Generalization Performance

Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

May 30, 2008 / Statistical Learning Theory II

Outline

1. Bounding risk
   - Risk and empirical risk
   - Concentration inequalities
   - Bound for finite function class
2. Risk bound for infinite function class
   - Techniques for infinite function class
   - Rademacher average, growth function, and VC-dimension
3. Risk bound for SVM
   - Risk bound for SVM

Risk and empirical risk I: Terminology

- Supervised learning:
  - $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$: data. i.i.d. sample.
  - $X_i \in \mathcal{X}$: input, $Y_i \in \mathcal{Y}$: output.
  - $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$: function class.
  - Choose f from $\mathcal{F}$ so that $Y_i \approx f(X_i)$.
- Risk and empirical risk
  - Loss function $\ell(y, f)$: measure discrepancy of Y_i and $f(X_i)$.
  - Risk: the purpose of learning is to minimize the risk;

$$L(f) = E[\ell(Y, f(X))] \qquad (f \in \mathcal{F}).$$

  - Empirical risk:

$$L_n(f) = \widehat{E}_n[\ell(Y, f(X))] = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) \qquad (f \in \mathcal{F}).$$

  - Learning must be done with data:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} L_n(f).$$

Risk and empirical risk II: Example of loss function

- Mean square error.
  - $\ell(y, f) = (y - f)^2$.
  - Empirical risk
    $\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$ (least mean square).
  - Risk = $E[(Y - f(X))^2]$.
- 0-1 loss. $y, f(x) \in \{\pm 1\}$.
  - $\ell(y, f) = \frac{1 - yf(x)}{2}$.
  - Empirical risk = ratio of errors:
    $$\widehat{E}_n[\ell(Y, f(X))] = \frac{1}{n}|\{i \mid Y_i \neq f(X_i)\}|.$$
  - Risk = mean error rate: $E[\ell(Y, f(X))] = \Pr(Y \neq f(X))$.
- Log likelihood
  - $\ell(y, f) = -\log p(y|f)$.
  - Empirical risk = - Empirical log likelihood.
  - Risk = - Expected log likelihood.

Risk and empirical risk III: Two approaches

- Goal: What can we say about $L(\widehat{f})$?

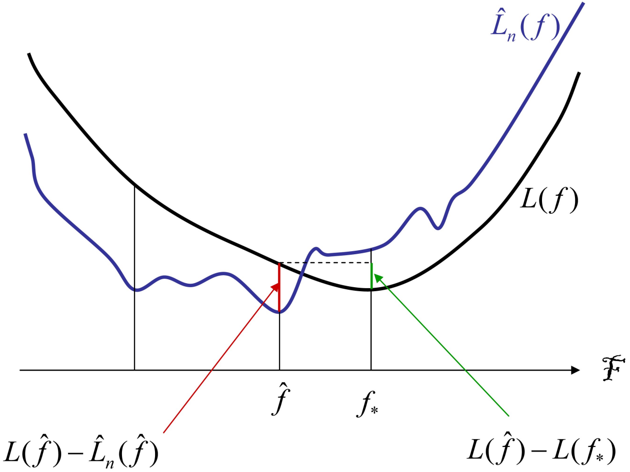
$$L(\widehat{f}) - \underbrace{\widehat{L}_n(\widehat{f})}_{\text{known}} = \underbrace{E[\ell(Y,\widehat{f}(X))|\mathcal{D}] - \widehat{E}_n[\ell(Y,\widehat{f}(X))]}_{?}.$$

- Approaches to analysis.
  - Asymptotic expansion of the expectation:

    e.g. $$E_{\mathcal{D}}\big[E[\ell(Y,\widehat{f}(X))] - \widehat{E}_n[\ell(Y,\widehat{f}(X))]\big] = \frac{A}{n} + \ldots$$

    $\Longrightarrow$ AIC.
  - Bounding risk:

    e.g. $$\Pr\big(E[\ell(Y,\widehat{f}(X))|\mathcal{D}] \le \widehat{E}_n[\ell(Y,\widehat{f}(X))] + \varepsilon\big)$$
    $$\le \Pr\Big(\sup_{f\in\mathcal{F}}\big(E[\ell(Y,f(X))] - \widehat{E}_n[\ell(Y,f(X))]\big) \le \varepsilon\Big) \quad \le \alpha e^{-\beta\varepsilon^2 n}.$$


$\hat{L}_n(f)$
$L(f)$
$\mathcal{F}$
$\hat{f}$
f_*
$L(\hat{f}) - \hat{L}_n(\hat{f})$
$L(\hat{f}) - L(f_*)$

Risk and empirical risk IV

- This lecture explains the latter approach
  - The bound applies for all n, not asymptotics.
  - Just a bound, but often derives a useful information in its functional form.
  - Can be applied to complex methods, such as SVM, AdaBoost.
  - Note: the loss function of SVM $(1 - yf(x))_+$ is not differentiable.

