

Stability Bounds for non-i.i.d. Processes

Mehryar Mohri Afshin Rostamizadeh

Courant Institute of Mathematical Sciences
New York University

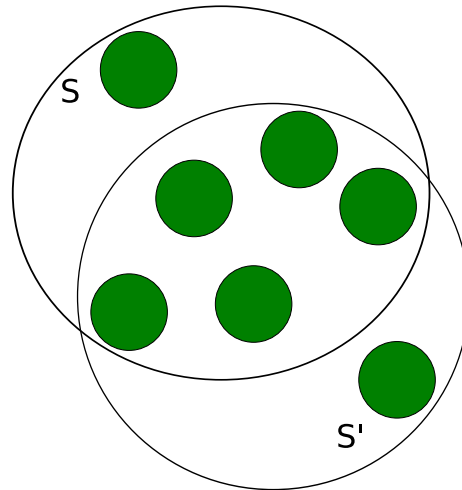
Motivation

- Learning algorithm analysis generally does not consider non-i.i.d. data.
- Many times the assumption of i.i.d. data may not hold, or may not be tested.
- We give new algorithm-specific bounds for a more general non-i.i.d. scenario.
- We will assume the standard *stationary* and *mixing* non-i.i.d. scenario.



$\tilde{\beta}$ -stability

- Let S and S' be two training sets drawn from $Z^m \in X^m \times Y^m$, the set of labeled points.



- An algorithm is called $\tilde{\beta}$ -stable if for every S and S' that differ by one point the following inequality holds,

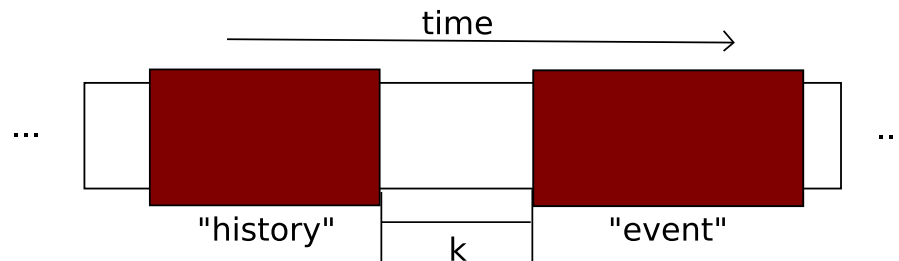
$$\sup_{z \in Z} |c(h_S, z) - c(h_{S'}, z)| \leq \tilde{\beta}$$

Mixing & Stationarity

- Although our sample is no longer i.i.d., we will assume it is *mixing*.
- We define two mixing coefficients below,

$$\beta(k) = \sup_n \mathbb{E} \sup_{B \in \sigma_{-\infty}^n} \sup_{A \in \sigma_{n+k}^{\infty}} \left| \Pr[A | B] - \Pr[A] \right|$$

$$\varphi(k) = \sup_{n, A \in \sigma_{n+k}^{\infty}, B \in \sigma_n^{-\infty}} \left| \Pr[A | B] - \Pr[A] \right|.$$



- A sequence is stationary if an event's distribution does not change with time.

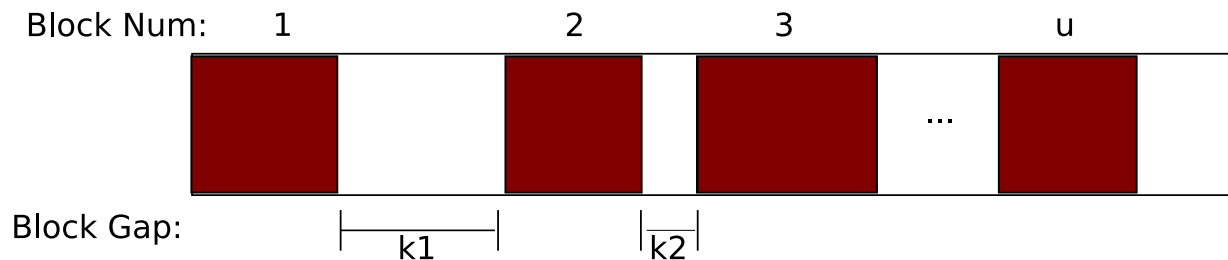
$$\forall i, j, \Pr[z_i] = \Pr[z_j]$$

The Independent-Block Method

- Theorem (Yu, 1994): For function h , bounded by M , that is defined on set of μ dependent "blocks",

$$|E[h] - \tilde{E}[h]| \leq (\mu - 1)M\beta^*$$

where \tilde{E} represents an expectation with respect to independent blocks.



