ELE530: Introduction to Statistical Learning Theory

Fernando Pérez-Cruz

Princeton University and University Carlos III in Madrid

March 2nd, 2009

Statistical Learning Theory

- Minimum risk:

$$R(\delta) = \sum_y \int L(\delta(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x}.$$

- $p(\mathbf{x}, y)$ is unknown, we are given $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ from it.
- Empirical risk:

$$R_{emp}(\delta) = \frac{1}{n} \sum_{i=1}^{n} L(\delta(\mathbf{x}_i), y_i)$$

- For a given δ the $R_{emp}(\delta)$ converges to $R(\delta)$ by the law of large number. But for $\delta^n = \arg\min R_{emp}(\delta)$ it does not need to.
- Example: $\delta(\mathbf{x}_i) = \begin{cases} y_i, & \mathbf{x} = \mathbf{x}_i \\ 1, & \text{otherwise} \end{cases}$.

Hoeffding Bound

- We define $\xi_i = |\delta(\mathbf{x}_i) - y_i|$, it measures the error in each $\mathbf{x}_i$.

$$\xi_i \in \{0, 1\}$$

- For independent samples, $\xi_1, \ldots, \xi_n$ are independent Bernoulli trials.
- Theorem: Let $\{\mathbf{x}_i\}_{i=1}^n$ be n independent samples from a bounded random variable ξ, support in $[a, b]$. Then for any ϵ:

$$P\left[\frac{1}{n}\sum_i \xi_i - E[\xi] \geq \epsilon\right] \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

$$P\left[E[\xi] - \frac{1}{n}\sum_i \xi_i \geq \epsilon\right] \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

Consistency

- For any δ and applying the Hoeffding bound:

$$P\left[|R_{emp}(\delta) - R(\delta)| \geq \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right)$$

- The empirical risk converges exponentially fast to the true risk.
- So why did we say we needed some else?

Consistency

- For any δ and applying the Hoeffding bound:

$$P\left[|R_{emp}(\delta) - R(\delta)| \geq \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right)$$

- The empirical risk converges exponentially fast to the true risk.
- So why did we say we needed some else?
- We do not want any δ to converge to the true risk.
- We want $\delta^n = \arg\min R_{emp}(\delta)$ to converge to the true risk.

Consistency

- For any δ and applying the Hoeffding bound:

$$P\left[|R_{emp}(\delta) - R(\delta)| \geq \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right)$$

- The empirical risk converges exponentially fast to the true risk.
- So why did we say we needed some else?
- We do not want any δ to converge to the true risk.
- We want $\delta^n = \arg\min R_{emp}(\delta)$ to converge to the true risk.
- We have selected δ^n, to minimize the mean of ξ_i.

Consistency

- For any δ and applying the Hoeffding bound:

$$P\left[|R_{emp}(\delta) - R(\delta)| \geq \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right)$$

- The empirical risk converges exponentially fast to the true risk.
- So why did we say we needed some else?
- We do not want any δ to converge to the true risk.
- We want $\delta^n = \arg\min R_{emp}(\delta)$ to converge to the true risk.
- We have selected δ^n, to minimize the mean of ξ_i.
- Therefore they are no longer independent Bernoulli trials.