- The techniques explained here use the notion of Rademacher average [BBM02].
  For more classical background, see [Vap98].

- Comment on terminology:[1]
  - Risk = generalization error, prediction error, (expected log likelihood), etc.
  - Empirical risk = empirical error, training error, (empirical log likelihood), etc.

[1] The terminology in statistical learning theory is slightly different from statistics.

1. Bounding risk
   - Risk and empirical risk
   - Concentration inequalities
   - Bound for finite function class
2. Risk bound for infinite function class
   - Techniques for infinite function class
   - Rademacher average, growth function, and VC-dimension
3. Risk bound for SVM
   - Risk bound for SVM

Empirical mean and expectation

Before considering

$$\sup_{f\in\mathcal{F}} E[\ell(Y,f(X))] - \widehat{E}_n[\ell(Y,f(X))],$$

review the behavior of

$$E[\ell(Y,f(X))] - \widehat{E}_n[\ell(Y,f(X))] = E[Z] - \tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i.$$

- The law of large numbers (Z_i: i.i.d.)

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i \quad \longrightarrow \quad E[Z] \quad a.e.(n\to\infty)$$

- Central limit theorem (Z_i: i.i.d.)

$$\sqrt{n}\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i\big) - E[Z]\big) \Longrightarrow N(0, \mathrm{Var}[Z]) \quad (n\to\infty)$$

- How about

$$\Pr\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i - E[Z] \geq \varepsilon\big) \quad ?$$

Hoeffding's inequality

Theorem (Hoeffding's inequality)

$X_1, \ldots, X_n$: independent random variables, $X_i \in [a_i, b_i]$. Then, for any $\varepsilon > 0$,

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - E[X] > \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

and

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i - E[X] < -\varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- Proof is omitted (see e.g. [vdVW96]), since this is a corollary to McDiamid's inequality).
- Example:
  If $\ell(y, f) \in [0, 1]$, then for any $f \in \mathcal{F}$,

$$\Pr\left(|\widehat{L}_n(f) - L(f)| > \varepsilon\right) \leq 2e^{-2\varepsilon^2 n}.$$

Azuma-Hoeffding's/McDiamid's inequality

Theorem (Azuma-Hoeffding's/McDiamid's inequality)

$X_1, \ldots, X_n$: *independent random variables on* $\mathcal{X}$.
$f : \mathcal{X}^n \to \mathbb{R}$: *measurable function.*
Assume for each i *there exists* $c_i > 0$ *such that for any* $x_1, \ldots, x_n, x_i'$

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i,$$

then

$$\Pr\big(f(X_1, \ldots, X_n) - E[f(X_1, \ldots, X_n)] > \varepsilon\big) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\Pr\big(f(X_1, \ldots, X_n) - E[f(X_1, \ldots, X_n)] < -\varepsilon\big) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

Proof I

Remark. $f(x_1, \ldots, x_n) = \sum_{i=1}^n X_i$ and $c_i = b_i - a_i$ prove Hoeffding's inequality.

proof. Let

$$\begin{aligned} V_i &= E[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i] - E[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{i-1}] \\ &= E[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i] \\ &\qquad - E_{X_i}[E[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i] \mid X_1, \ldots, X_{i-1}] \end{aligned}$$

Then,

$$\sum_{i=1}^n V_i = f - E[f], \quad \text{and} \quad E[V_i \mid X_1, \ldots, X_{i-1}] = 0.$$

By Markov's inequality with e^{tx} $(t > 0)$,

$$\begin{aligned} &\Pr(f - E[f] > \varepsilon) = \Pr(\textstyle\sum_{i=1}^n V_i > \varepsilon) \\ &\le \inf_{t>0} e^{-t\varepsilon} E\left[e^{t \sum_{i=1}^n V_i}\right] \\ &= \inf_{t>0} e^{-t\varepsilon} E\left[E_{X_n}\left[e^{t \sum_{i=1}^n V_i} \mid X_1, \ldots, X_{n-1}\right]\right] \\ &= \inf_{t>0} e^{-t\varepsilon} E\left[e^{t \sum_{i=1}^{n-1} V_i} E_{X_n}\left[e^{tV_n} \mid X_1, \ldots, X_{n-1}\right]\right] \qquad [V_1, \ldots, V_{n-1} \perp\!\!\!\perp X_n]. \end{aligned}$$

Proof II

Let

$$L_i \equiv \inf_x V_i(x_1, \dots, x_{i-1}, x) \le V_i \le \sup_x V_i(x_1, \dots, x_{i-1}, x) \equiv U_i.$$

By the assumption, it is easy to see

$$U_i - L_i \le c_i.$$

From the lemma shown below, $E[e^{tV_n} \mid X_1, \dots, X_{n-1}] \le e^{t^2 c_n^2/8}$. Thus,

$$\Pr(f - E[f] > \varepsilon) \le \inf_{t>0} e^{-t\varepsilon} E\left[e^{t \sum_{i=1}^{n-1} V_i}\right] e^{-t^2 c_n^2/8}.$$

Repeating the same argument $n-1$ times,

$$\Pr(f - E[f] > \varepsilon) \le \inf_{t>0} e^{-t\varepsilon} e^{-t^2 \sum_{i=1}^{n} c_i^2/8}.$$

The optimal choice $t = 4\varepsilon / \sum_{i=1}^n c_i^2$ gives

$$\Pr\Big(f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)] > \varepsilon\Big) \le \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

The second inequality is obtained by replacing f with $-f$.

