

统计学习理论

张超

统计学习遇到的问题

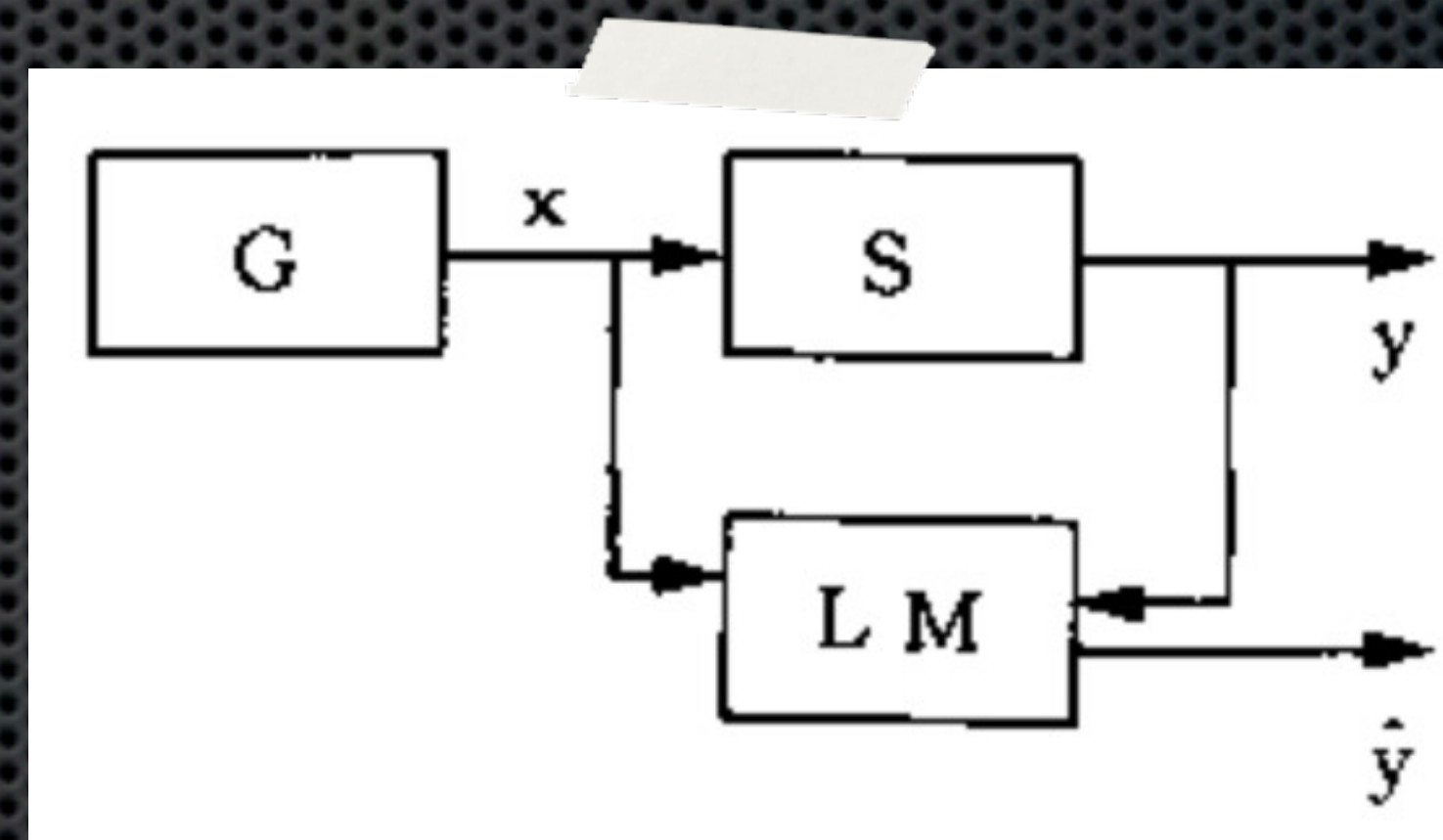
- 学习应使用的数据量
- 学习机器的收敛速度（计算复杂度）
- 学习过程中机器出现的错误
- 学习得到的函数与目标函数的差异及带来的错误率

没有什么比一个好的理论更实用了

—— Vladimir N. Vapnik

函数估计模型

- 产生器G: $x \sim F(x)$
- 训练器S: $F(y|x)$
- 学习机器LM: $f(x, \alpha), \alpha \in \Lambda$
- 学习问题:
 - $(x_1, y_1), \dots, (x_l, y_l) \sim F(x, y) (\text{i.i.d})$



经验风险最小化

- 记 $(x_i, y_i) = z_i$, 损失函数: $Q(z, \alpha) = L(y, f(x, \alpha)), \alpha \in \Lambda$
- 度量损失的风险泛函: $R(\alpha) = \int Q(z, \alpha) dF(z)$
- 经验风险泛函: $R_{\text{emp}}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha)$
- 经验风险最小化 (ERM) 原则:
 - 使用 $R_{\text{emp}}(\alpha)$ 来替换 $R(\alpha)$
 - 用使 $R_{\text{emp}}(\alpha)$ 最小的函数 $Q(z, \alpha')$ 来逼近使 $R(\alpha)$ 最小的函数 $Q(z, \alpha_0)$

统计学习理论的四个部分

1. 一个基于ERM原则的学习过程具有一致性的条件是什么？

→ 学习过程一致性的理论

2. 学习过程收敛的速度（样本数目）有多快？

→ 学习过程收敛速度的非渐进理论

3. 如何控制学习过程的推广能力？

→ 控制学习过程的推广能力的理论

4. 怎样构造能够控制推广能力的算法？

→ 构造学习算法的理论

学习过程一致性的理论

- ERM原则的一致性:

- 若 $R(\alpha') \xrightarrow[l \rightarrow \infty]{P} \inf R(\alpha)$ 和 $R_{\text{emp}}(\alpha') \xrightarrow[l \rightarrow \infty]{P} \inf R(\alpha)$ 两序列收敛于同一极限
- 称ERM原则对函数集 $Q(z, \alpha), \alpha \in \Lambda$ 和概率分布函数 $F(z)$ 是一致的