Consistency

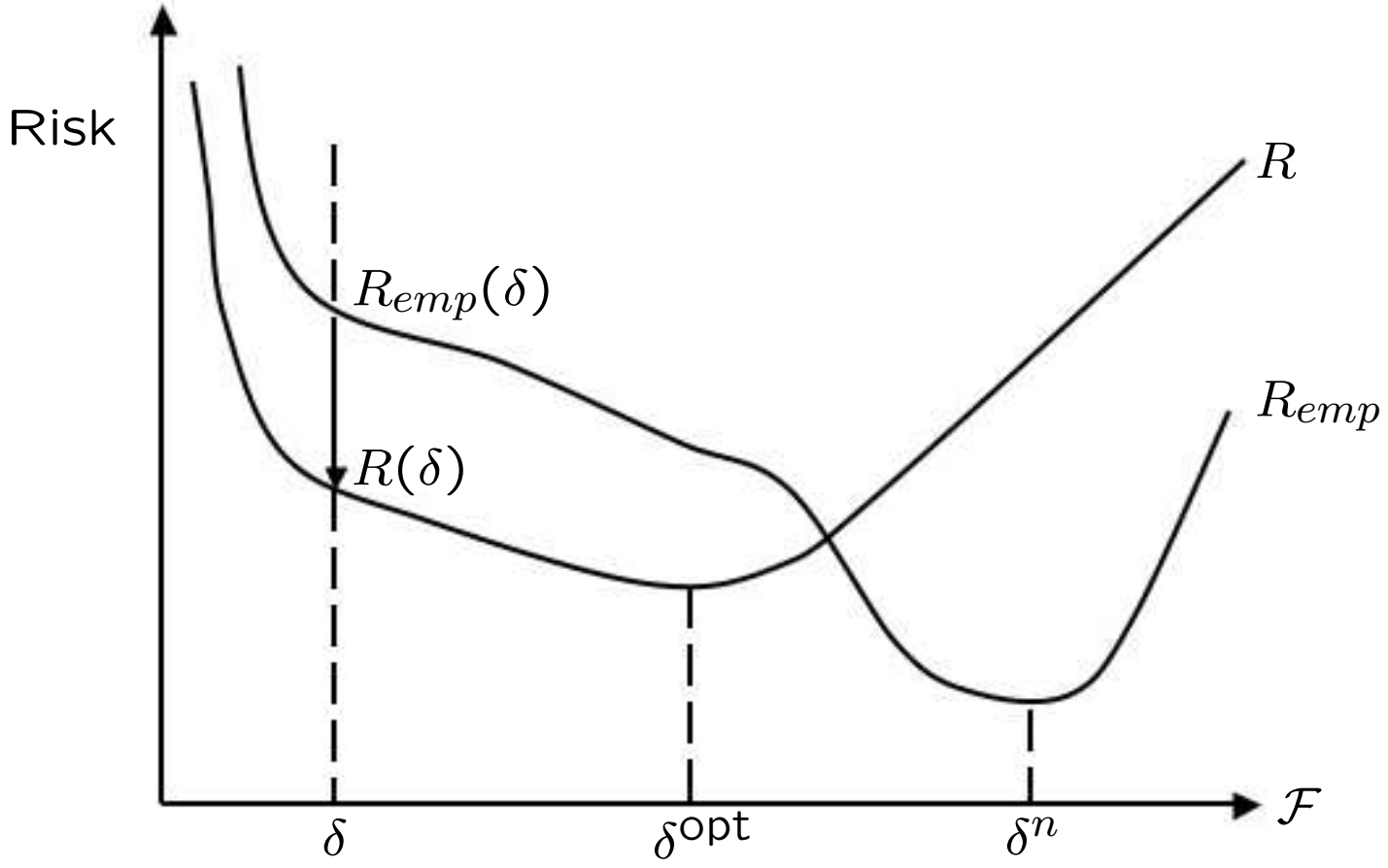
- We want that as n tends to infinity that

$$R_{emp}(\delta^n) \quad \longrightarrow \quad R(\delta^{\text{opt}})$$

  where $R(\delta^{\text{opt}}) = \arg\min R(\delta)$.

- As n grows we want to do as good as we we did know $p(\mathbf{x}, y)$.
- This cannot be done if we do not restrict the class of functions $\mathcal{F}$ from which we draw δ.
- We will see that we need a *worse case analysis*.
- That all functions in $\mathcal{F}$ converge to the true risk.
- We can only do that if we restrict the class of functions $\mathcal{F}$.