Lemmas

Lemma (Hoeffding's lemma)

Let X be a random variable with $E[X] = 0$ and $a \le X \le b$. Then for any $t > 0$,

$$E[e^{tX}] \le e^{t^2(b-a)^2/8}.$$

Proof omitted (exercise).

Lemma (Markov's inequality)

Let X be a random variable such that $X \ge 0$. Then, for any $\varepsilon > 0$

$$\Pr(X \ge \varepsilon) \le \frac{E[X]}{a}.$$

Bound for finite function class I

The simplest case: $|\mathcal{F}| < \infty$ (finite class). $\ell(y, f) \in [0, 1]$.

- For each $f \in \mathcal{F}$,

$$\Pr\Big(E[\ell(Y, f(X))] - \widehat{E}_n[\ell(Y, f(X))] \geq \varepsilon\Big) \leq e^{-2\varepsilon^2 n}.$$

- From $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$,

$$\Pr\Big(\sup_{f\in\mathcal{F}}\big\{E[\ell(Y, f(X))] - \widehat{E}_n[\ell(Y, f(X))]\big\} \geq \varepsilon\Big) \leq |\mathcal{F}|e^{-2\varepsilon^2 n}.$$

- Let $\delta = |\mathcal{F}|e^{-2\varepsilon^2 n}$.
  With probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}}\big\{E[\ell(Y, f(X))] - \widehat{E}_n[\ell(Y, f(X))]\big\} \leq \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}.$$

Bound for finite function class II

Two results:

- Estimation of the risk by the empirical risk.
  With probability at least $1-\delta$,

$$L(\widehat{f}) \leq \widehat{L}_n(\widehat{f}) + \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}.$$

- The difference from the optimal risk.
  $f_* = \arg\min_{f\in\mathcal{F}} L(f)$. With probability at least $1-2\delta$,

$$L(\widehat{f}) \leq L(f_*) + \sqrt{\frac{\log(1/\delta)}{2n}} + \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}.$$

  Proof.

$$\begin{aligned} L(\widehat{f}) &= (L(\widehat{f}) - \widehat{L}_n(\widehat{f})) + (\widehat{L}_n(\widehat{f}) - \widehat{L}_n(f_*)) + (\widehat{L}_n(f_*) - L(f_*)) + L(f_*) \\ &\leq (\text{uniform bound}) + (\leq 0) + (\text{Hoeffding}). \end{aligned}$$

Extension of risk bound to infinite classes

We wish to extend the uniform bound to an infinite function class $\mathcal{F}$;

$$\sup_{f \in \mathcal{F}} \left\{ E[\ell(Y, f(X))] - \widehat{E}_n[\ell(y, f(X))] \right\}.$$

Consider in general $\mathcal{G} \subset \{g : \mathcal{Z} \to [0,1]\}$ and

$$\sup_{g \in \mathcal{G}} \left\{ E[g(Z)] - \widehat{E}_n[g(Z)] \right\}.$$

Ex. $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{G} = \ell_{\mathcal{F}} = \{\ell(y, f(x)) \mid f \in \mathcal{F}\}$.

The method consists of three steps:

1. Concentration by Azuma-Hoeffding's inequality.
2. Symmetrization for removing $E[g]$.
3. Bounding Rademacher average.

Step 1: Concentration

- Define

$$h(z_1, \ldots, z_n) = \sup_{g \in \mathcal{G}} \left\{ E[g(Z)] - \tfrac{1}{n} \textstyle\sum_{i=1}^{n} g(z_i) \right\}.$$

- h satisfies the condition

$$|h(z_1, \ldots, z_{i-1}, z_i, \ldots, z_n) - h(z_1, \ldots, z_{i-1}, z_i', \ldots, z_n)| \leq 1/n.$$

- Apply Azuma-Hoeffding's inequality to h:
  With probability $\geq 1 - \delta$,

$$\sup_{g \in \mathcal{G}} \left\{ E[g(Z)] - \widehat{E}_n[g(Z)] \right\} \leq E\left[\sup_{g \in \mathcal{G}} \left\{ E[g(Z)] - \widehat{E}_n[g(Z)] \right\} \right] + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Step 2: Symmetrization - (1)

- We wish to have
$$E\Big[\sup_{g\in\mathcal{G}}\{E[g(Z)] - \widehat{E}_n[g(Z)]\}\Big]$$
converge to zero.
- Symmetrization.
$Z_1', \ldots, Z_n'$: an i.i.d. sample with the same distribution as Z_i.

$$E\Big[\sup_{g\in\mathcal{G}}\{E[g(Z)] - \widehat{E}_n[g(Z)]\}\Big] = E\Big[\sup_{g\in\mathcal{G}}\{E[\tfrac{1}{n}\textstyle\sum_{i=1}^n g(Z_i')] - \tfrac{1}{n}\sum_{i=1}^n g(Z_i)]$$

$$= E\Big[\sup_{g\in\mathcal{G}} E\big[\tfrac{1}{n}\textstyle\sum_{i=1}^n (g(Z_i') - g(Z_i)) \mid Z\big]\Big]$$

$$\leq E\Big[E\big[\sup_{g\in\mathcal{G}}\{\tfrac{1}{n}\textstyle\sum_{i=1}^n (g(Z_i') - g(Z_i))\} \mid Z\big]\Big] \qquad \text{[convexity of } \sup\text{]}$$

$$= E\Big[\sup_{g\in\mathcal{G}}\{\tfrac{1}{n}\textstyle\sum_{i=1}^n (g(Z_i') - g(Z_i))\}\Big]$$

- This removes the *infinite sample* $E[g]$, and makes a bound with a finite sample.

Step 2: Symmetrization - (2)

- We wish to remove the double sample Z_i and Z_i'.
- Rademacher variables: i.i.d. random variable $\sigma_i \in \{\pm 1\}$ with probability 1/2 for each value.
- Note: By the symmetry,

$$\sum_{i=1}^n (g(Z_i') - g(Z_i)) \quad \text{and} \quad \sum_{i=1}^n \sigma_i (g(Z_i') - g(Z_i))$$

have the same law.
Hence,

$$\begin{aligned} & E\left[\sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n (g(Z_i') - g(Z_i)) \right\}\right] \\ &= E\left[\sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n \sigma_i (g(Z_i') - g(Z_i)) \right\}\right] \\ &\leq E\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i')\right] + E\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i)\right] \\ &= 2E\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i)\right]. \end{aligned}$$

Step 3: Rademacher average

$$E\left[\sup_{g\in\mathcal{G}}\left\{E[g(Z)] - \widehat{E}_n[g(Z)]\right\}\right] \le 2E\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\right].$$

- Rademacher average:

$$R_n(\mathcal{G}) \equiv E\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\right].$$

- Empirical Rademacher average:

$$\widehat{R}_n(\mathcal{G}) \equiv E\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i) \mid Z_1,\ldots,Z_n\right].$$

- Note: $E[\sigma_i g(Z_i)] = 0$. Thus, $\frac{1}{n}\sum_i \sigma_i g(Z_i)$ must be small.
- $R_n(\mathcal{G})$ ($\widehat{R}_n(\mathcal{G})$) represents the complexity of the function class $\mathcal{G}$.

Example: $\mathcal{G} \subset \{g : \{Z_1,\ldots,Z_n\} \to \{\pm 1\}\}$. Regard σ_i as a label of Z_i.

$$\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i) = \frac{1}{n}\sum_{i=1}^n (1 - 2I_{\{\sigma_i \ne g(Z_i)\}}) = 1 - 2\widehat{L}_n(g).$$

$$R_n(\mathcal{G}) = 1 - 2 \times \text{(expected minimum empirical loss)}.$$

Risk bound for infinite classes

We have obtained: With probability $\geq 1-\delta$,

$$\sup_{g\in\mathcal{G}}\{E[g(Z)] - \widehat{E}_n[g(Z)]\} \leq 2R_n(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Two consequences:

$\ell(y,f)\in[0,1]$, and let $\ell_{\mathcal{F}} = \{\ell(y,f(x)) \mid f\in\mathcal{F}\}$.

- Estimation of the risk by the empirical risk.
  With probability at least $1-\delta$,

$$L(\widehat{f}) \leq \widehat{L}_n(\widehat{f}) + 2R_n(\ell_{\mathcal{F}}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

- The difference from the best possible risk.
  $f_* = \arg\min_{f\in\mathcal{F}} L(f)$. With probability at least $1-2\delta$,

$$L(\widehat{f}) \leq L(f_*) + 2R_n(\ell_{\mathcal{F}}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Relations between $R_n(\ell_{\mathcal{F}})$ and $R_n(\mathcal{F})$

The bound includes $R_n(\ell_{\mathcal{F}})$.
It is often related to $R_n(\mathcal{F})$, which is easier to analyze.

- 0-1 loss: $\mathcal{F} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$, $\ell(y, f) = \frac{1-yf}{2}$.
- Fact: for 0-1 loss,

$$R_n(\ell_{\mathcal{F}}) = \frac{1}{2} R_n(\mathcal{F}).$$

Proof.

$$\begin{aligned} R_n(\ell_{\mathcal{F}}) &= E\left[\sup_{f \in \mathcal{F}} \textstyle\sum_{i=1}^n \sigma_i \dfrac{1 - Y_i f(X_i)}{2}\right] \\ &= \frac{1}{2} E\left[\sup_{f \in \mathcal{F}} \textstyle\sum_{i=1}^n (-\sigma_i Y_i) f(X_i)\right] \\ &= \frac{1}{2} R_n(\mathcal{F}) \qquad [(-\sigma_i Y_i) \text{ works as a Rademacher variable}] \end{aligned}$$

Bounding Rademacher average I

How to bound the Rademacher average ?

$$R_n(\mathcal{G}) = E\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^n \sigma_i g(Z_i)\right]$$

Assume $\mathcal{G} \subset \{g : \mathcal{Z} \to \{\pm 1\}\}$.

Note: $\mathcal{G}$ affects on $R_n(\mathcal{G})$ only through $(g(Z_1), \ldots, g(Z_n)) \in \{\pm 1\}^n$.

We can use the following lemma.

Lemma (Massart [Mas])

A: finite subset of $\mathbb{R}^n$. Assume $\max_{a\in A} \|a\| \le R$.
Then

$$E\left[\max_{a\in A} \sum_{i=1} \sigma_i a_i\right] \le R\sqrt{2\log|A|},$$

where σ_i are Rademacher variables.

Bounding Rademacher average II

For $Z_1^n = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, define

$$\mathcal{G}_{|Z_1^n} = \big\{ \big(g(Z_1), \ldots, g(Z_n)\big) \in \{\pm 1\}^n \mid g \in \mathcal{G} \big\}.$$

Fact:

$$R_n(\mathcal{G}) \le \sqrt{\frac{2E[\log |\mathcal{G}_{|Z_1^n}|]}{n}} \le \sqrt{\frac{2 \log E[|\mathcal{G}_{|Z_1^n}|]}{n}}.$$

Proof.

$$\begin{aligned}
R_n(\mathcal{G}) &= E\big[\sup_{a \in \mathcal{G}_{|Z_1^n}} \tfrac{1}{n} \textstyle\sum_{i=1}^n \sigma_i a_i\big] \\
&= E\Big[E\big[\sup_{a \in \mathcal{G}_{|Z_1^n}} \tfrac{1}{n} \textstyle\sum_{i=1}^n \sigma_i a_i \mid Z_1^n\big]\Big] \\
&\le \frac{1}{\sqrt{n}} E\Big[\sqrt{2 \log |\mathcal{G}_{|Z_1^n}|}\Big] && \text{[Massart's lemma]} \\
&\le \sqrt{\frac{2E[\log |\mathcal{G}_{|Z_1^n}|]}{n}} && \text{[concavity of } \sqrt{\ }\text{]} \\
&\le \sqrt{\frac{2 \log E[|\mathcal{G}_{|Z_1^n}|]}{n}} && \text{[concavity of } \log\text{]}
\end{aligned}$$

Proof of Massart's lemma

Proof.

Let $s > 0$.

$$
\begin{aligned}
\exp(sE[\max_a \textstyle\sum_i \sigma_i a_i]) &\le E[\exp(s\max_a \textstyle\sum_i \sigma_i a_i)] \qquad \text{[convexity of } \exp(sz)\text{]} \\
&= E[\max_a \exp(s\textstyle\sum_i \sigma_i a_i)] \\
&\le E[\textstyle\sum_a \exp(s\textstyle\sum_i \sigma_i a_i)] \qquad [\max \longrightarrow \textstyle\sum] \\
&= \textstyle\sum_a E\left[\prod_{i=1}^n e^{s\sigma_i a_i}\right] \qquad \text{[independence of } \sigma_i\text{]} \\
&= \textstyle\sum_{a\in A} \prod_{i=1}^n E\left[e^{s\sigma_i a_i}\right] \\
&\le \textstyle\sum_{a\in A} \prod_{i=1}^n \exp(s^2 4a_i^2/8) \\
&\qquad\qquad \text{[Hoeffding's lemma, } \sigma_i a_i \in [-a_i, a_i]\text{]} \\
&= |A| \exp(s^2 R^2/2).
\end{aligned}
$$

Take the optimal $s = \sqrt{\frac{2\log|A|}{R^2}}$. Then,

$$
E[\max_a \textstyle\sum_i \sigma_i a_i] \le R\sqrt{2\log|A|}.
$$

Distribution-free bound: Growth function

Let $\mathcal{G} \subset \{g : \mathcal{Z} \to \{\pm 1\}\}$.

Definition. Growth function

$$\Pi_{\mathcal{G}}(n) = \max\{|\mathcal{G}_{|Z_1^n}| \in \mathbb{N} \mid Z_1^n = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n\}.$$

$\Pi_{\mathcal{G}}(n)$ is monotonically decreasing w.r.t. n.

Definition. Vapnik-Chervonenkis (VC) dimension

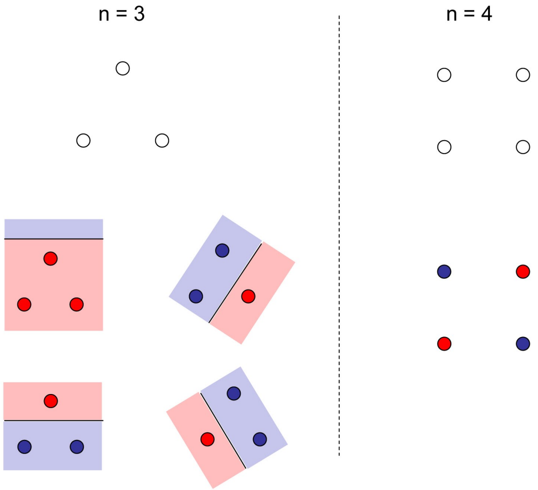
$$\text{dim}_{VC}(\mathcal{G}) = \max\{n \in \mathbb{N} \mid \Pi_{\mathcal{G}}(n) = 2^n\}$$

Example: linear threshold functions on $\mathbb{R}^d$.

$$\mathcal{G} = \{sgn(w^T x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\},$$

$$\text{dim}_{VC}(\mathcal{G}) = d + 1.$$

d = 2



Sauer's lemma

Theorem (Sauer's lemma)

$\mathcal{G} \subset \{g : \mathcal{Z} \to \{\pm 1\}\}$. dim$_{VC}(\mathcal{G}) = d$. Then,

$$\Pi_{\mathcal{G}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

and for $n \geq d$,

$$\Pi_{\mathcal{G}}(n) \leq \left(\frac{en}{d}\right)^{d}.$$

Corollary (Distribution-free bound of Rademacher average)

$\mathcal{G} \subset \{g : \mathcal{Z} \to \{\pm 1\}\}$. dim$_{VC}(\mathcal{G}) = d$. Then,

$$R_n(\mathcal{G}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{G}}(n)}{n}} \leq \sqrt{\frac{2d(\log n + \log(e/d))}{n}}.$$

For the proof of Sauer's lemma, see [Vap98].

Bound of risk I

$\mathcal{F} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$. $\dim_{VC}(\mathcal{F}) = d$.
Recall $R_n(\ell_{\mathcal{F}}) = \frac{1}{2} R_n(\mathcal{F}) \leq \frac{1}{2}\sqrt{\frac{2d\log n + 2d\log(e/d))}{n}}$ for $n \geq d$.

Distribution-free bound of risk.

- Estimation of the risk by the empirical risk.
  With probability at least $1-\delta$,

$$L(\widehat{f}) \leq \widehat{L}_n(\widehat{f}) + \sqrt{\frac{2d\log n + 2d\log(e/d))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

- The difference from the best possible risk.
  $f_* = \arg\min_{f\in\mathcal{F}} L(f)$. With probability at least $1-2\delta$,

$$L(\widehat{f}) \leq L(f_*) + \sqrt{\frac{2d\log n + 2d\log(e/d))}{n}} + 2\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Bound of risk II

- Risk bound:
  With Probability $\geq 1-\delta$,

$$L(\widehat{f}) \leq \widehat{L}_n(\widehat{f}) + \sqrt{\frac{2d\log n}{n} + \frac{2d\log(e/d))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- AIC:

$$E_{\mathcal{D}}[L(\widehat{f})] \approx E_{\mathcal{D}}[L(\widehat{f})] + \frac{\text{\# parameters}}{n}.$$

- MDL:

$$\text{MDL} = E_{\mathcal{D}}[L(\widehat{f})] + \frac{\text{\# parameters}\log n}{n}.$$

Properties of Rademacher average I

1. $$\mathcal{F} \subset \mathcal{G} \quad \Longrightarrow \quad R_n(\mathcal{F}) \subset R_n(\mathcal{G}).$$

2. $$R_n(c\mathcal{F}) = |c| R_n(\mathcal{F}),$$
   where $c \in \mathbb{R}$ and $c\mathcal{F} = \{cf \mid f \in \mathcal{F}\}$.

3. For $F + g = \{f + g \mid f \in \mathcal{F}\}$,
   $$R_n(\mathcal{F} + g) = R_n(\mathcal{F}).$$

4. Assume $-\mathcal{F} = \mathcal{F}$. Then,
   $$R_n(\text{co}\mathcal{F}) = R_n(\mathcal{F}),$$
   where co$\mathcal{F} = \{\sum_{i=1}^m a_i f_i \mid f_i \in \mathcal{F}, a_i \geq 0, \sum_{i=1}^n a_i = 1\}$.

Properties of Rademacher average II

5. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ $(i = 1, \ldots, n)$ be Lipschitz continuous with Lipschitz constant b, i.e.,

$$|\phi_i(x) - \phi_i(y)| \le b|x - y| \qquad (\forall x, y).$$

Then,

$$E\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \phi_i(f(X_i))\Big] \le b\, E\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\Big] = bR_n(\mathcal{F}),$$

where σ_i are Rademacher constants.

Proof is omitted. For (5), see [LT91], Th.4.12.

Rademacher average vs distribution-free bound

How to measure the complexity of function classes.

- VC-dimension is simple and easy to compute or bound for many function classes.
- VC-dimension does not take the distribution of X into account.

- Rademacher average includes the distribution of X.
- It may not be easy to compute.
- Various useful properties. (For Rademacher averages, see [BM02], [LT91].)

Mini-summary on risk bound

- With probability $\geq 1 - \delta$,

$$\text{(Risk)} \leq \text{(Empirical risk)} + \text{(Complexity of } \mathcal{F}\text{)} + \Theta(1/\delta).$$

- The bound applies to all n, but usually meaningful for large n.
- The functional form of the complexity term reflects the property of the function class and learning method.
- Rademacher average represents the complexity term. It is upper bounded by using VC dimension.

Review of risk bound I

- Assume loss function $\ell(y, f) \in [0, 1]$, and let $\ell_{\mathcal{F}} = \{\ell(y, f(x)) \mid f \in \mathcal{F}\}$.
- Risk: the purpose of learning is to minimize the risk;

$$L(f) = E[\ell(Y, f(X))] \qquad (f \in \mathcal{F}).$$

- Empirical risk:

$$\widehat{L}_n(f) = \widehat{E}_n[\ell(Y, f(X))] = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) \qquad (f \in \mathcal{F}).$$

- Learning:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} L_n(f).$$

Review of risk bound II

- Estimation of the risk by the empirical risk.
  With probability at least $1-\delta$,

$$L(\widehat{f}) \leq \widehat{L}_n(\widehat{f}) + 2R_n(\ell_{\mathcal{F}}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

- The difference from the best possible risk.
  $f_* = \arg\min_{f\in\mathcal{F}} L(f)$. With probability at least $1-2\delta$,

$$L(\widehat{f}) \leq L(f_*) + 2R_n(\ell_{\mathcal{F}}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}.$$

- Rademacher average $R_n(\mathcal{G})$ expresses the complexity of $\mathcal{G}$.

$$R_n(\mathcal{G}) = E\left[\sup_{g\in\mathcal{G}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i g(Z_i)\right],$$

  where $\sigma_i \in \{\pm 1\}$ are Rademacher variables (i.i.d. and $\Pr(\sigma_i = 1) = 1/2$.

Hinge loss and 0-1 loss I

Binary classification. $y \in \{\pm 1\}$.

- 0-1 loss:

$$\ell_{01}(y, f) = (1 - y\,\mathrm{sgn}(f))/2.$$

- Risk is often evaluated with 0-1 loss in classification.

$$L(f) = E[\ell_{01}(y, f(X))] = E[Y \neq \mathrm{sgn}(f(X))].$$

- Hinge loss (soft margin loss)

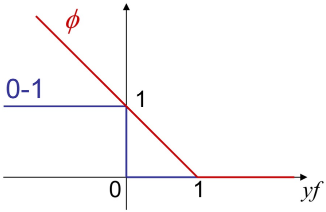
$$\ell_{hinge}(y, f) = \phi(fy), \quad \phi(t) = (1 - t)_+$$

  used for representing the constraints of soft-margin SVM.
- *c.f.* SVM

$$\min \widehat{E}_n[\phi(Y_i f(X_i))] + \frac{\lambda}{2}\|f\|^2.$$

Hinge loss and 0-1 loss II

- Truncated hinge loss:
$$\tilde{\phi}(t) = \min(1, \phi(t)).$$
- $\tilde{\phi}$ satisfies $\tilde{\phi}(yf) \in [0, 1]$.
The results on the uniform bound are applicable.
- Relation:
$$\ell_{01}(y, f(x)) \leq \tilde{\phi}(yf(x)) \leq \phi(yf(x)).$$



Uniform bound with hinge loss

- $$L(f) = E[\ell_{01}(Y, f(X))] \leq E[\tilde{\phi}(Yf(X))].$$

- With probability $\geq 1 - \delta$,

  $$\sup_{f \in \mathcal{F}} \{E[\tilde{\phi}(Yf(X))] - \widehat{E}_n[\tilde{\phi}(Yf(X))]\} \leq 2R_n(\ell_{\tilde{\phi},\mathcal{F}}) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

  where $\ell_{\tilde{\phi},\mathcal{F}} = \{\tilde{\phi}(yf(x)) \mid f \in \mathcal{F}\}$.

- As a result, With probability $\geq 1 - \delta$,

  $$L(f) \leq \underbrace{\widehat{E}_n[\phi(Yf(X))]}_{\text{empirical hinge loss}} + 2R_n(\ell_{\tilde{\phi},\mathcal{F}}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

  for any $f \in \mathcal{F}$.

Uniform bound for SVM

- Recall margin = $1/\|w\|$ (w: weight of linear classifier).
- Set the function class

$$\mathcal{F}_r = \{f \in \mathcal{H}_k \mid \|f\|_{\mathcal{H}_k} \leq r\}$$

and consider

$$\min_{f \in \mathcal{H}_k} \widehat{E}_n[\phi(Yf(X))] \qquad \text{subj. to } f \in \mathcal{F}_r.$$

(Slightly different from the original SVM.)

Lemma

$$R_n(\ell_{\tilde{\phi},\mathcal{F}_r}) \leq R_n(\mathcal{F}_r) \leq r\sqrt{\frac{E[k(X,X)]}{n}}.$$

Risk bound for SVM

Theorem

Let $\mathcal{F}_r = \{f \in \mathcal{H}_k \mid \|f\|_{\mathcal{H}_k} \le r\}$.
With probability $\ge 1 - \delta$,

$$L(f) \le \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+ + 2r\sqrt{\frac{E[k(X,X)]}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

for any $f \in \mathcal{F}_r$.

- The risk is smaller for a class of larger margin (smaller r), assuming that the empirical error is the same.
- The complexity term of the function class does not depend on the dimensionality ($\approx$ number of parameters), but only on the norm.

Proof of Lemma I

1. $R_n(\ell_{\tilde{\phi},\mathcal{F}_r}) \leq R_n(\mathcal{F}_r)$.

By definition,

$$R_n(\ell_{\tilde{\phi},\mathcal{F}_r}) = E\left[\sup_{f\in\mathcal{F}_r} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\tilde{\phi}(Y_i f(X_i))\right].$$

Since $\tilde{\phi}$ is Lipschitz continuous

$$|\tilde{\phi}(t_1) - \tilde{\phi}(t_2)| \leq |t_1 - t_2|,$$

(see Properties of Rademacher averages (5))

$$E\left[\sup_{f\in\mathcal{F}_r} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\tilde{\phi}(Y_i f(X_i))\right] \leq E\left[\sup_{f\in\mathcal{F}_r} \frac{1}{n}\sum_{i=1}^{n}\sigma_i Y_i f(X_i)\right] = R_n(\mathcal{F}_r).$$

The last equality holds because $\sigma_i Y_i$ are Rademacher variables.

Proof of Lemma II

2. $R_n(\mathcal{F}_r) \le r\sqrt{E[k(X,X)]/n}$.

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i f(X_i) = \left\langle \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i k(\cdot, X_i), f \right\rangle \le \|f\| \left\| \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i k(\cdot, X_i)\right\|.$$

Thus,

$$R_n(\mathcal{F}_r) \le rE\left\| \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i k(\cdot, X_i)\right\|.$$

$$\begin{aligned}
&\left(E\left\| \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i k(\cdot, X_i)\right\|\right)^2 \\
&\le E\left\| \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i k(\cdot, X_i)\right\|^2 && [E|\varphi| \le (E|\varphi|^2)^{1/2}] \\
&= E\left[\tfrac{1}{n^2}\textstyle\sum_{i,j=1}^n \sigma_i\sigma_j k(X_i, X_j)\right] \\
&= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n E\left[k(X_i, X_i)\right] + \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \sum_{j\ne i} E[\sigma_i]E[\sigma_j]E[k(X_i, X_j)] \\
&= \tfrac{1}{n}E[k(X,X)] + 0.
\end{aligned}$$

□

More on the bound for SVM etc.

- The previous theorem does not reflect the learning of SVM rigorously;
  the bound is determined as a result of learning, not a priori.

- More rigorous approaches:
  - Bound by fat shattering dimension [BST99].
  - Luckiness framework [Her01].

- Other topics:
  - Generalization of boosting.
  - Relation to the uniform convergence of empirical process (covering number, entropy integral, etc.).

References I

P. Bartlett, O. Bousquet, and S. Mendelson.
Localized rademacher complexities.
In *Proceedings of the 15th annual conference on Computational Learning Theory*, pages 44–58, 2002.

Peter L. Bartlett and Shahar Mendelson.
Rademacher and gaussian complexities: Risk bounds and structural results.
Jounral of Machine Learning Research, 3:463–482, 2002.

Peter Bartlett and John Shawe-Taylor.
Generalization performance of support vector machines and other pattern classifiers.
pages 43–54, 1999.

Ralf Herbrich.
Learning Kernel Classifiers: Theory and Algorithms.
Cambridge, MA, USA, 2001.

References II

Michel Ledoux and Michel Talagrand.
Probability in Banach Spaces.
Springer-Verlag, 1991.

Pascal Massart.
Some applications of concentration inequalities to statistics.
Annales de la faculté des sciences de Toulouse Sér. 6, 9(2):245–303.

Vladimir N. Vapnik.
Statistical Learning Theory.
Wiley-Interscience, 1998.

Ard van der Vaart and Jon A. Wellner.
Weak convergence and empirical processes.
Springer Verlag, 1996.