- ERM原则一致性的充要条件:

- 当 $Q(z, \alpha), \alpha \in \Lambda$ 使得 $A \leq R(\alpha) \leq B$ 时,
- ERM原则一致的充要条件是 $\lim_{l \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon \right\} = 0, \forall \varepsilon > 0$
一致单边收敛成立
- 把一致性转化为一边单边收敛问题

学习过程一致性的理论（续）

- 示性损失函数集的熵:

- $N^\Lambda(z_1, \dots, z_l)$: 示性函数集中的函数能把样本分成多少种不同分类

- 熵: $H^\Lambda(l) = E \ln N^\Lambda(z_1, \dots, z_l)$, 反映期望的多样性

- 有界实损失函数集的熵:

- $N^\Lambda(\varepsilon; z_1, \dots, z_l)$: $(Q(z_1, \alpha), \dots, Q(z_l, \alpha)), \alpha \in \Lambda$ 上最小 ε -网格的元素数目

- VC熵: $H^\Lambda(\varepsilon; l) = E \ln N^\Lambda(\varepsilon; z_1, \dots, z_l)$

学习过程一致性的理论 (续2)

- $\lim_{l \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon \right\} = 0, \forall \varepsilon > 0$
 - 充分条件: $\lim_{l \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} |R(\alpha) - R_{\text{emp}}(\alpha)| > \varepsilon \right\} = 0 \Leftrightarrow \lim_{l \rightarrow \infty} \frac{H^\Lambda(\varepsilon; l)}{l} = 0$
 - 定义 $Q^*(z, \alpha^*), \alpha^* \in \Lambda^*$, 对每一个 $Q(z, \alpha)$ 存在一个 $Q^*(z, \alpha^*)$
 - A. $Q(z, \alpha) - Q^*(z, \alpha^*) \geq 0, \forall z$
 - B. $\int (Q(z, \alpha) - Q^*(z, \alpha^*)) dF(z) \leq \delta$
- 充要条件: $\forall \delta, \eta, \varepsilon > 0, \exists Q^*(z, \alpha^*), \alpha^* \in \Lambda^*$, 使得 $\lim_{l \rightarrow \infty} \frac{H^{\Lambda^*}(\varepsilon; l)}{l} < \eta$

收敛速度非渐进性理论

- 无限样本学习过程的收敛速度:

- 退火的VC熵: $H_{\text{ann}}^{\Lambda}(l) = \ln EN^{\Lambda}(z_1, \dots, z_l)$

- 学习过程收敛速度快 $\Leftarrow \lim_{l \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda}(l)}{l} = 0$

- 一致性和收敛速度不依赖概率测度:

- 生长函数: $G^{\Lambda}(l) = \ln \sup_{z_1, \dots, z_l} N^{\Lambda}(z_1, \dots, z_l)$

- ERM学习过程对任何概率测度具有一致性 $\Leftrightarrow \lim_{l \rightarrow \infty} \frac{G^{\Lambda}(l)}{l} = 0$

- ERM学习过程收敛速度快 $\Leftarrow \lim_{l \rightarrow \infty} \frac{G^{\Lambda}(l)}{l} = 0$

- $H^{\Lambda}(l) \leq H_{\text{ann}}^{\Lambda}(l) \leq G^{\Lambda}(l)$

收敛速度非渐进性理论 (续)

- 有限样本学习过程的收敛速度（下界也存在）：

1. 示性损失函数依赖分布的上界：

$$\begin{aligned} & \blacksquare P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right| > \varepsilon \right\} \leq 4e \left\{ \left(\frac{H_{\text{ann}}^{\Lambda}(2l)}{l} - \varepsilon^2 \right)^l \right\} \\ & \blacksquare P \left\{ \sup_{\alpha \in \Lambda} \frac{\left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right|}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \leq 4e \left\{ \left(\frac{H_{\text{ann}}^{\Lambda}(2l)}{l} - \frac{\varepsilon^2}{4} \right)^l \right\} \end{aligned}$$

收敛速度非渐进性理论 (续2)

- 有限样本学习过程的收敛速度 (下界也存在):

2.对实函数集的推广: $Q(z, \alpha), \alpha \in \Lambda$ 的示性函数集 $I(z, \alpha, \beta), \alpha \in \Lambda, \beta \in \mathcal{B}$

1)完全有界实损失函数集

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right| > \varepsilon \right\} \leq 4e \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2l)}{l} - \frac{\varepsilon^2}{(B-A)^2} \right)^l$$

2)完全有界非负实损失函数集

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{\left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right|}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \leq 4e \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2l)}{l} - \frac{\varepsilon^2}{4B} \right)^l$$

收敛速度非渐进性理论 (续3)

- 有限样本学习过程的收敛速度 (下界也存在):

3.收敛速度与分布 $F(z)$ 无关的上界: 用 $G^{\Lambda, \mathcal{B}}(2l)$ 替换 $H_{\text{ann}}^{\Lambda, \mathcal{B}}(2l)$

1)完全有界实损失函数集

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right| > \varepsilon \right\} \leq 4e \left(\frac{G^{\Lambda, \mathcal{B}}(2l)}{l} - \frac{\varepsilon^2}{(B-A)^2} \right)^l$$

2)完全有界非负实损失函数集

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{\left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right|}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \leq 4e \left(\frac{G^{\Lambda, \mathcal{B}}(2l)}{l} - \frac{\varepsilon^2}{4B} \right)^l$$

推广能力的界的理论

- ERM学习机器推广能力的概念性界:
 - 最小经验风险 $R_{\text{emp}}(\alpha')$ 的函数 $Q(z, \alpha')$ 的期望风险 $R(\alpha')$ 是多少?
 - 对任意函数集 $R_{\text{emp}}(\alpha')$ 与真正的期望风险最小 $\inf_{\alpha \in \Lambda} R(\alpha)$ 的接近程度怎样?