Consistency



Consistency

- For any $\delta \in \mathcal{F}$:

$$R(\delta) - R(\delta^{\text{opt}}) \geq 0$$

- For any $\delta \in \mathcal{F}$:

$$R_{emp}(\delta) - R_{emp}(\delta^{n}) \geq 0$$

Consistency

- For δ^n:

$$R(\delta^n) - R(\delta^{\text{opt}}) \geq 0$$

- For δ^{opt}:

$$R_{emp}(\delta^{\text{opt}}) - R_{emp}(\delta^n) \geq 0$$

Consistency

- For any δ^n:

$$R(\delta^n) - R(\delta^{\text{opt}}) \geq 0$$

- For any δ^{opt}:

$$R_{emp}(\delta^{\text{opt}}) - R_{emp}(\delta^n) \geq 0$$

- Therefore:

$$\begin{aligned} 0 &\leq R(\delta^n) - R(\delta^{\text{opt}}) + R_{emp}(\delta^{\text{opt}}) - R_{emp}(\delta^n) \\ &= R(\delta^n) - R_{emp}(\delta^n) + R_{emp}(\delta^{\text{opt}}) - R(\delta^{\text{opt}}) \\ &\leq \sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) + R_{emp}(\delta^{\text{opt}}) - R(\delta^{\text{opt}}) \end{aligned}$$

Consistency

- δ^{opt} does not depend on the n training samples, and

$$|R_{emp}(\delta^{\text{opt}}) - R(\delta^{\text{opt}})| \quad \xrightarrow{P} \quad 0 \qquad \text{as} \quad n \to \infty$$

  by the law of large numbers.

- If $\mathcal{F}$ is such that:

$$\sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

  the minimum of the empirical risk converges to the true risk.

- One-sided uniform convergence is a sufficient condition for learning consistency.

Consistency

- δ^{opt} does not depend on the n training samples, and

$$|R_{emp}(\delta^{\text{opt}}) - R(\delta^{\text{opt}})| \quad \xrightarrow{P} \quad 0 \qquad \text{as} \quad n \to \infty$$

  by the law of large numbers.

- If $\mathcal{F}$ is such that:

$$\sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

  the minimum of the empirical risk converges to the true risk.

- One-sided uniform convergence is a sufficient condition for learning consistency.

- It is also necessary (but we will not get to that).

Convergence Bounds

- For a given $\mathcal{F}$ and training set, how do we know if we achieve one-sided convergence?

$$\sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

- Can we get a simple bound that tell us when this converge occurs?

Convergence Bounds

- For a given $\mathcal{F}$ and training set, how do we know if we achieve one-sided convergence?

$$\sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

- Can we get a simple bound that tell us when this converge occurs?
- Are these bounds problem dependent?

Convergence Bounds

- For a given $\mathcal{F}$ and training set, how do we know if we achieve one-sided convergence?

$$\sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

- Can we get a simple bound that tell us when this converge occurs?
- Are these bounds problem dependent?
- How fast is that convergence?

Convergence Bounds

- For a given $\mathcal{F}$ and training set, how do we know if we achieve one-sided convergence?

$$\sup_{\delta \in \mathcal{F}} (R(\delta^n) - R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

- Can we get a simple bound that tell us when this converge occurs?
- Are these bounds problem dependent?
- How fast is that convergence?
- Vapnik and Chervonenkis addressed these issues and gave us the main results in learning theory (VC Theory).

Symmetrization

- Lemma: For $m\epsilon^2 \geq 2$

$$P\left(\sup_{\delta\in\mathcal{F}}(R(\delta) - R_{emp}(\delta)) > \epsilon\right) \leq 2P\left(\sup_{\delta\in\mathcal{F}}(R'_{emp}(\delta) - R_{emp}(\delta)) > \epsilon/2\right)$$

  where the empirical risk on the right hand side is computed for two independent sets of n training data.

- The consequence of this lemma is that we can bound the supremum over infinite many functions, by a supremum over a finite number of functions.

- Why?

Shattering Coefficient

- We are given 2 sets with n samples:

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \qquad \{(\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{2n}, y_{2n})\}$$

- We are given and infinite function class $\mathcal{F}$.
- But over these set we only have at most 2^{2n} different function.
- Why?
- Let see a simple example with $n = 2$.

Shattering Coefficient

- We are given 2 sets with n samples:

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \qquad \{(\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{2n}, y_{2n})\}$$

- We are given and infinite function class $\mathcal{F}$.
- But over these set we only have at most 2^{2n} different function.
- Why?
- Let see a simple example with $n = 2$.
- Definition: The shattering coefficient $\mathcal{N}(\mathcal{F}, n)$ is the number of ways that the function class $\mathcal{F}$ can separate the n patterns in two ways.

