Rademacher Bounds for Non-i.i.d. Processes

Afshin Rostamizadeh

Joint work with:
Mehryar Mohri

Background

Background

- Generalization Bounds - How well can we estimate an algorithm's true performance based on a finite sample?

$$\widehat{\mathrm{err}}(h, S) \overset{?}{\approx} \mathrm{E}_S[\widehat{\mathrm{err}}(h, S)] = \mathrm{err}(h, S)$$

Background

- Generalization Bounds - How well can we estimate an algorithm's true performance based on a finite sample?

$$\widehat{\mathrm{err}}(h,S) \overset{?}{\approx} \mathrm{E}_S[\widehat{\mathrm{err}}(h,S)] = \mathrm{err}(h,S)$$

- Usually, generalization bounds take the form:

$$\forall h \in H, \forall S \in (X \times Y)^m$$

$$\mathrm{err}(h,S) \leq \widehat{\mathrm{err}}(h,S) + \mathrm{complexity}(H)$$

Background

- Generalization Bounds - How well can we estimate an algorithm's true performance based on a finite sample?

$$\widehat{\mathrm{err}}(h, S) \overset{?}{\approx} \mathrm{E}_S[\widehat{\mathrm{err}}(h, S)] = \mathrm{err}(h, S)$$

- Usually, generalization bounds take the form:

$$\forall h \in H, \forall S \in (X \times Y)^m$$

$$\mathrm{err}(h, S) \leq \widehat{\mathrm{err}}(h, S) + \mathrm{complexity}(H)$$

Explicit trade-off in choice of *H*.

Background

Background

- One natural measure, Rademacher Complexity.

Background

- One natural measure, Rademacher Complexity.
- Define Rademacher r.v. as: $\Pr(\sigma_i = \pm 1) = 1/2$

Background

- One natural measure, Rademacher Complexity.
- Define Rademacher r.v. as: $\Pr(\sigma_i = \pm 1) = 1/2$
  - Empirical:

$$\widehat{\mathfrak{R}}_S(H) = \frac{2}{m} \mathrm{E}_\sigma \left[\sup_{h \in H} \left| \sum_{i=1}^{m} \sigma_i h(z_i) \right| \middle| S = (z_1, \ldots, z_m) \right]$$

Background

- One natural measure, Rademacher Complexity.
- Define Rademacher r.v. as: $\Pr(\sigma_i = \pm 1) = 1/2$
  - Empirical:

$$\widehat{\mathfrak{R}}_S(H) = \frac{2}{m} \mathrm{E}_\sigma \left[\sup_{h \in H} \left| \sum_{i=1}^{m} \sigma_i h(z_i) \right| \middle| S = (z_1, \ldots, z_m) \right]$$

  - Actual: $\mathfrak{R}_m(H) = \mathrm{E}_S[\widehat{\mathfrak{R}}_S]$

Background

- One natural measure, Rademacher Complexity.
- Define Rademacher r.v. as: $\Pr(\sigma_i = \pm 1) = 1/2$
  - Empirical:

$$\widehat{\mathfrak{R}}_S(H) = \frac{2}{m}\mathrm{E}_\sigma\left[\sup_{h \in H}\left|\sum_{i=1}^{m}\sigma_i h(z_i)\right| \middle| S = (z_1, \ldots, z_m)\right]$$

  - Actual: $\mathfrak{R}_m(H) = \mathrm{E}_S[\widehat{\mathfrak{R}}_S]$
  - Intuitively, this measures the ability of a hypothesis class to fit uniform random noise.

Background

- One natural measure, Rademacher Complexity.
- Define Rademacher r.v. as: $\Pr(\sigma_i = \pm 1) = 1/2$
  - Empirical:

$$\widehat{\mathfrak{R}}_S(H) = \frac{2}{m} \mathrm{E}_\sigma \left[\sup_{h \in H} \left| \sum_{i=1}^{m} \sigma_i h(z_i) \right| \middle| S = (z_1, \ldots, z_m) \right]$$

  - Actual: $\mathfrak{R}_m(H) = \mathrm{E}_S[\widehat{\mathfrak{R}}_S]$
  - Intuitively, this measures the ability of a hypothesis class to fit uniform random noise.
  - Can be measured from data, tighter bounds.

Background

Background

- Rademacher Generalization Bounds (0/1 loss)

  [Bartlett, Mendelson '01, Koltchinskii, Panchenko '00]:

$$\forall h \in H, \forall S \in (X \times Y)^m$$

$$\mathrm{err}(h, S) \leq \widehat{\mathrm{err}}(h, S) + \frac{\widehat{\mathfrak{R}}_S(H)}{2} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Background

- Rademacher Generalization Bounds (0/1 loss)

  [Bartlett, Mendelson '01, Koltchinskii, Panchenko '00]:

$$\forall h \in H, \forall S \in (X \times Y)^m$$

$$\text{err}(h,S) \le \widehat{\text{err}}(h,S) + \frac{\widehat{\mathfrak{R}}_S(H)}{2} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

- CRITICAL Assumption: The sample must be identically and independently distributed (i.i.d.).

Motivation

Motivation

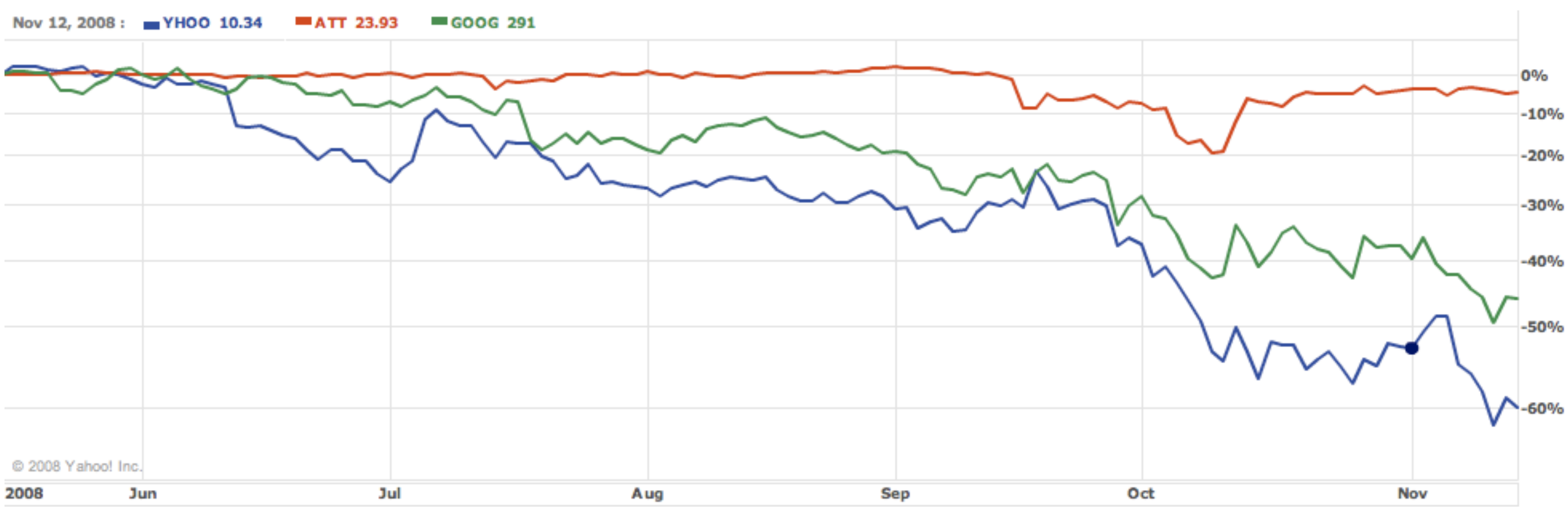
- In practice data often contains dependencies. Here we consider temporal dependancies:

Motivation

- In practice data often contains dependencies. Here we consider temporal dependancies:
    - We assume the distribution to be mixing; implies a dependence which weakens over time.

Motivation

- In practice data often contains dependencies. Here we consider temporal dependancies:
  - We assume the distribution to be mixing; implies a dependence which weakens over time.
  - Natural in the context of time-series analysis (i.e. stock market quotes).



Motivation

Motivation

- We MUST deal with non-i.i.d. data!

Motivation

- We MUST deal with non-i.i.d. data!
- Often in practice, simply use algorithm designed for the i.i.d. case. Many times, such algorithms still perform well

Motivation

- We MUST deal with non-i.i.d. data!
- Often in practice, simply use algorithm designed for the i.i.d. case. Many times, such algorithms still perform well
- How can we justify/guarantee this performance?

Motivation

- We MUST deal with non-i.i.d. data!
- Often in practice, simply use algorithm designed for the i.i.d. case. Many times, such algorithms still perform well
- How can we justify/guarantee this performance?
- Give generalization bounds that do NOT assume i.i.d. data!

Motivation

- We MUST deal with non-i.i.d. data!
- Often in practice, simply use algorithm designed for the i.i.d. case. Many times, such algorithms still perform well
- How can we justify/guarantee this performance?
- Give generalization bounds that do NOT assume i.i.d. data!
- Here we present general proof techniques useful for mixing processes.

Motivation

- We MUST deal with non-i.i.d. data!
- Often in practice, simply use algorithm designed for the i.i.d. case. Many times, such algorithms still perform well
- How can we justify/guarantee this performance?
- Give generalization bounds that do NOT assume i.i.d. data!
- Here we present general proof techniques useful for mixing processes.
- We extend useful properties of Rademacher complexity to this non-i.i.d. setting.

Definitions

Definitions

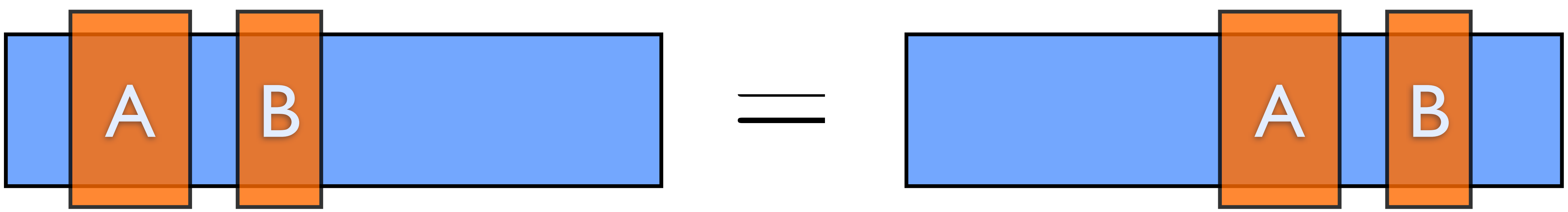
- We make the standard assumption of Stationarity.

[Stationarity] A sequence of random variables $= Z_{t}{}_{t=-\infty}^{\infty}$ is said to be *stationary* if for any t and non-negative integers m and k, the random vectors $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ have the same distribution.

Definitions

- We make the standard assumption of Stationarity.

[Stationarity] A sequence of random variables $= Z_t{}_{t=-\infty}^{\infty}$ is said to be *stationary* if for any t and non-negative integers m and k, the random vectors $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ have the same distribution.

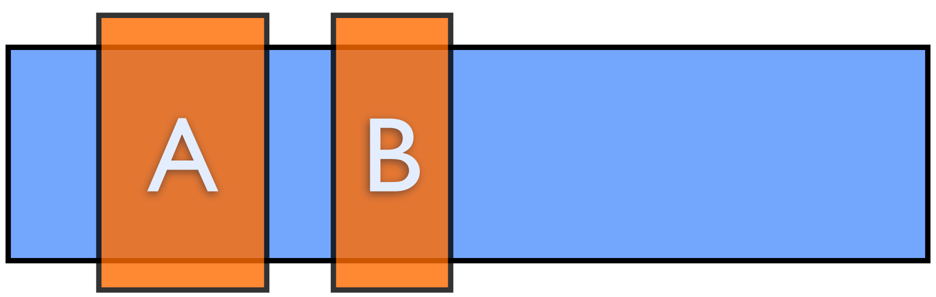


$\mathrm{P}(\mathrm{x}_t = B | x_{t-s} = A)$ $\mathrm{P}(\mathrm{x}_{t+k} = B | x_{t-s+k} = A)$

Definitions

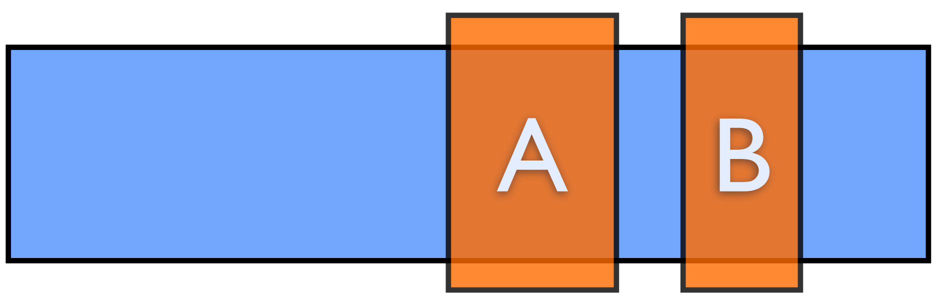
- We make the standard assumption of Stationarity.

[Stationarity] A sequence of random variables $= Z_t{}_{t=-\infty}^{\infty}$ is said to be *stationary* if for any t and non-negative integers m and k, the random vectors $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+m+k})$ have the same distribution.



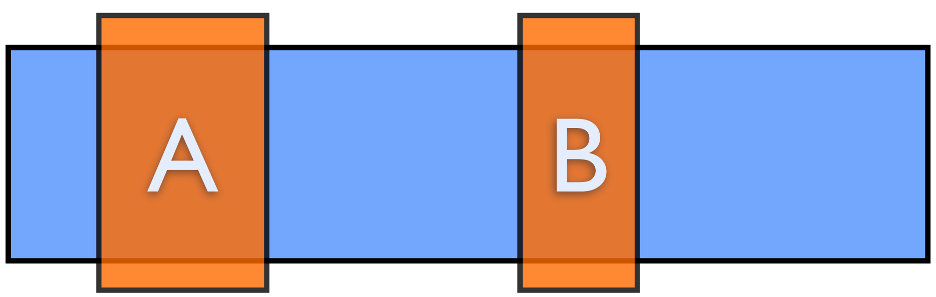
$\mathrm{P}(\mathrm{x}_t = B | x_{t-s} = A)$

$$=$$



$\mathrm{P}(\mathrm{x}_{t+k} = B | x_{t-s+k} = A)$

$$\neq$$



$\mathrm{P}(\mathrm{x}_{t+k} = B | x_{t-s+l} = A)$

(relative distance matters)

Definitions

Definitions

- We quantify dependence with natural β-mixing coefficient:

[β-mixing] Let $= Z_t{}_{t=-\infty}^{\infty}$ be a stationary sequence of random variables. Let σ_i^j denote the σ-algebra generated by the random variables Z_k, $i \leq k \leq j$. The β-mixing coefficient of the stochastic process is defined as

$$\beta(k) = \sup_n {}_{B \in \sigma_{-\infty}^n} \Big[\sup_{A \in \sigma_{n+k}^{\infty}} \Big| \Pr[A \mid B] - \Pr[A] \Big| \Big].$$

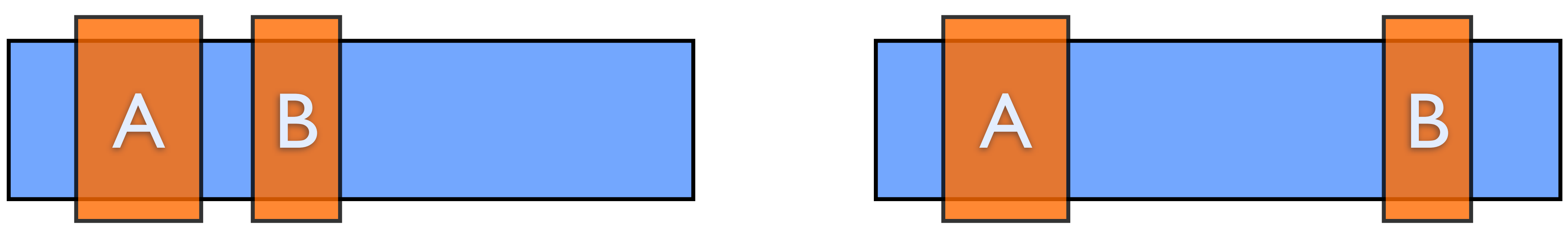
Definitions

- We quantify dependence with natural β-mixing coefficient:

[β-mixing] Let $= Z_t{}_{t=-\infty}^{\infty}$ be a stationary sequence of random variables. Let σ_i^j denote the σ-algebra generated by the random variables Z_k, $i \leq k \leq j$. The β-mixing coefficient of the stochastic process is defined as

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

Mixing implies:



$$\Big| \Pr[B|A] - \Pr[B] \Big| \quad \geq \quad \Big| \Pr[B|A] - \Pr[B] \Big|$$

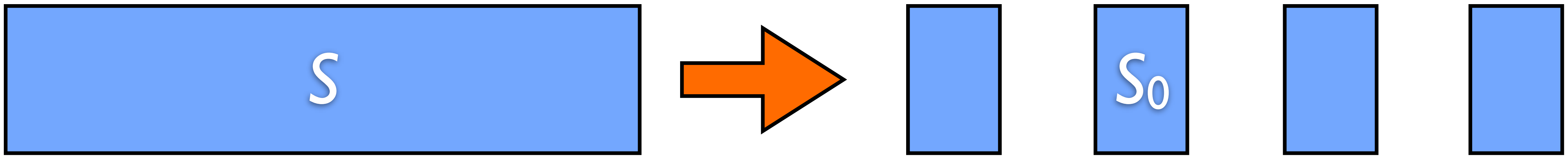
Proof Strategy

Proof Strategy

- Reduce dependent scenario to the independent case.

Proof Strategy

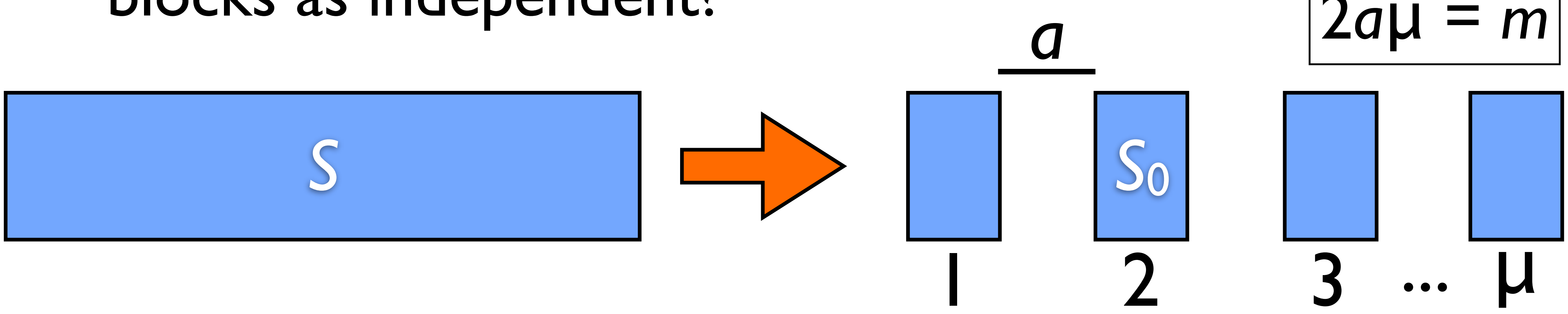
- Reduce dependent scenario to the independent case.
- If we introduce gaps in the sequence, can we treat the blocks as independent?



Proof Strategy

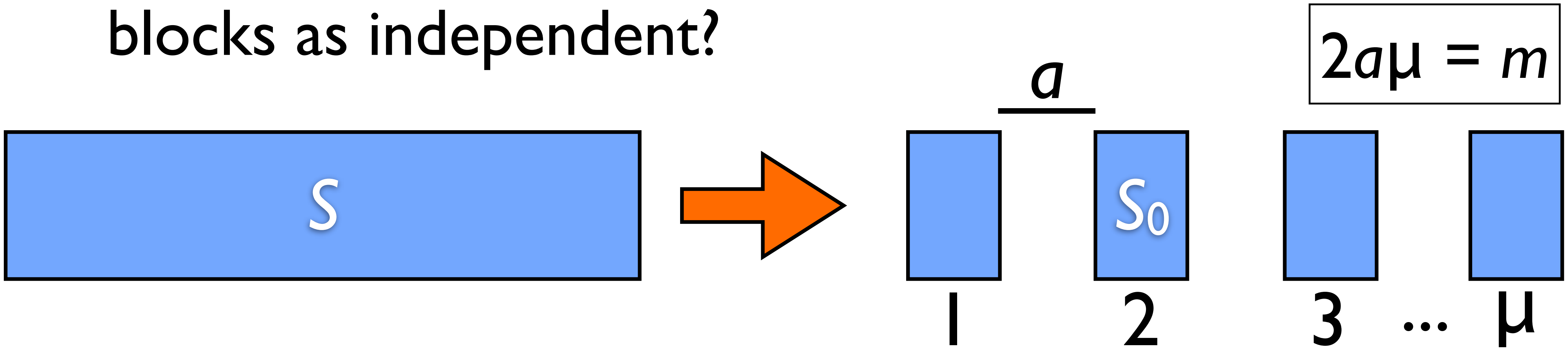
- Reduce dependent scenario to the independent case.
- If we introduce gaps in the sequence, can we treat the blocks as independent?

$\boxed{2a\mu = m}$



Proof Strategy

- Reduce dependent scenario to the independent case.
- If we introduce gaps in the sequence, can we treat the blocks as independent?



- β-mixing assumption allows us to exactly bound this approximation.

Lemma [Yu, '94] Let S_0 be defined as above and let h be a function of S_0 that is bounded by M and then,

$$|\mathrm{E}_{S_0}[h] - \mathrm{E}_{\tilde{S}_0}[h]| \leq (\mu - 1)M\beta(a) \ ,$$

where E_{S_0} (resp. $\mathrm{E}_{\tilde{S}_0}$) denotes the expectation with respect to the dependent (resp. independent) block sequence.

Proof Strategy

Proof Strategy

- In i.i.d. case, apply McDiarmid's inequality to:

$$\Phi(S) = \sup_{h \in H} R(h) - \widehat{R}_S(h)$$

where,

$$\widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} h(z_i) \qquad R(h) = \mathrm{E}_S[\widehat{R}_S(h)]$$

Proof Strategy

- In i.i.d. case, apply McDiarmid's inequality to:

$$\Phi(S) = \sup_{h \in H} R(h) - \widehat{R}_S(h)$$

where,

$$\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} h(z_i) \qquad R(h) = \mathrm{E}_S[\widehat{R}_S(h)]$$

- We apply it to i.i.d. blocks, extending *H* in a natural way.

Define $h_a(B) = \frac{1}{a}\sum_{i=1}^{a} h(z_i)$ for any block $B = (z_1, \ldots, z_a) \in Z^a$, and define H_a as the set of all block-based hypotheses h_a generated from $h \in H$.

Preparation Step

Preparation Step

- Re-write in terms of blocks:

Preparation Step

- Re-write in terms of blocks:

$$\begin{aligned}
\Pr_S[\Phi(S) > \epsilon] &= \Pr_S[\sup_h (R(h) - \widehat{R}_S(h)) > \epsilon] \\
&= \Pr_S\left[\sup_h \left(\frac{R(h) - \widehat{R}_{S_0}(h)}{2} + \frac{R(h) - \widehat{R}_{S_1}(h)}{2}\right) > \epsilon\right] && (\text{def. of } \widehat{R}_S(h)) \\
&\leq \Pr_S[\Phi(S_0) + \Phi(S_1) > 2\epsilon] && (\text{def. of } \Phi) \\
&\leq \Pr_{S_0}[\Phi(S_0) > \epsilon] + \Pr_{S_1}[\Phi(S_1) > \epsilon] && (\text{union bound}) \\
&= 2 \Pr_{S_0}[\Phi(S_0) > \epsilon] && (\text{stationarity}) \\
&= 2 \Pr_{S_0}[\Phi(S_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon']. && (\text{def. of } \epsilon')
\end{aligned}$$

Preparation Step

- Re-write in terms of blocks:

$$\boxed{\epsilon' = \epsilon - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)]}$$

$$\begin{aligned}
\Pr_S[\Phi(S) > \epsilon] &= \Pr_S[\sup_h (R(h) - \widehat{R}_S(h)) > \epsilon] \\
&= \Pr_S\Big[\sup_h \Big(\frac{R(h) - \widehat{R}_{S_0}(h)}{2} + \frac{R(h) - \widehat{R}_{S_1}(h)}{2}\Big) > \epsilon\Big] && (\text{def. of } \widehat{R}_S(h)) \\
&\leq \Pr_S[\Phi(S_0) + \Phi(S_1) > 2\epsilon] && (\text{def. of } \Phi) \\
&\leq \Pr_{S_0}[\Phi(S_0) > \epsilon] + \Pr_{S_1}[\Phi(S_1) > \epsilon] && (\text{union bound}) \\
&= 2 \Pr_{S_0}[\Phi(S_0) > \epsilon] && (\text{stationarity}) \\
&= 2 \Pr_{S_0}[\Phi(S_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon']. && (\text{def. of } \epsilon')
\end{aligned}$$

Preparation Step

- Re-write in terms of blocks:

$$\boxed{\epsilon' = \epsilon - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)]}$$

$$\begin{aligned}
\Pr_S[\Phi(S) > \epsilon] &= \Pr_S[\sup_h (R(h) - \widehat{R}_S(h)) > \epsilon] \\
&= \Pr_S\Big[\sup_h \Big(\frac{R(h) - \widehat{R}_{S_0}(h)}{2} + \frac{R(h) - \widehat{R}_{S_1}(h)}{2}\Big) > \epsilon\Big] && (\text{def. of } \widehat{R}_S(h)) \\
&\leq \Pr_S[\Phi(S_0) + \Phi(S_1) > 2\epsilon] && (\text{def. of } \Phi) \\
&\leq \Pr_{S_0}[\Phi(S_0) > \epsilon] + \Pr_{S_1}[\Phi(S_1) > \epsilon] && (\text{union bound}) \\
&= 2 \Pr_{S_0}[\Phi(S_0) > \epsilon] && (\text{stationarity}) \\
&= 2 \Pr_{S_0}[\Phi(S_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon']. && (\text{def. of } \epsilon')
\end{aligned}$$

- Obtain independent blocks, using Yu's Lemma:

$$2 \Pr_{S_0}[\Phi(S_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon'] \leq 2 \Pr_{\tilde{S}_0}[\Phi(\tilde{S}_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon'] + 2(\mu - 1)\beta(a)$$

Concentration Bound

Concentration Bound

- Now, apply McDiarmid's inequality to i.i.d. blocks:

Concentration Bound

- Now, apply McDiarmid's inequality to i.i.d. blocks:

Changing a block $\tilde{Z}_k$ of the sample $\tilde{S}_0$ can change $\Phi(\tilde{S}_0)$ by at most $\frac{1}{\mu}|h(\tilde{Z}_k)| \leq M/\mu$ and by McDiarmid's inequality, the following holds for any $\epsilon > 2(\mu - 1)M\beta(a)$:

$$\Pr_{\tilde{S}_0}[\Phi(\tilde{S}_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon']$$

$$\leq \exp\left(\frac{-2\epsilon'^2}{\sum_{i=1}^{\mu}(M/\mu)^2}\right) = \exp\left(\frac{-2\mu\epsilon'^2}{M^2}\right)$$

Concentration Bound

- Now, apply McDiarmid's inequality to i.i.d. blocks:

Changing a block $\tilde{Z}_k$ of the sample $\tilde{S}_0$ can change $\Phi(\tilde{S}_0)$ by at most $\frac{1}{\mu}|h(\tilde{Z}_k)| \leq M/\mu$ and by McDiarmid's inequality, the following holds for any $\epsilon > 2(\mu - 1)M\beta(a)$:

$$\Pr_{\tilde{S}_0}[\Phi(\tilde{S}_0) - \mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] > \epsilon']$$

$$\leq \exp\left(\frac{-2\epsilon'^2}{\sum_{i=1}^{\mu}(M/\mu)^2}\right) = \exp\left(\frac{-2\mu\epsilon'^2}{M^2}\right)$$

- So far, we have:

$$\Pr_{S}[\Phi(S) > \epsilon] \leq 2\exp\left(\frac{-2\mu\epsilon'^2}{M^2}\right) + 2(\mu - 1)\beta(a),$$

Bounding the Expectation

Bounding the Expectation

- To make the bound useful, we must bound the expectation (over i.i.d. blocks):

