^T \ell(h, Z_t)}{(\mathcal{L}_{\Pi,\alpha} + \tau)^{1/\alpha}} \right) \quad (18)$$

where $\mathcal{L}_{\Pi,\alpha}(h) = E_{\Pi}[\ell(h, Z)^{\alpha}]$

Result for $\alpha = 2$ (Corollary 3): With probability $1 - \delta$, for all $h \in \mathcal{H}$:

$$\mathcal{L}_{\Pi}(h) \leq \sum_{t=1}^T \ell(h, Z_t) + 2\sqrt{\mathcal{L}_{\Pi,2}(h) B_m \Gamma_0(2B_m)} \quad (19)$$

where $\Gamma_0(\epsilon) = \frac{1}{2} + \sqrt{1 + \frac{1}{2} \log \frac{1}{\epsilon}}$ and $B_m = \sqrt{\frac{2 \log E_{\Pi}[S_Q(Z_1^T)] + \log \frac{1}{\delta}}{m}}$

References

- [1] Kuznetsov, V., & Mohri, M. (2016). Generalization Bounds for Non-stationary Mixing Processes. Machine Learning.
- [2] Mohri, M., & Rostamizadeh, A. (2009). Rademacher complexity bounds for non-i.i.d. processes. In Advances in Neural Information Processing Systems.
- [3] Mohri, M., & Muñoz, A. (2012). New analysis and algorithm for learning with drifting distributions. In Algorithmic Learning Theory.
- [4] Bartlett, P. L., Bousquet, O., & Mendelson, S. (2005). Local Rademacher complexities. Ann. Statist. 33(4):1497–1537.
- [5] Rakhlin, A., Sridharan, K., & Tewari, A. (2015). Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields.