Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha

Lecture 9: Statistical Learning Theory (Capacity)

- General model of learning & ERM (Vapnik 0.1-1.11)
- Consistency (Vapnik 3.1-3.2.1)
- Uniform Convergence (Vapnik 3.3, 3.4, 3.7)
- Entropy, Capacity (Vapnik 3.7, 3.10, 3.13)
- Bounds (Vapnik 4.1, 4.8)
- VC Dimension (Vapnik 4.9.1, 4.11)
- Structural Risk Minimization (SRM)

Empirical Processes

- Consider a sequence of random variables which depends both on the pdf and the set of functions:

$$r_\ell = \sup_\alpha \left| R(\alpha) - R_{emp}(\alpha_\ell) \right|$$

$$= \sup_\alpha \left| \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell} \sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right|$$

$$r_\ell^+ = \sup_\alpha \left(R(\alpha) - R_{emp}(\alpha_\ell) \right)_+ \qquad (u)_+ = \begin{cases} u & if\ u > 0, \\ 0 & otherwise. \end{cases}$$

- We call this sequence a *one-sided (two-sided) empirical process*
- Why are we concerned with one-sided process?
- Looking for consistency results in minimizing risk!

Uniform Convergence

- We want conditions for convergence (in probability):
- Two sided:

$$P\left\{ \sup_{\alpha} \left| \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell} \sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right| > \varepsilon \right\} \xrightarrow[\ell \to \infty]{} 0$$

- One-sided:

$$P\left\{ \sup_{\alpha} \left(\int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell} \sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right) > \varepsilon \right\} \xrightarrow[\ell \to \infty]{} 0$$

- We call these relations *uniform (two/one-sided) convergence of means to their mathematical expectation over a given set of functions*
- *Lets just say uniform convergence or U.C*
- How do we know that such convergence is equivalent to strict consistency?

Key Equivalence Theorem

- **Key Theorem:** suppose that for all functions in the set $\{L(\mathbf{z},\alpha)\}$ and all PDFs in the set $\{F(\mathbf{z})\}$ the inequalities below hold true

$$c \leq \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) \leq C$$

Then,

For any pdf in the set $\{F(\mathbf{z})\}$, the ERM method is strictly consistent on $\{L(\mathbf{z},\alpha)\}$

IF AND ONLY IF

For any pdf in the set $\{F(\mathbf{z})\}$, one-sided U.C takes place on the set $\{L(\mathbf{z},\alpha)\}$

Law of Large Numbers

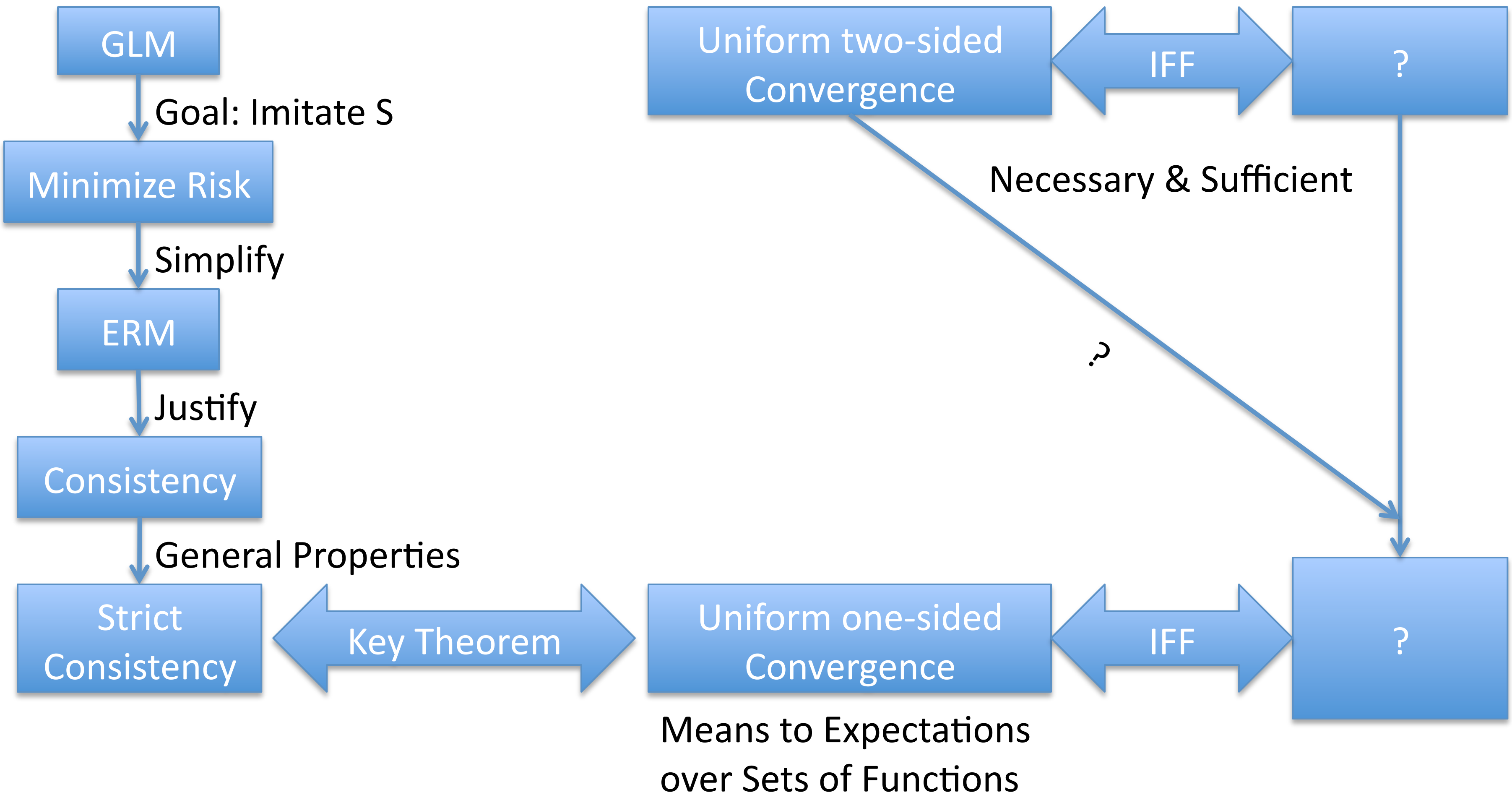
- Law of Large Numbers (LLN): the sequence of means converges to expectation of a random variable as the number of examples increases
- Strong LLN: A.S convergence
- Weak LLN: convergence in probability
- Uniform LLN: generalization for functions (instead of variables)

- Problem: ULLN applies to one function, we have sets of functions!
  - ➢ LLN can be applied if we fix “alpha”. We have a sup over the set of all alphas
  - ➢ Moreover we can have sets with infinite number of elements!
- Solution: need to generalize LLN to functional space
- Note: Glivenko – Cantelli theorem shows that ULLN holds for specific sets of functions (with bounds on asymptotic rate of convergence)

Recap

- We are interested in conditions for (strict) consistency of ERM
- Key Theorem proves that we should demonstrate conditions for uniform one-sided convergence
- We already have results (LLN) that demonstrate conditions for two-sided convergence
- But we have a more general case (sets of functions)
- Approach: find conditions for two-sided U.C and then obtain corresponding conditions for one-sided U.C

Road Map (2)


GLM
Goal: Imitate S
Minimize Risk
Simplify
ERM
Justify
Consistency
General Properties
Strict Consistency
Key Theorem
Uniform two-sided Convergence
IFF
?
Necessary & Sufficient
?
Uniform one-sided Convergence
IFF
?
Means to Expectations over Sets of Functions

Indicator Functions

- Until now we didn't care about the specific properties of the set $\{L(\mathbf{z},\alpha)\}$
- To describe conditions for (two-sided) U.C, consider indicator functions:

$$L(y,g(\mathbf{x},\alpha)) = \begin{cases} 0 & \text{if } y = g \\ 1 & \text{if } y \neq g \end{cases}$$

- We are now considering convergence of frequencies to probabilities:

$$P\left\{ \sup_{\alpha} \left| \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell}\sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right| > \varepsilon \right\} \xrightarrow[\ell\to\infty]{} 0$$

$$P\left\{ \sup_{\alpha} \left| P\{L(\mathbf{z},\alpha) > 0\} - v_{\ell}\{L(\mathbf{z},\alpha) > 0\} \right| > \varepsilon \right\} \xrightarrow[\ell\to\infty]{} 0$$

$$P\left\{ \sup_{\alpha} \left| p_{L>0} - v_{\ell} \right| > \varepsilon \right\} \xrightarrow[\ell\to\infty]{} 0$$

Notation

• For indicator functions we assume that g(x, alpha) outputs the class label (not a real value). For simplicity assume its a binary class label {0,1}.

$$L(y, g(\mathbf{x}, \alpha)) = \begin{cases} 0 & if \;\; y = g \\ 1 & if \;\; y \neq g \end{cases}$$

• We are now considering convergence of frequencies to probabilities, therefore by v_{L} we denote the frequencies and by p_{L} the probabilities of {L > 0}. This is the same as frequencies/probabilities of {L = 1} for binary classification, in other words counting the number of mistakes.

$$P\left\{ \sup_{\alpha} \left| P\{L(\mathbf{z}, \alpha) > 0\} - v_{\ell}\{L(\mathbf{z}, \alpha) > 0\} \right| > \varepsilon \right\} \xrightarrow[\ell \to \infty]{} 0$$

$$P\left\{ \sup_{\alpha} \left| p_{L>0} - v_{\ell} \right| > \varepsilon \right\} \xrightarrow[\ell \to \infty]{} 0$$

Case 1: One Function

- Suppose our set of functions contains just one function (one set of parameters)

$$\alpha \in \Lambda, |\Lambda| = 1 \Rightarrow \sup_{\alpha} \equiv \sup_{\alpha=\alpha^*}$$

- The supremum disappears
- Special case of LLN: just like tossing a coin
- We know that the frequencies converge to the probability as $\ell \to \infty$
- Moreover, we know the rate of convergence (Chernoff bound):

$$P\left\{ \sup_{\alpha} \left| P\left\{ L\left(\mathbf{z},\alpha^*\right) > 0 \right\} - v_\ell \left\{ L\left(\mathbf{z},\alpha^*\right) > 0 \right\} \right| > \varepsilon \right\} \xrightarrow[\ell \to \infty]{} 0$$

$$P\left\{ \left| p_{L>0} - v_\ell \right| > \varepsilon \right\} \xrightarrow[\ell \to \infty]{} 0$$

$$P\left\{ \left| p_{L>0} - v_\ell \right| > \varepsilon \right\} \leq 2\exp\left\{ -2\varepsilon^2 \ell \right\}$$

Chernoff Bounds

- Consider m independent coin flips (Bernoulli trials). Let S denote the # of heads observed, and let μ denote the expected value of S
  - ➢ What is the probability that S deviates from its mean by an amount ε?
- Another way to ask the same question: consider success probability p^ = S/m instead of S (actual number)
  - ➢ How fast does the estimate p^ converge to p as a function of m?
- Notation:

$$S = X_1 + \ldots + X_m,\ X_i \in \{0,1\},\ 0 \le \varepsilon \le 1$$

$$\Pr[X_i = 1] = p, \quad \mu = E[S] = pm, \quad p^{\wedge} = S/m$$

- Additive Form:

$$\Pr[S > (p+\varepsilon)m] \le \exp\left\{-2\varepsilon^2 m\right\} \qquad \Pr[S < (p-\varepsilon)m] \le \exp\left\{-2\varepsilon^2 m\right\}$$

$$\Pr[S < (p-\varepsilon)m] \Rightarrow \Pr[\frac{S}{m} < (p-\varepsilon)] \Rightarrow \Pr[p^{\wedge} < (p-\varepsilon)] \Rightarrow \Pr[p - p^{\wedge} > \varepsilon]$$

Chernoff Bounds

- Notation: $S = X_1 + \ldots + X_m,\ X_i \in \{0,1\},\ 0 \le \varepsilon \le 1$

$$\Pr[X_i = 1] = p, \quad \mu = E[S] = pm, \quad \hat{p} = \frac{S}{m}$$

- Additive Form:

$$\Pr[S > (p+\varepsilon)m] \le \exp\left\{-2\varepsilon^2 m\right\} \qquad \Pr[S < (p-\varepsilon)m] \le \exp\left\{-2\varepsilon^2 m\right\}$$

$$\Pr[\hat{p} - p > \varepsilon] \le \exp\left\{-2\varepsilon^2 m\right\} \qquad \Pr[p - \hat{p} > \varepsilon] \le \exp\left\{-2\varepsilon^2 m\right\}$$

$$\Rightarrow \Pr[\left|p - \hat{p}\right| > \varepsilon] = \Pr[\hat{p} - p > \varepsilon] + \Pr[p - \hat{p} > \varepsilon] \le 2\exp\left\{-2\varepsilon^2 m\right\}$$

Case 2: Finite Number of Functions

- Suppose our set contains N functions (where N is finite)

$$\alpha_{1,\ldots,N} \in \Lambda, |\Lambda| = N \Rightarrow \sup_{\alpha} \equiv \max_{\alpha}$$

- Easy to generalize case 1 using Chernoff bounds:

$$P\left\{ \max_{1 \le k \le n} \left| P\{L(\mathbf{z},\alpha_k) > 0\} - v_\ell\{L(\mathbf{z},\alpha_k) > 0\} \right| > \varepsilon \right\}$$

$$\le \sum_{k=1}^{N} P\left\{ \left| p_{L>0}(k) - v_\ell(k) \right| > \varepsilon \right\}$$

$$\le 2N \exp\left\{-2\varepsilon^2 \ell\right\} = 2\exp\left\{\ln N - 2\varepsilon^2 \ell\right\} = 2\exp\left\{\left(\frac{\ln N}{\ell} - 2\varepsilon^2\right)\ell\right\}$$

- What's the point behind the last manipulation?

Case 3: Infinite Number (idea)

- For U.C to take place we need the relation below to be satisfied

$$P\left\{ \max_{1 \le k \le n} \left| P\{L(\mathbf{z},\alpha_k) > 0\} - v_\ell\{L(\mathbf{z},\alpha_k) > 0\} \right| > \varepsilon \right\} \le 2\exp\left\{ \left(\frac{\ln N}{\ell} - 2\varepsilon^2 \right) \ell \right\}$$

$$\forall \varepsilon: \quad P\{ |\circ| > \varepsilon \} \xrightarrow[\ell \to \infty]{} 0 \quad \Leftrightarrow \quad \frac{\ln N}{\ell} \xrightarrow[\ell \to \infty]{} 0$$

- Obviously holds when N is finite. Can we generalize to infinite number of events?
- Lets introduce a new concept:
  - ➢ Set may contain infinite number of events/functions, but only a finite number of clusters of events is distinguishable for a given sample (of ℓ examples)
  - ➢ Distinguishable if there exist (at least) one element in the sample that belongs to one event but not to the other
  - ➢ Idea: denote number of clusters by N^, show that ln(N^) must increase slowly (not exponentially) as the sample size grows for U.C to hold

Entropy (Information Theory)

• Entropy is a measure of uncertainty of a random variable

• Another meaning: expected value of the information contained in a message (introduced by Claude Shannon developing communication theory, 1948)

• For a random variable X with n outcomes {x1,….,xn} the entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i)$$

• Can easily generalize to infinite outcomes (integral instead of sum)

• The higher the entropy value, the more uncertain we are about the outcome of the variable for a given trial/draw

Entropy of a Function Set

- Consider an arbitrary sequence of iid generated vectors $\{z_1,\ldots,z_\ell\}$
- Using our set of indicator functions, determine a set of binary vectors:

$$q(\alpha) = [L(z_1,\alpha),\ldots,L(z_\ell,\alpha)]$$

- For any fixed alpha, q(alpha) determines some vertex of the unit cube
- Denote the number of different vertices induced by the sample & function set as:

$$N^\wedge(z_1,\ldots,z_\ell) \leq 2^\ell$$

- *Random Entropy* (of the set of indicator functions on the given sample):

$$H^\wedge(z_1,\ldots,z_\ell) = \ln N^\wedge(z_1,\ldots,z_\ell)$$

- *Entropy* (of the set of indicator functions on samples of size ℓ):

$$H^\wedge(\ell) = E\left[H^\wedge(z_1,\ldots,z_\ell)\right] = \int H^\wedge(z_1,\ldots,z_\ell)\,dF(z_1,\ldots,z_\ell)$$

U.C (2-sided) Theorem

- **Theorem:**

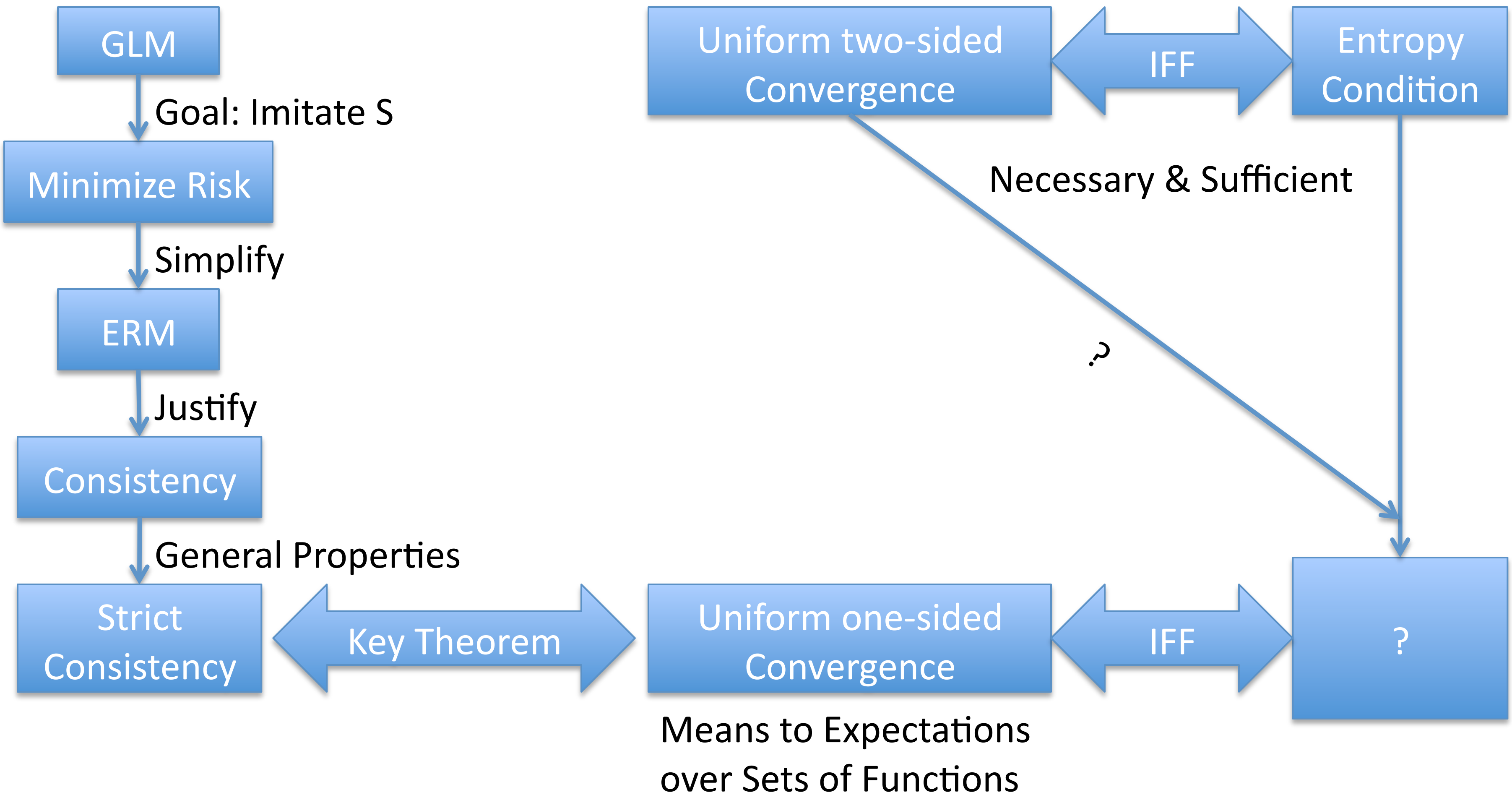
Two-sided U.C over the set of indicator functions takes place

$$P\left\{ \sup_{\alpha} \left| \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell}\sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right| > \varepsilon \right\} \xrightarrow[\ell\to\infty]{} 0$$

IF AND ONLY IF

$$\frac{H^{\wedge}(\ell)}{\ell} \xrightarrow[\ell\to\infty]{} 0$$

Road Map (3)


GLM
Goal: Imitate S
Minimize Risk
Simplify
ERM
Justify
Consistency
General Properties
Strict Consistency
Key Theorem
Uniform one-sided Convergence
Means to Expectations over Sets of Functions
IFF
?
Uniform two-sided Convergence
IFF
Entropy Condition
Necessary & Sufficient
?