Bounding the Expectation

- To make the bound useful, we must bound the expectation (over i.i.d. blocks):

$$
\begin{aligned}
\mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] \leq & \mathrm{E}_{\tilde{S}_0,\tilde{S}'_0}[\sup_{h\in H} \widehat{R}_{\tilde{S}'_0}(h) - \widehat{R}_{\tilde{S}_0}(h)] \\
= & \mathrm{E}_{\tilde{S}_0,\tilde{S}'_0}\left[\sup_{h_a\in H_a} \frac{1}{\mu}\sum_{i=1}^{\mu} h_a(Z_i) - h_a(Z'_i)\right] && \text{(def. of } \widehat{R}) \\
= & \mathrm{E}_{\tilde{S}_0,\tilde{S}'_0,\sigma}\left[\sup_{h_a\in H_a} \frac{1}{\mu}\sum_{i=1}^{\mu} \sigma_i(h_a(Z_i) - h_a(Z'_i))\right] && \text{(Rad. var.'s)} \\
\leq & \mathrm{E}_{\tilde{S}_0,\tilde{S}'_0,\sigma}\left[\sup_{h_a\in H_a} \frac{1}{\mu}\sum_{i=1}^{\mu} \sigma_i h_a(Z_i)\right] \\
& + \mathrm{E}_{\tilde{S}_0,\tilde{S}'_0,\sigma}\left[\sup_{h_a\in H_a} \frac{1}{\mu}\sum_{i=1}^{\mu} \sigma_i h_a(Z'_i)\right] && \text{(prop. of sup)} \\
= & \, 2\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h_a\in H_a} \frac{1}{\mu}\sum_{i=1}^{\mu} \sigma_i h_a(Z_i)\right].
\end{aligned}
$$

Bounding the Expectation

Bounding the Expectation

- Would like complexity of H not H_a:

Bounding the Expectation

- Would like complexity of *H* not H_a:

$$\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h_a\in H_a}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h_a(Z_i)\right]=\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h\in H}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i\frac{1}{a}\sum_{j=1}^{a}h(z_j^{(i)})\right]$$

Bounding the Expectation

- Would like complexity of *H* not H_a:

$$\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h_a \in H_a} \frac{2}{\mu} \sum_{i=1}^{\mu} \sigma_i h_a(Z_i)\right] = \mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h \in H} \frac{2}{\mu} \sum_{i=1}^{\mu} \sigma_i \frac{1}{a} \sum_{j=1}^{a} h(z_j^{(i)})\right]$$

- Almost, looks like Rad. comp. but σ's are shared.

Bounding the Expectation

- Would like complexity of *H* not H_a:

$$\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h_a \in H_a} \frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h_a(Z_i)\right] = \mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h \in H} \frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i \frac{1}{a}\sum_{j=1}^{a} h(z_j^{(i)})\right]$$

- Almost, looks like Rad. comp. but σ's are shared.

$$\begin{aligned}
\mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] &\leq \mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h \in H} \frac{1}{a}\sum_{j=1}^{a}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h(z_j^{(i)})\right] && \text{(reversing order of sums)} \\
&\leq \frac{1}{a}\sum_{j=1}^{a}\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h \in H}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h(z_j^{(i)})\right] && \text{(convexity of sup)} \\
&= \frac{1}{a}\sum_{j=1}^{a}\mathrm{E}_{\tilde{S}_0^j,\sigma}\left[\sup_{h \in H}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h(z_j^{(i)})\right] && \text{(marginalization)} \\
&= \mathrm{E}_{\tilde{S}_\mu,\sigma}\left[\sup_{h \in H}\frac{2}{\mu}\sum_{z \in \tilde{S}_\mu}\sigma_i h(z)\right] \leq \mathfrak{R}_\mu^{\tilde{D}}(H).
\end{aligned}$$

Bounding the Expectation

- Would like complexity of H not H_a:

$$\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h_a\in H_a}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h_a(Z_i)\right] = \mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h\in H}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i\frac{1}{a}\sum_{j=1}^{a}h(z_j^{(i)})\right]$$

- Almost, looks like Rad. comp. but σ's are shared.

$$\begin{aligned}
\mathrm{E}_{\tilde{S}_0}[\Phi(\tilde{S}_0)] \leq & \mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h\in H}\frac{1}{a}\sum_{j=1}^{a}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h(z_j^{(i)})\right] && \text{(reversing order of sums)}\\
\leq & \frac{1}{a}\sum_{j=1}^{a}\mathrm{E}_{\tilde{S}_0,\sigma}\left[\sup_{h\in H}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h(z_j^{(i)})\right] && \text{(convexity of sup)}\\
= & \frac{1}{a}\sum_{j=1}^{a}\mathrm{E}_{\tilde{S}_0^j,\sigma}\left[\sup_{h\in H}\frac{2}{\mu}\sum_{i=1}^{\mu}\sigma_i h(z_j^{(i)})\right] && \text{(marginalization)}\\
= & \mathrm{E}_{\tilde{S}_\mu,\sigma}\left[\sup_{h\in H}\frac{2}{\mu}\sum_{z\in\tilde{S}_\mu}\sigma_i h(z)\right] \leq \mathfrak{R}_\mu^{\tilde{D}}(H).
\end{aligned}$$

$\tilde{D}$ Denotes i.i.d. distribution

Bound So Far

Bound So Far

- We have a bound in terms of the Rad. complexity.

With probability at least $1-\delta$ and $\delta' = \delta - 2(\mu - 1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_{\mu}^{\tilde{D}}(H) + M\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}}$$

Bound So Far

- We have a bound in terms of the Rad. complexity.

With probability at least $1-\delta$ and $\delta' = \delta - 2(\mu-1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_\mu^{\tilde{D}}(H) + M\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}}$$

- Would like bounds in terms of empirical Rademacher complexity. It has many benefits:

Bound So Far

- We have a bound in terms of the Rad. complexity.

With probability at least $1-\delta$ and $\delta' = \delta - 2(\mu-1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_\mu^{\tilde{D}}(H) + M\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}}$$

- Would like bounds in terms of empirical Rademacher complexity. It has many benefits:
  - Can be measured from data (tighter bounds).

Bound So Far

- We have a bound in terms of the Rad. complexity.

With probability at least $1-\delta$ and $\delta' = \delta - 2(\mu-1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_\mu^{\tilde{D}}(H) + M\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}}$$

- Would like bounds in terms of empirical Rademacher complexity. It has many benefits:
  - Can be measured from data (tighter bounds).
  - Can be related to other complexity measures (e.g. VC-dimension).