- 记 $\mathcal{E} = 4 \frac{G^{\Lambda, \mathcal{B}}(2l) - \ln(\frac{\eta}{4})}{l}$

推广能力的界的理论（续）

- ERM学习机器推广能力的概念性界:

1. 完全有界损失函数集:

- 1) 下面不等式至少以 $1 - \eta$ 的概率对 $Q(z, \alpha), \alpha \in \Lambda$ 中所有函数成立

- $R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B - A}{2} \sqrt{\mathcal{E}}$

与一致收敛速度的界等价

- $R_{\text{emp}}(\alpha) - \frac{B - A}{2} \sqrt{\mathcal{E}} \leq R(\alpha)$

- 2) 下面不等式至少以 $1 - 2\eta$ 的概率对使经验风险最小的函数 $Q(z, \alpha')$ 成立

- $R(\alpha') - \inf_{\alpha \in \Lambda} R(\alpha) \leq (B - A) \sqrt{\frac{-\ln \eta}{2l}} + \frac{B - A}{2} \sqrt{\mathcal{E}}$

推广能力的界的理论 (续2)

- ERM学习机器推广能力的概念性界:

2.完全有界非负损失函数集:

1)下面不等式至少以 $1 - \eta$ 的概率对 $Q(z, \alpha), \alpha \in \Lambda$ 中所有函数成立

$$\blacksquare R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}}{2} \left[1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}}} \right] \quad \text{相对偏差一致收敛速度}$$

2)下面不等式至少以 $1 - 2\eta$ 的概率对使经验风险最小的函数 $Q(z, \alpha')$ 成立

$$\blacksquare R(\alpha') - \inf_{\alpha \in \Lambda} R(\alpha) \leq B \sqrt{\frac{-\ln \eta}{2l}} + \frac{B\mathcal{E}}{2} \left[1 + \sqrt{1 + \frac{4}{\mathcal{E}}} \right]$$

推广能力的界的理论 (续3)

- ERM学习机器推广能力的构造性界:

- 任何生长函数,

- 或者满足 $G^\Lambda(l) \leq h \left(\ln \frac{l}{h} + 1 \right)$

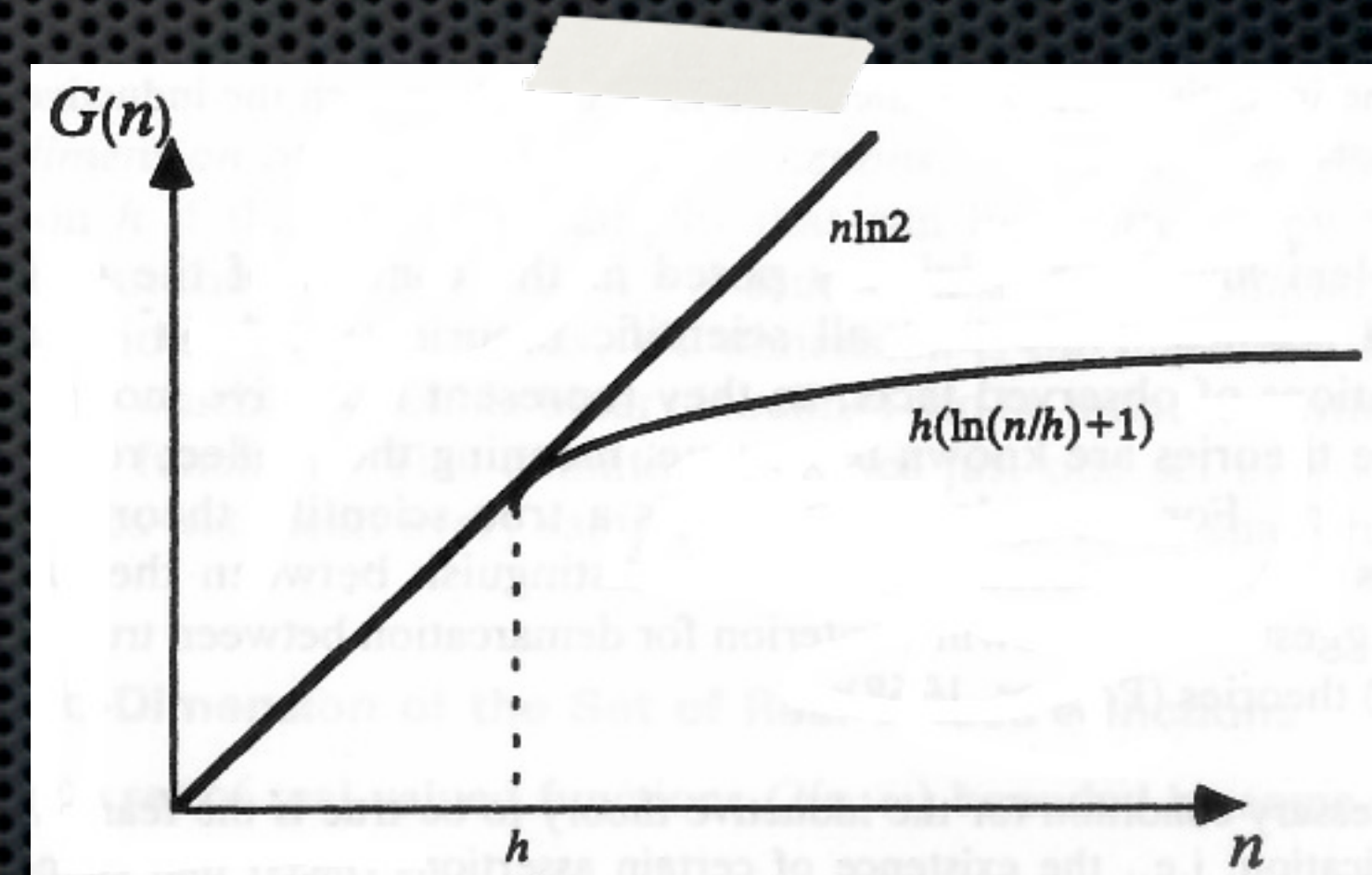
- 或者满足 $G^\Lambda(l) = l \ln 2$

- h 是整数, 满足 $l = h$ 时, $G^\Lambda(h) = h \ln 2$ 且

$$G^\Lambda(h+1) < (h+1) \ln 2$$

- 示性函数集的生长函数是线性的, VC维无穷大

- 示性函数集的生长函数以 h 的对数函数为界, VC维等于 h

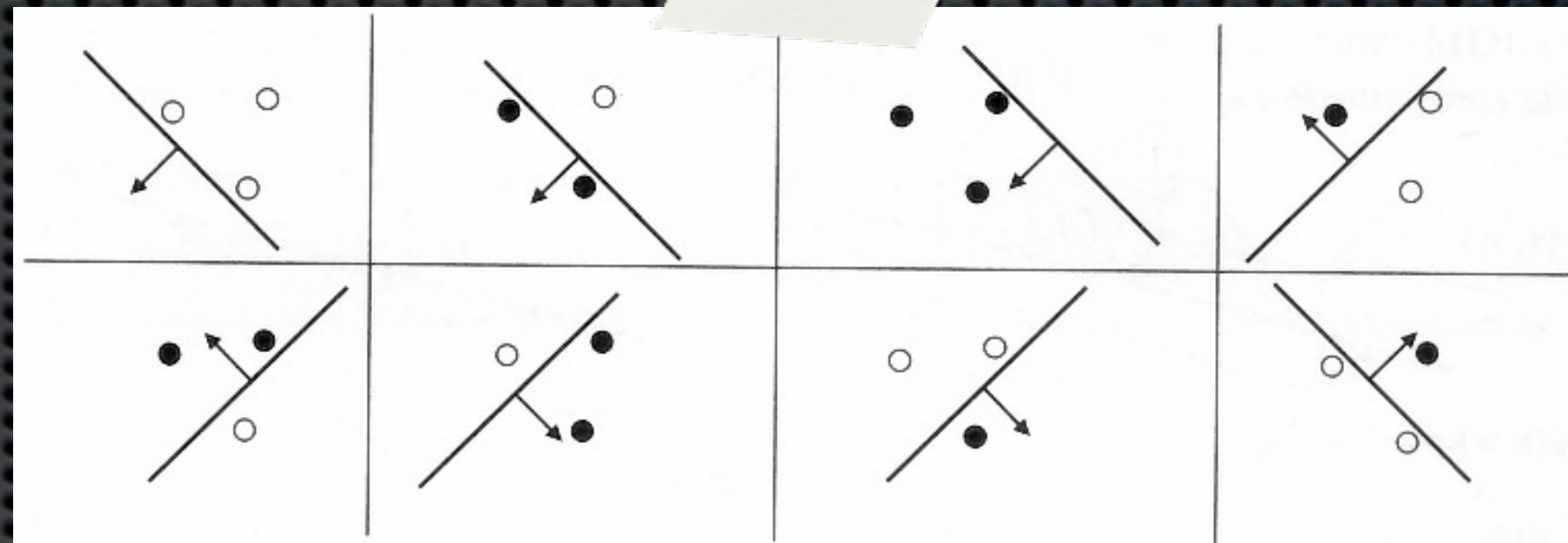


推广能力的界的理论 (续4)

▪ 示性函数集 $Q(z, \alpha), \alpha \in \Lambda$ 的 VC 维:

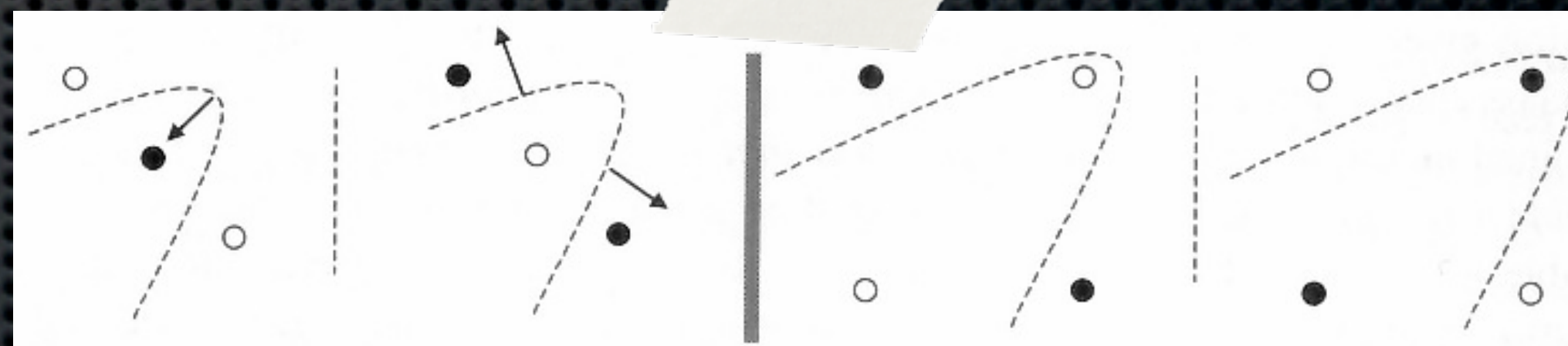
▪ 被集合中函数以 2^h 种方式分成两类的向量 $z_1, \dots, z_h$ 的最大数目 h

▪ 如果对任意 n , 总存在一个 n 个向量的集合可以被函数集打散, 则 VC 维无穷大



▪ 实函数集 $A \leq Q(z, \alpha) \leq B, \alpha \in \Lambda$ 的 VC 维:

▪ $I(z, \alpha, \beta), \alpha \in \Lambda, \beta \in (A, B)$



推广能力的界的理论 (续5)

- ERM学习机器推广能力的构造性界:

- $Q(z, \alpha), \alpha \in \Lambda$ VC维有限, 设为 h

- 函数数目无限: 记 $\mathcal{E} = 4 \frac{h \left(\ln \frac{2l}{h} + 1 \right) - \ln \left(\frac{\eta}{4} \right)}{l}$

- 函数数目为 N : 记 $\mathcal{E} = 2 \frac{\ln N - \ln \eta}{l}$

- 用 $h \left(\ln \frac{2l}{h} + 1 \right)$ 代替 $G^{\Lambda, \mathcal{B}}(2l)$

推广能力的界的理论 (续6)

- ERM学习机器推广能力的构造性界:

1. 完全有界函数集

1) 下面不等式至少以 $1 - \eta$ 的概率对 $Q(z, \alpha), \alpha \in \Lambda$ 中所有函数成立

$$\begin{aligned} \checkmark R(\alpha) &\leq R_{\text{emp}}(\alpha) + \frac{B - A}{2} \sqrt{\mathcal{E}} \\ \blacksquare R(\alpha) &\geq R_{\text{emp}}(\alpha) - \frac{B - A}{2} \sqrt{\mathcal{E}} \end{aligned}$$

2) 下面不等式至少以 $1 - 2\eta$ 的概率对使经验风险最小的函数 $Q(z, \alpha')$ 成立

$$\blacksquare R(\alpha') - \inf_{\alpha \in \Lambda} R(\alpha) \leq (B - A) \sqrt{\frac{-\ln \eta}{2l}} + \frac{B - A}{2} \sqrt{\mathcal{E}}$$

推广能力的界的理论 (续7)

2. 无界非负函数集

1) 下面不等式至少以 $1 - \eta$ 的概率对 $Q(z, \alpha), \alpha \in \Lambda$ 中所有函数成立

$$\sqrt{R(\alpha)} \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}}{2} \left[1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}}} \right]$$

2) 下面不等式至少以 $1 - 2\eta$ 的概率对使经验风险最小的函数 $Q(z, \alpha')$ 成立

$$\blacksquare R(\alpha') - \inf_{\alpha \in \Lambda} R(\alpha) \leq B \sqrt{\frac{-\ln \eta}{2l}} + \frac{B\mathcal{E}}{2} \left[1 + \sqrt{1 + \frac{4}{\mathcal{E}}} \right]$$

- 以上的界是不能被显著改进的
- 函数集的VC维影响了学习机器的推广性能

控制推广能力的理论

- 函数集的结构:

- 容许结构: 设函数的集合 S 具有函数子集 S_k 嵌套构成的结构

- $S_1 \subset S_2 \subset \cdots \subset S_k \cdots \subset S$

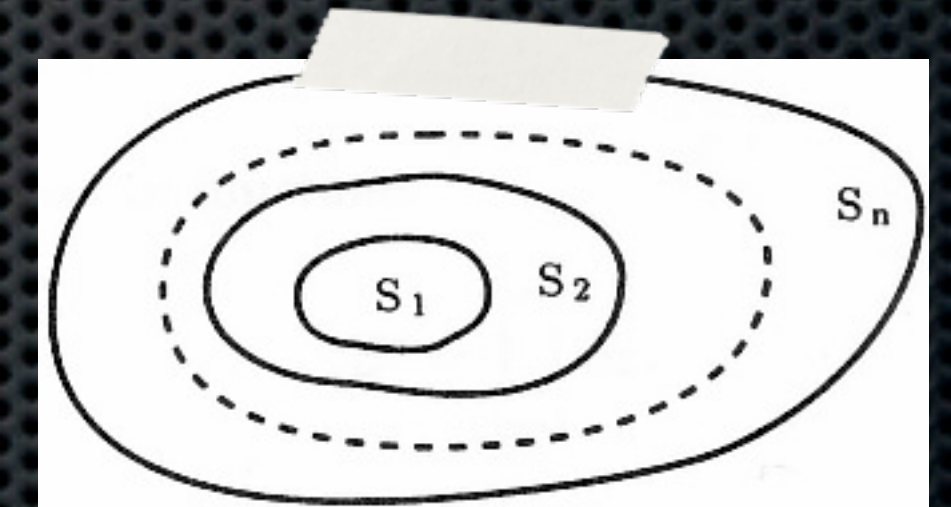
- $S_k = \{Q(z, \alpha), \alpha \in \Lambda_k\}$

- 1. 每个函数集的VC维都是有限的, 且 $h_1 \leq h_2 \leq \cdots \leq h_k \leq \cdots$

- 2. 结构中的任何元素 (函数子集)

- 或包含完全有界的正函数集 $0 < Q(z, \alpha) < B_k, \alpha \in \Lambda_k$

- 或包含对一定 (p, τ_k) , 满足 $\sup_{\alpha \in \Lambda_k} \frac{\left(\int Q^p(z, \alpha) dF(z)\right)^{\frac{1}{p}}}{\int Q(z, \alpha) dF(z)} \leq \tau_k, p > 2$ 的集合



控制推广能力的理论（续）

- 结构风险最小化（SRM）原则:

$$\mathcal{E} = 4 \frac{h \left(\ln \frac{2l}{h} + 1 \right) - \ln \left(\frac{\eta}{4} \right)}{l}$$

- $R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B-A}{2} \sqrt{\mathcal{E}}$
- 以及 $R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}}{2} \left[1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}}} \right]$
 - $\frac{l}{h}$ 较大时, $\mathcal{E}$ 较小, 经验风险接近期望风险, ERM准则合理
 - $\frac{l}{h}$ 较小时, 小的经验风险并不能保证小的期望风险合理
- 将上面式子记为 $R(\alpha) \leq R_{\text{emp}}(\alpha) + \Phi\left(\frac{h}{l}\right)$
- 求 $\min(R_{\text{emp}}(\alpha) + \Phi\left(\frac{h}{l}\right))$ 称作结构风险最小原则

控制推广能力的理论（续）

- 结构风险最小化（SRM）原则:

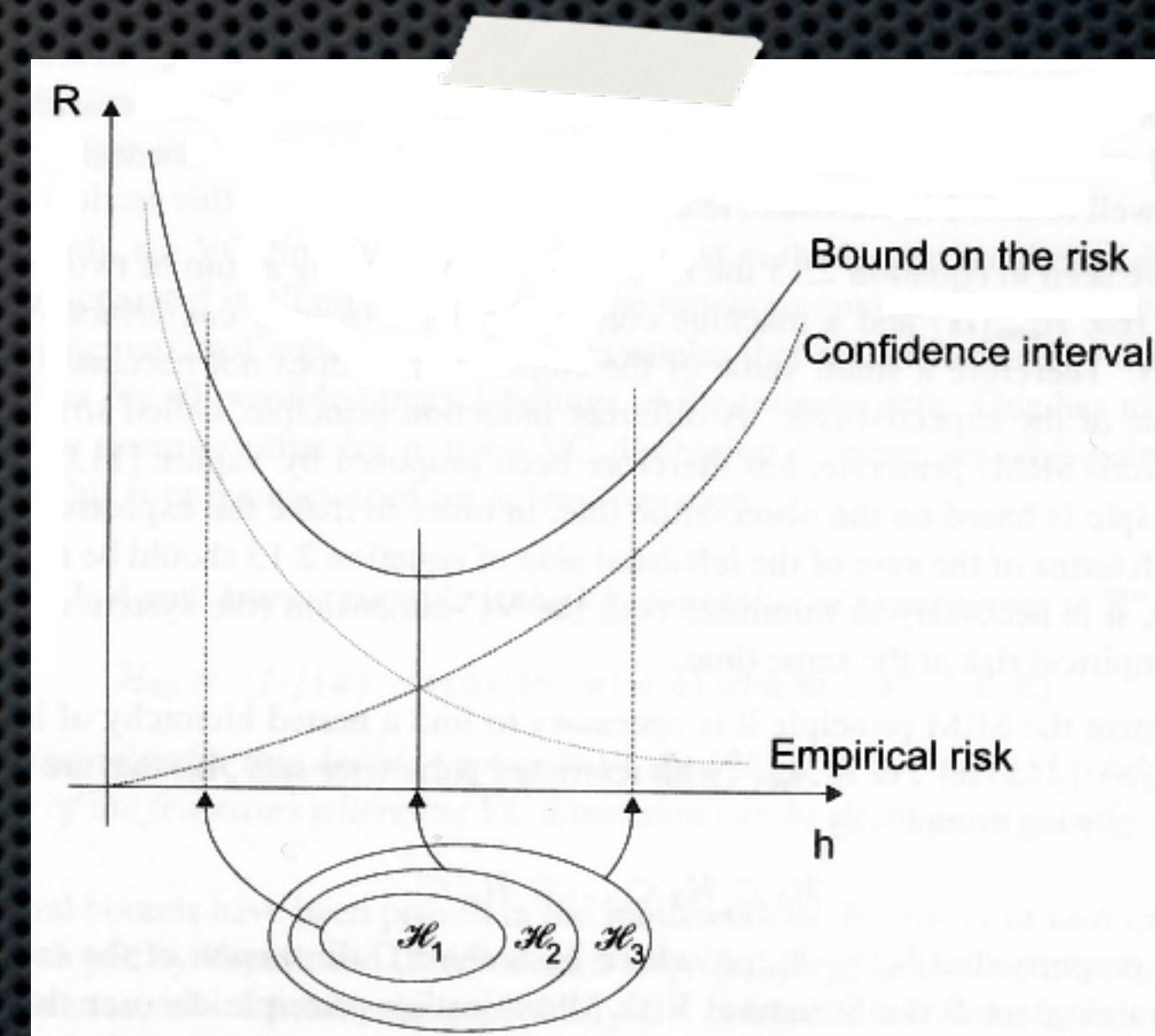
- $R(\alpha) \leq R_{\text{emp}}(\alpha) + \Phi\left(\frac{h}{l}\right)$

- $\Phi\left(\frac{h}{l}\right)$ 是置信范围，由函数集VC维决定（机器设计）

- 根据推广能力选择最好的函数子集

- $R_{\text{emp}}(\alpha)$ 取决函数集中特定函数，由训练方法决定

- 根据经验风险最小在函数子集中选择最好的函数



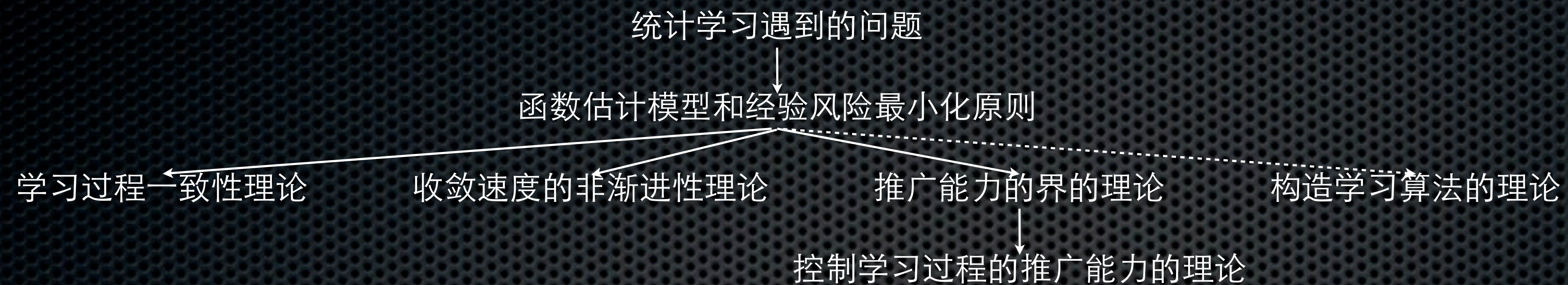
一些思考

- ✓ 对ERM做了深入分析，并在理论上找到其适用范围
- ✓ 包括控制小样本 ($\frac{l}{h} < 20$) 学习机器推广能力的原则
- ✓ 在最坏情况下进行推导，结论具有一般性
- ✓ 除理论分析外，还包括实用的算法 (SVM)
 - 最坏情况下推到得到的结论实际中有时是无用的
 - 不包括计算复杂度方面的内容