- The k_i represent the space between the set of points in block i and $i + 1$. In the theorem, $\beta^* = \sup_i \beta(k_i)$.

Φ Lipschitz Condition

- Let S and S^i differ only in point i . We bound the R and \hat{R} terms separately.

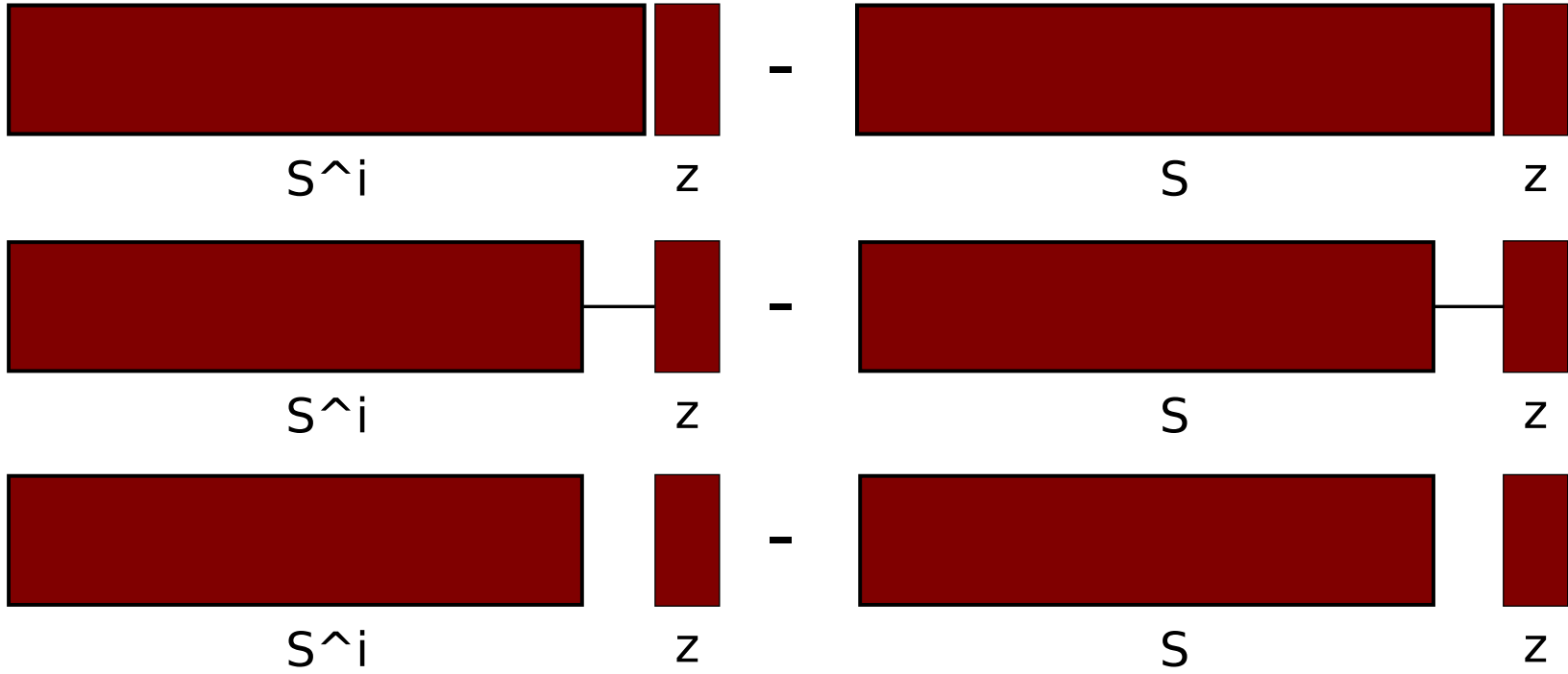
$$\begin{aligned}(1) \quad |\hat{R}(h_S) - \hat{R}(h_{S^i})| &= \frac{1}{m} \sum_{j=1}^m |c(h_S, z_j) - c(h_{S^i}, z'_j)| \\ &= \frac{1}{m} \sum_{j \neq i} |c(h_S, z_j) - c(h_{S^i}, z'_j)| + \frac{1}{m} |c(h_S, z_i) - c(h_{S^i}, z'_i)| \\ &\leq \hat{\beta} + \frac{M}{m}.\end{aligned}$$

$$\begin{aligned}(2) \quad |R(h_S) - R(h_{S^i})| &\leq |\tilde{R}(h_{S_b}) - \tilde{R}(h_{S_b^i})| + 2b\hat{\beta} + 2\beta(b)M \\ &= \mathbb{E}_{\tilde{z}}[c(h_{S_b}, \tilde{z}) - c(h_{S_b^i}, \tilde{z})] + 2b\hat{\beta} + 2\beta(b)M \\ &\leq \hat{\beta} + 2b\hat{\beta} + 2\beta(b)M.\end{aligned}$$

- (1) + (2) shows $|\Phi(S) - \Phi(S^i)| \leq (b+1)2\hat{\beta} + 2\beta(b)M + \frac{M}{m}$.