Uniform Convergence Bound

- We first pick a maximal set of functions

$$\left\{\delta_1, \ldots, \delta_{\mathcal{N}(\mathcal{F}, Z_{2n})}\right\}$$

where $Z_{2n} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{2n}, y_{2n})\}$ (Specific set of samples from $p(\mathbf{x}, y)$).

- Like the δ_j are Z_{2n} dependent, we are going to apply a random permutation σ to Z_{2n}. We will consider instead:

$$P\left(\sup_{\delta\in\mathcal{F}}(R'^{\sigma}_{emp}(\delta) - R^{\sigma}_{emp}(\delta)) > \epsilon/2\right)$$
$$= \int P\left(\sup_{\delta\in\mathcal{F}}(R'^{\sigma}_{emp}(\delta) - R^{\sigma}_{emp}(\delta) > \epsilon/2 | Z_{2n}\right) dP(Z_{2n})$$

Erroneous Events

- We look for the permutations σ such that

$$C_\epsilon = \left\{ \sigma \middle| \sup_{\delta \in \left\{\delta_1,\ldots,\delta_{\mathcal{N}(\mathcal{F},Z_{2n})}\right\}} (R_{emp}'^{\sigma}(\delta) - R_{emp}^{\sigma}(\delta)) > \epsilon/2 \right\}$$

Which can be expressed as:

$$C_\epsilon = \bigcup_{j=1}^{\mathcal{N}(\mathcal{F},Z_{2n})} C_\epsilon(\delta_j)$$

where

$$C_\epsilon(\delta_j) = \left\{ \sigma \middle| (R_{emp}'^{\sigma}(\delta_j) - R_{emp}^{\sigma}(\delta_j)) > \epsilon/2 \right\}$$

Union Convergence Bounds

- For $Z_{2n} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{2n}, y_{2n})\}$, the functions δ_j are fixed.
- The law of large number holds for each δ_i.
- For $\xi_i^\sigma = |\delta(\mathbf{x}_i^\sigma) - y_i^\sigma|$, we can apply the Hoeffding bound.

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - \frac{1}{n}\sum_{i=n+1}^{2n}\xi_i \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

- Therefore:

$$P\left(C_\epsilon(\delta_i)\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8}\right).$$

Union Convergence Bounds

$$P\left(\sup_{\delta\in\mathcal{F}}(R'^{\sigma}_{emp}(\delta)-R^{\sigma}_{emp}(\delta))>\epsilon/2\right)$$

$$=\int P\left(\sup_{\delta\in\mathcal{F}}(R'^{\sigma}_{emp}(\delta)-R^{\sigma}_{emp}(\delta)>\epsilon/2|Z_{2n}\right)dP(Z_{2n})$$

$$=\int P\left(C_\epsilon\right)dP(Z_{2n})=\int P\left(\bigcup_{j=1}^{\mathcal{N}(\mathcal{F},Z_{2n})}C_\epsilon(\delta_j)\right)dP(Z_{2n})$$

$$\leq\int\sum_{j=1}^{\mathcal{N}(\mathcal{F},Z_{2n})}P\left(C_\epsilon(\delta_j)\right)dP(Z_{2n})$$

$$\leq\int\mathcal{N}(\mathcal{F},Z_{2n})2\exp\left(-\frac{n\epsilon^2}{8}\right)dP(Z_{2n})$$

$$=2E[\mathcal{N}(\mathcal{F},Z_{2n})]\exp\left(-\frac{n\epsilon^2}{8}\right)$$

Union Convergence Bounds

- Finally we can bound the one-sided uniform convergence:

$$P\left(\sup_{\delta\in\mathcal{F}}(R(\delta)-R_{emp}(\delta))>\epsilon\right)\leq 4\exp\left(\log E[\mathcal{N}(\mathcal{F},Z_{2n})]-\frac{n\epsilon^2}{8}\right)$$

- If $E[\mathcal{N}(\mathcal{F},Z_{2n})]$ does not grows exponentially with n, then

$$\sup_{\delta\in\mathcal{F}}(R(\delta^n)-R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

- Does this makes sense?

Union Convergence Bounds

- Finally we can bound the one-sided uniform convergence:

$$P\left(\sup_{\delta\in\mathcal{F}}(R(\delta)-R_{emp}(\delta))>\epsilon\right)\leq 4\exp\left(\log E[\mathcal{N}(\mathcal{F},Z_{2n})]-\frac{n\epsilon^2}{8}\right)$$

- If $E[\mathcal{N}(\mathcal{F},Z_{2n})]$ does not grows exponentially with n, then

$$\sup_{\delta\in\mathcal{F}}(R(\delta^n)-R_{emp}(\delta^n)) \quad \xrightarrow{P} \quad 0$$

- Does this makes sense?
  - If $E[\mathcal{N}(\mathcal{F},Z_{2n})]=2^{2n}$, it means that we have a function $\delta_j\in\mathcal{F}$ that gives zero error on every Z_{2n}.
  - We can explain any data. No generalization capabilities in $\mathcal{F}$.

Confidence Interval

- We set the excess probability to ρ:

$$P\left(\sup_{\delta\in\mathcal{F}}(R(\delta)-R_{emp}(\delta))>\epsilon\right)\leq 4\exp\left(\log E[\mathcal{N}(\mathcal{F},Z_{2n})]-\frac{n\epsilon^2}{8}\right)=\rho$$

- Then:

$$\epsilon=\sqrt{\frac{8}{m}\left(\log E[\mathcal{N}(\mathcal{F},Z_{2n})]-\frac{4}{\rho}\right)}$$

- Therefore with probability $1-\rho$

$$R(\delta)\leq R_{emp}(\delta)+\sqrt{\frac{8}{m}\left(\log E[\mathcal{N}(\mathcal{F},Z_{2n})]-\frac{4}{\rho}\right)}$$

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

Annealed VC Entropy

- For $Z_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
- Vapnik and Chervonenkis defined the annealed entropy as:

$$H_{\mathcal{F}}^{\text{ann}}(n) = \log E[\mathcal{N}(\mathcal{F}, Z_n)]$$

- For uniform two-sided exponential convergence:

$$P\left(\sup_{\delta \in \mathcal{F}} |R(\delta) - R_{emp}(\delta)| > \epsilon\right) \leq 4 \exp\left((H_{\mathcal{F}}^{\text{ann}}(2n)/n - \epsilon^2)m\right)$$

- It is necessary and sufficient that

$$\lim_{n \to \infty} H_{\mathcal{F}}^{\text{ann}}(n)/n = 0$$

VC Entropy

- Vapnik and Chervonenkis defined entropy as:

$$H_{\mathcal{F}}(n) = E[\log \mathcal{N}(\mathcal{F}, Z_n)] \leq H_{\mathcal{F}}^{\text{ann}}(n)$$

- For uniform two-sided convergence:

$$\lim_{n\to\infty} P\left(\sup_{\delta\in\mathcal{F}} |R(\delta) - R_{emp}(\delta)| > \epsilon\right) = 0$$

- It is necessary and sufficient that

$$\lim_{n\to\infty} H_{\mathcal{F}}(n)/n = 0$$

- We do not know the speed of convergence.
- Both of these convergence modes depend on the $p(\mathbf{x}, y)$.

Growth function

- Vapnik and Chervonenkis defined the growth function as:

$$G_{\mathcal{F}}(n) = \max_{Z_n} \log \mathcal{N}(\mathcal{F}, Z_n) = \log \mathcal{N}(\mathcal{F}, n)$$

- For uniform two-sided exponential convergence for any $p(\mathbf{x}, y)$:

$$\lim_{n\to\infty} P\left(\sup_{\delta\in\mathcal{F}} |R(\delta) - R_{emp}(\delta)| > \epsilon\right) = 0$$