统计学习遇到的问题



函数估计模型和经验风险最小化原则



统计学习遇到的问题

函数估计模型和经验风险最小化原则

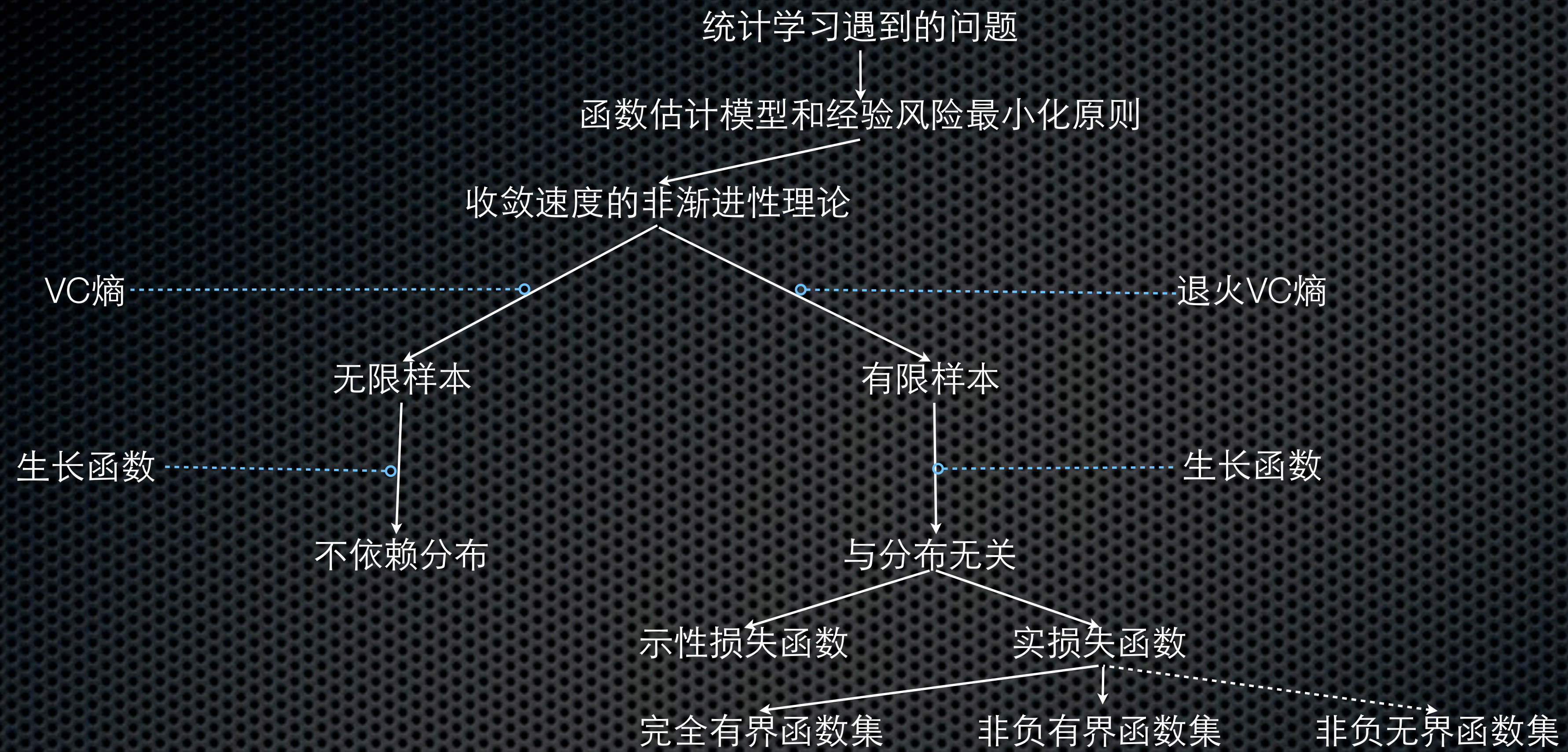
学习过程 ← 一致性理论

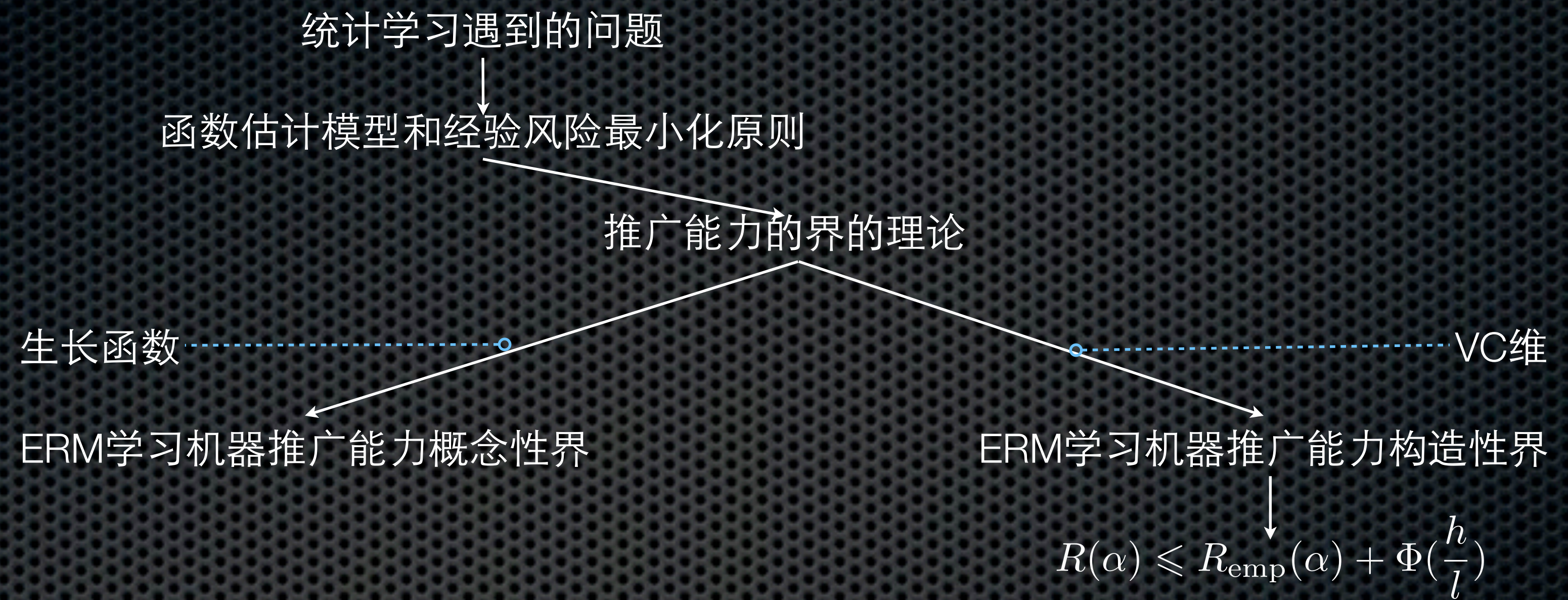
ERM原则一致性 (什么时候合理)