Bound So Far

- We have a bound in terms of the Rad. complexity.

With probability at least $1-\delta$ and $\delta' = \delta - 2(\mu-1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_\mu^{\tilde{D}}(H) + M\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}}$$

- Would like bounds in terms of empirical Rademacher complexity. It has many benefits:
  - Can be measured from data (tighter bounds).
  - Can be related to other complexity measures (e.g. VC-dimension).
  - Can be bounded for specific hypotheses.

Bound So Far

Bound So Far

- Using similar techniques, we can show $\mathfrak{R}_{\mu}^{\tilde{D}}$ is close to $\widehat{\mathfrak{R}}_{S_\mu}$:

Bound So Far

- Using similar techniques, we can show $\mathfrak{R}_\mu^{\tilde{D}}$ is close to $\widehat{\mathfrak{R}}_{S_\mu}$:

With probability at least $1-\delta$ and $\delta' = \delta/2 - 2(\mu-1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \widehat{\mathfrak{R}}_{S_\mu}(H) + 3M\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}},$$

Bound So Far

- Using similar techniques, we can show $\mathfrak{R}_\mu^{\tilde{D}}$ is close to $\widehat{\mathfrak{R}}_{S_\mu}$:

With probability at least $1-\delta$ and $\delta' = \delta/2 - 2(\mu-1)\beta(a)$, the following inequality holds for all hypotheses $h \in H$:

$$R(h) \leq \widehat{R}_S(h) + \widehat{\mathfrak{R}}_{S_\mu}(H) + 3M\sqrt{\frac{\log\frac{2}{\delta'}}{2\mu}},$$

- Kernel hypotheses bounds [Bartlett, Mendelson '01]:

In the case of classification with hypotheses based on a kernel K and a weight vector w bounded by B, $\|w\| \leq B$, the empirical Rademacher complexity can be bounded as follows: $\widehat{\mathfrak{R}}_{S_\mu}(H) \leq \frac{2B}{\mu}\sqrt{[K]}$

Classification Bound

Classification Bound

- Let H be the set of hypotheses $\left\{(x,y) \in Z \mapsto y \sum_{i=1}^{m} \alpha_i K(x_i, x) : \sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \leq 1\right\}$.

Classification Bound

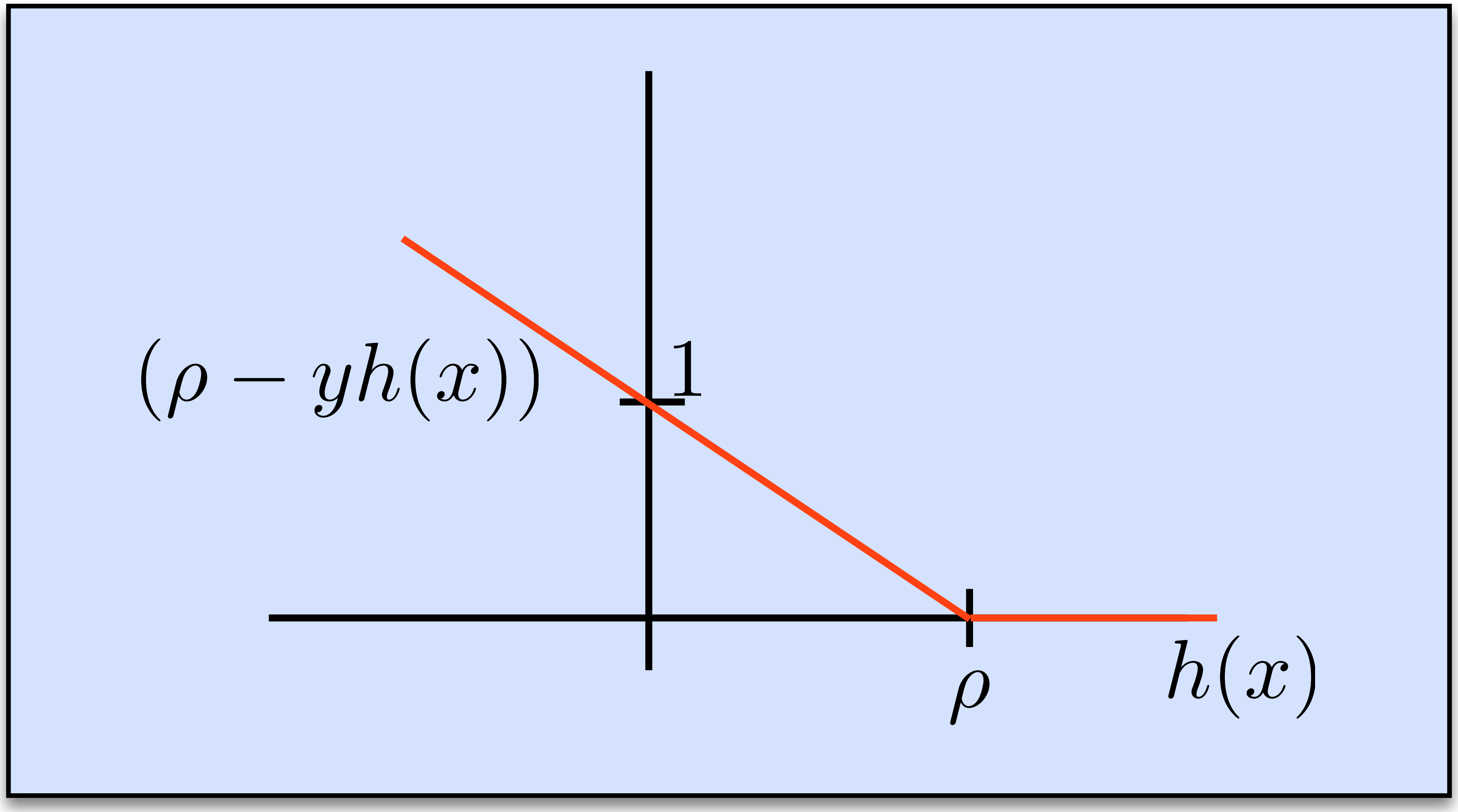
- Let H be the set of hypotheses $\left\{(x, y) \in Z \mapsto y \sum_{i=1}^{m} \alpha_i K(x_i, x) : \sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \leq 1\right\}$.