Φ Lipschitz Condition

Illustration of blocking steps used to bound $|\hat{R}(h_S) - \hat{R}(h_{S'})|$.



— signifies dependence

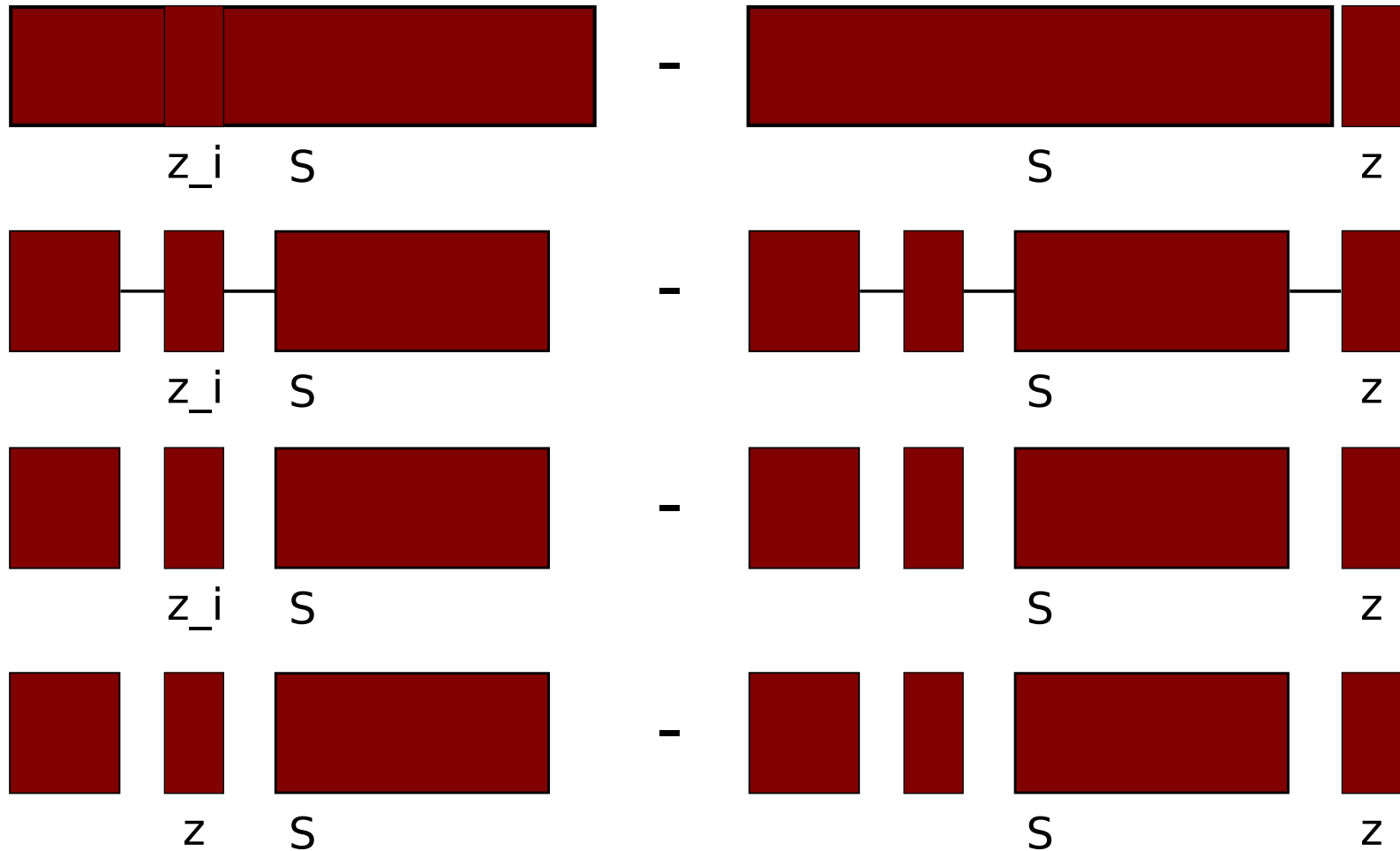
Bound on $E[\Phi]$

- Define $R(h) = 1/m \sum_i c(h, z_i)$, $\hat{R}(h) = E_z[c(h, z)]$.
- Define $\Phi(S) = R(h_S) - \hat{R}(h_S)$, we then bound $E_S[\Phi(S)]$.