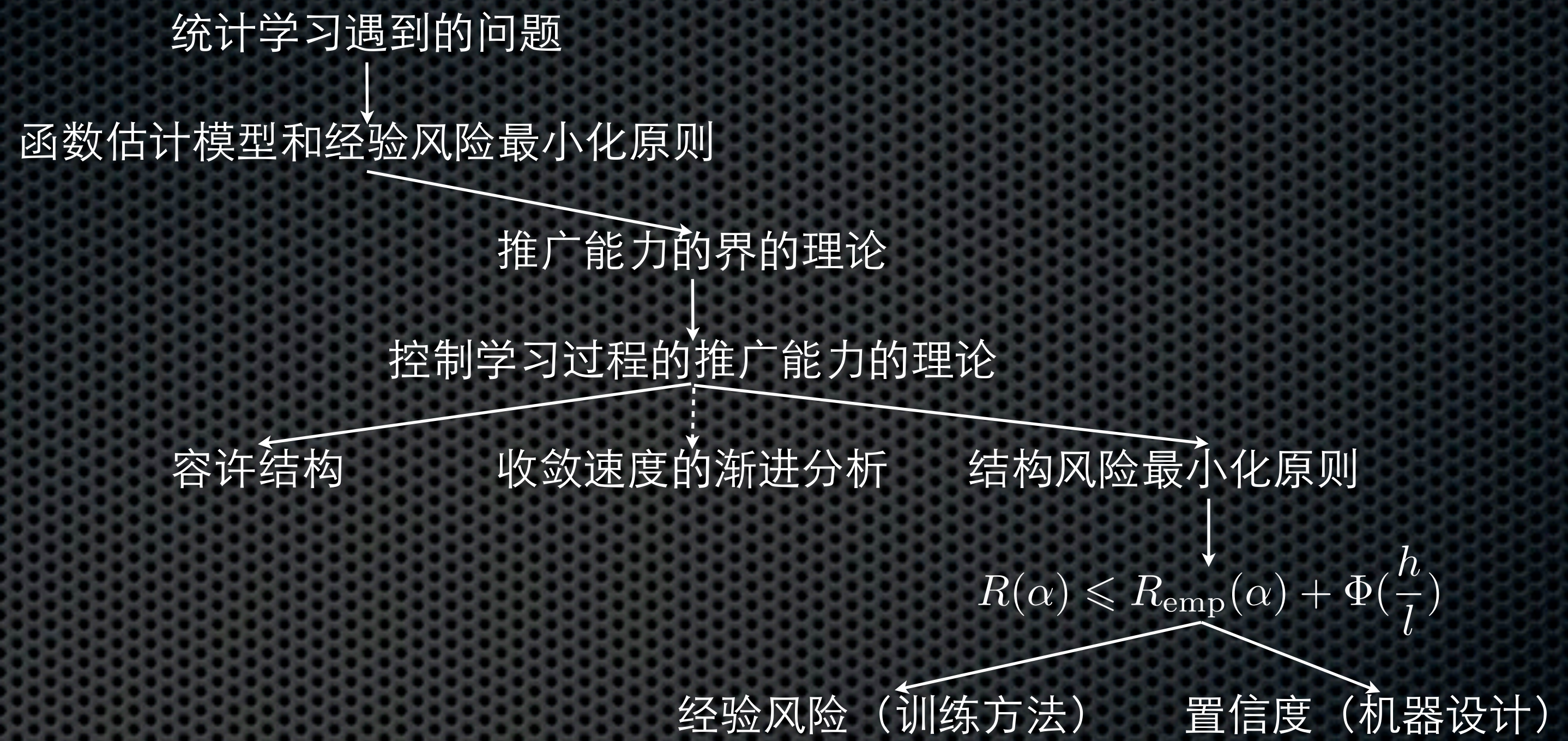
充要条件 (变为一致单边收敛问题)

VC熵 (期望的多样性)

$$\lim_{l \rightarrow \infty} \frac{H^{\Lambda^*}(\varepsilon; l)}{l} < \eta$$







参考文献

- ✦ V. Kecman. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. 2001, The MIT Press.
- ✦ Filip M. Mulier, Vladimir S. Cherkassky. Learning from Data. 1998, Wiley-Interscience.
- ✦ Rychetsky, Matthias. Algorithms and Architectures for Machine Learning based on Regularized Neural Networks and Support Vector Approaches. 2001, Shaker Verlag.
- ✦ Tom M. Mitchell. Machine Learning. 1997, McGraw Hill.
- ✦ V. Vapnik. The Nature of Statistical Learning Theory. 1995, Springer.
- ✦ V. Vapnik. Statistical Learning Theory. 1998, John Wiley and Sons.

谢谢！