- Let $\widehat{R}_S^\rho(h)$ denote the average amount by which $y_i h(x_i)$ deviates from the margin ρ: $\widehat{R}_S^\rho(h) = \frac{1}{m} \sum_{i=1}^{m} (\rho - y_i h(x_i))_+$.

Classification Bound

- Let H be the set of hypotheses $\left\{(x, y) \in Z \mapsto y \sum_{i=1}^{m} \alpha_i K(x_i, x) : \sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \leq 1\right\}$.

- Let $\widehat{R}_S^{\rho}(h)$ denote the average amount by which $y_i h(x_i)$ deviates from the margin ρ: $\widehat{R}_S^{\rho}(h) = \frac{1}{m} \sum_{i=1}^{m} (\rho - y_i h(x_i))_+$.



Classification Bound

Classification Bound

With probability at least $1-\delta$, the following inequality holds for all hypotheses $h \in H$:

$$\Pr[yh(x) \leq 0] \leq \frac{1}{\rho}\widehat{R}_S^{\rho}(h) + \frac{4}{\mu\rho}\sqrt{[\mathbf{K}]} + 3\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}},$$

where $\delta' = \delta/2 - 2(\mu-1)\beta(a)$, and $\mathbf{K}$ is the Gram matrix of the kernel K for the sample S.

Classification Bound

With probability at least $1-\delta$, the following inequality holds for all hypotheses $h \in H$:

$$\Pr[yh(x) \leq 0] \leq \frac{1}{\rho}\widehat{R}_S^{\rho}(h) + \frac{4}{\mu\rho}\sqrt{[\mathbf{K}]} + 3\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}},$$

where $\delta' = \delta/2 - 2(\mu-1)\beta(a)$, and $\mathbf{K}$ is the Gram matrix of the kernel K for the sample S.

- Now, we need to appropriately choose the parameters *a* and μ.

Classification Bound

With probability at least $1 - \delta$, the following inequality holds for all hypotheses $h \in H$:

$$\Pr[yh(x) \leq 0] \leq \frac{1}{\rho} \widehat{R}_S^{\rho}(h) + \frac{4}{\mu\rho} \sqrt{[\mathbf{K}]} + 3\sqrt{\frac{\log \frac{2}{\delta'}}{2\mu}},$$

where $\delta' = \delta/2 - 2(\mu - 1)\beta(a)$, and $\mathbf{K}$ is the Gram matrix of the kernel K for the sample S.

- Now, we need to appropriately choose the parameters *a* and μ.
- If we assume algebraic mixing, $\beta(a) := \beta_0 a^{-r}$, one suitable choice:

$$\mu = \frac{m^{\frac{2r+1}{2r+4}}}{2}$$

A Complete Bound

A Complete Bound

Assuming that the sample is drawn from a stationary algebraically β-mixing distribution, $\beta(a) = \beta_0 a^{-r}$, the following bound holds,

$$\Pr[yh(x) \leq 0] \leq \frac{1}{\rho}\widehat{R}_S^{\rho}(h) + \frac{8Rm^{\gamma_1}}{\rho} + 3m^{\gamma_2}\sqrt{\log \frac{2}{\delta'}},$$

where $\gamma_1 = \frac{1}{2}\big(\frac{3}{r+2} - 1\big)$, $\gamma_2 = \frac{1}{2}\big(\frac{3}{2r+4} - 1\big)$ and $\delta' = \delta/2 - \beta_0 m^{\gamma_1}$.

A Complete Bound

Assuming that the sample is drawn from a stationary algebraically β-mixing distribution, $\beta(a) = \beta_0 a^{-r}$, the following bound holds,

$$\Pr[yh(x) \leq 0] \leq \frac{1}{\rho} \widehat{R}_S^{\rho}(h) + \frac{8Rm^{\gamma_1}}{\rho} + 3m^{\gamma_2} \sqrt{\log \frac{2}{\delta'}},$$

where $\gamma_1 = \frac{1}{2}\left(\frac{3}{r+2} - 1\right)$, $\gamma_2 = \frac{1}{2}\left(\frac{3}{2r+4} - 1\right)$ and $\delta' = \delta/2 - \beta_0 m^{\gamma_1}$.

- As $r \rightarrow \infty$ and $\beta_0 \rightarrow 0$ (i.e. the i.i.d. scenario is approached), this bound has the same asymptotic behavior as the i.i.d. bound.

Summary

Summary

- Have given the first data-dependent bounds for a non-i.i.d. scenario.

Summary

- Have given the first data-dependent bounds for a non-i.i.d. scenario.
- First known margin-based classification bounds (can be extended to regression as well).

Summary

- Have given the first data-dependent bounds for a non-i.i.d. scenario.
- First known margin-based classification bounds (can be extended to regression as well).
- Can easily extend bounds to other complexity measures via Rademacher complexity.

Summary

- Have given the first data-dependent bounds for a non-i.i.d. scenario.
- First known margin-based classification bounds (can be extended to regression as well).
- Can easily extend bounds to other complexity measures via Rademacher complexity.
- Future work:

Summary

- Have given the first data-dependent bounds for a non-i.i.d. scenario.
- First known margin-based classification bounds (can be extended to regression as well).
- Can easily extend bounds to other complexity measures via Rademacher complexity.
- Future work:
  - Can we make use of the entire sample to computer empirical Rademacher complexity?

Summary

- Have given the first data-dependent bounds for a non-i.i.d. scenario.
- First known margin-based classification bounds (can be extended to regression as well).
- Can easily extend bounds to other complexity measures via Rademacher complexity.
- Future work:
  - Can we make use of the entire sample to computer empirical Rademacher complexity?
  - Can we strictly generalize the i.i.d. bound?