$$\begin{aligned} E_S[\Phi(S)] &= E_{S,z} \left[\frac{1}{m} \sum_{i=1}^m c(h_S, z_i) - c(h_S, z) \right] \\ &\leq E_{S_{i,b},z} \left[\frac{1}{m} \sum_{i=1}^m c(h_{S_{i,b}}, z_i) - c(h_{S_{i,b}}, z) \right] + 6b\hat{\beta} \\ &\leq E_{\tilde{S}_{i,b},\tilde{z}} \left[\frac{1}{m} \sum_{i=1}^m c(h_{\tilde{S}_{i,b}}, \tilde{z}_i) - c(h_{\tilde{S}_{i,b}}, \tilde{z}) \right] + 6b\hat{\beta} + 3\beta(b)M \\ &\leq E_{\tilde{S}_{i,b},\tilde{z}} \left[\frac{1}{m} \sum_{i=1}^m c(h_{\tilde{S}_{i,b}^i}, \tilde{z}) - c(h_{\tilde{S}_{i,b}}, \tilde{z}) \right] + 6b\hat{\beta} + 3\beta(b)M \\ &\leq \hat{\beta} + 6b\hat{\beta} + 3\beta(b)M \end{aligned}$$

Bound on $E[\Phi]$

Here we illustrate the use of the “blocking method”.



Non-i.i.d. McDiarmid's Inequality

- Theorem (Kontorovich and Ramanan, 2006): For an l -Lipshitz function $\Phi : Z^m \rightarrow \mathbb{R}$ the following holds for all $\epsilon > 0$,

$$\Pr_Z[|\Phi(Z) - \mathbb{E}[\Phi(Z)]| > \epsilon] \leq 2 \exp\left(\frac{-\epsilon^2}{2ml^2 \|\Delta_m\|_\infty^2}\right),$$

where $\|\Delta_m\|_\infty \leq 1 + 2 \sum_{k=1}^m \varphi(k)$.

- This allows us to use McDiarmid's inequality for non-iid processes, with an additional term that depends on mixing properties.

Main Bound

- For an algebraic ϕ -mixing sequence ($\phi(k) = \phi_0 k^{-r}$), with mixing exponent $r > 1$, the following holds for all $\epsilon > 0$,

$$\Pr_S \left[\left| R(h_S) - \hat{R}(h_S) \right| > \epsilon + \hat{\beta} + (r + 1)6M\varphi(b) \right] \\ \leq 2 \exp \left(\frac{-\epsilon^2 (4 + 2/(r - 1))^{-2}}{2m(2\hat{\beta} + (r + 1)2M\varphi(b) + M/m)^2} \right),$$

where $\varphi(b) = \varphi_0 \left(\frac{\hat{\beta}}{r\varphi_0 M} \right)^{r/(r+1)}$.

- For $r > 1$, $\|\Delta\|_\infty < 3 + 2/(r - 1)$, and the optimal value of $b = \left(\frac{\hat{\beta}}{r\varphi_0 M} \right)^{-1/(r+1)}$.

Applications

- Given a bounded output $Y = [0, B]$, and bounded kernel $K(x, x) < \kappa, \forall x$, with probability at least $1 - \delta$ the following holds,
- Support vector regression:

$$R(h_S) \leq \hat{R}(h_S) + \frac{13\kappa^2}{2\lambda m} + 5 \left(\frac{3\kappa^2}{\lambda} + \kappa \sqrt{\frac{B}{\lambda}} \right) \sqrt{\frac{2 \ln(1/\delta)}{m}}$$

Kernel Ridge Regression (Saunders et al., 1998):

$$R(h_S) \leq \hat{R}(h_S) + \frac{26\kappa^2 B^2}{\lambda m} + 5 \left(\frac{12\kappa^2 B^2}{\lambda} + \kappa \sqrt{\frac{B}{\lambda}} \right) \sqrt{\frac{2 \ln(1/\delta)}{m}}$$

Conclusion

- We have shown the first bounds for a family of $\tilde{\beta}$ -stable algorithms in a non-i.i.d. scenario.
- It is important to note that, when we have perfect mixing ($\beta(\cdot) = \phi(\cdot) = 0$), our bounds coincide with those shown in the i.i.d. scenario.
- We would like to extend the work into a purely β -mixing scenario, this involves finding concentration bounds for β -mixing.
- We need to further investigate whether the stability assumptions can be relaxed.