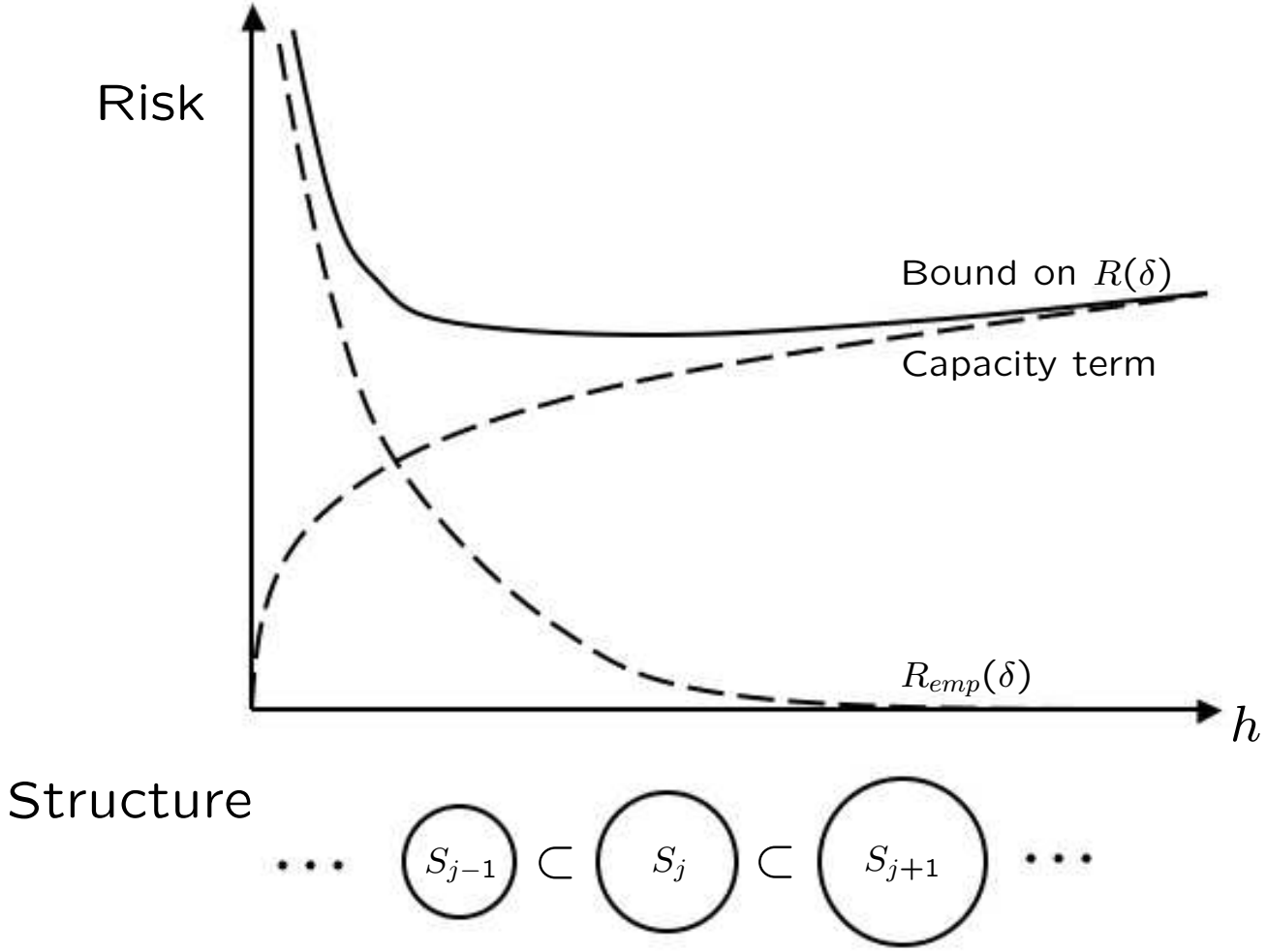
- It is necessary and sufficient that

$$\lim_{n\to\infty} G_{\mathcal{F}}(n)/n = 0$$

VC Dimension

- We can bound the growth function as:

$$G_{\mathcal{F}}(n) \leq \begin{cases} n \log(2), & n \leq h \\ h\left(\log \frac{n}{h} + 1\right), & n > h \end{cases}$$

- h is know as the VC dimension.
- If for a given $\mathcal{F}$, h is finite, we will have universal two-sided convergence of the empirical risk to the true risk.
- The converge will be exponentially fast.
- If h is infinite, learning might not occur and we might overfit for any n.

Structural Risk Minimization