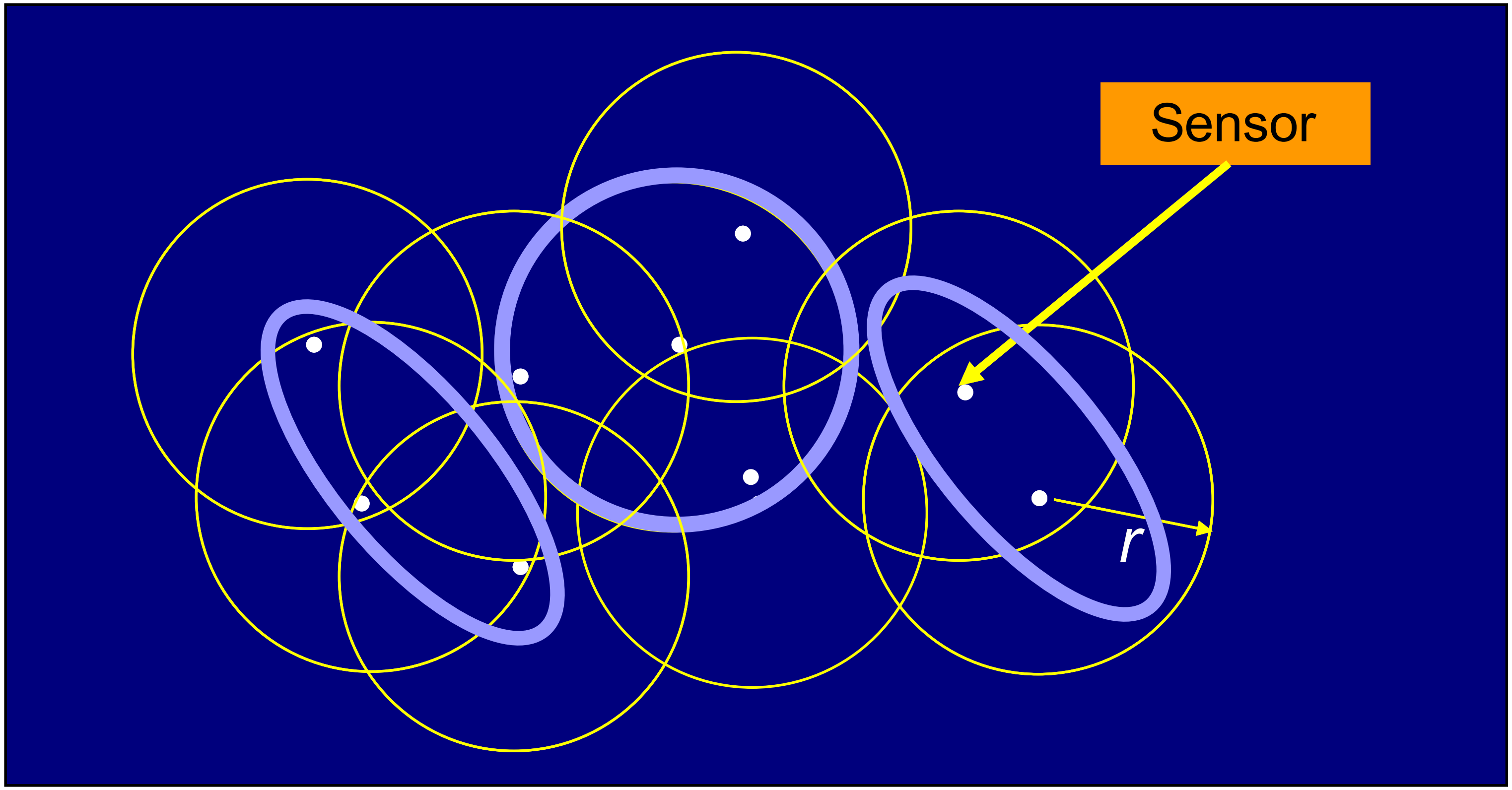
Lecture 5
Maximum Entropy and Minimum KL Divergence

Outline

1. Ill-posed Problems
2. Maximum Entropy
3. Minimal K-L Divergence
4. Connections between them
5. Intuition to ME Principle
6. Application of ME Principle

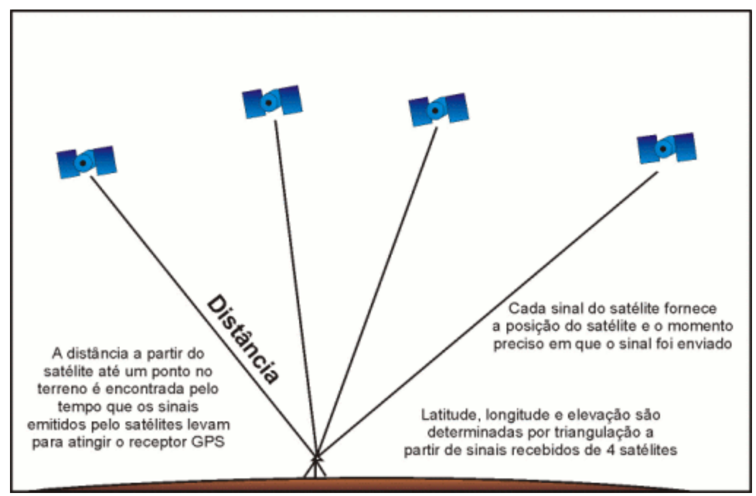
1 Ill-Posed Problem

- A simple example: sensor network localization



Sensor Network Localization

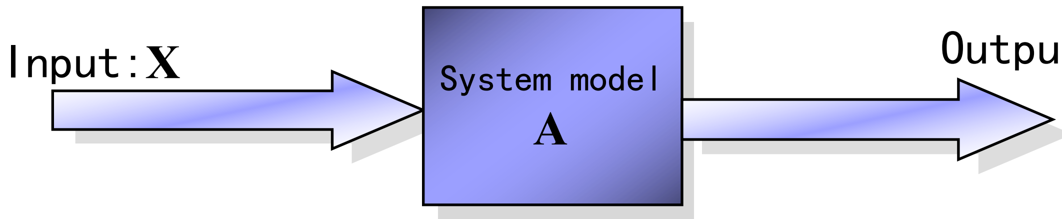
- Quasi-distance between some pairs of nodes
- Could we recover the coordinates of each nodes?
- GPS Revisited - over deterministic
- This problem – typically undeterministic



Regime of quantitative research

- Categories of problems
  - Direct problem: model the physical laws, determine the parameter of the model, input -> output
  - Reverse problem: based on observations, infer the system parameter and input
- ill-posed problems
  - Over-deterministic: too many clues.
  - Undeterministic: too little clues.

A simplified view



- Direct problem: know **X** and **A**, resolve **Y**
- Reverse problem: know **Y**, resolve **X** and **A**;
  know **Y** and **A**, resolve **X**

For over-deterministic problems

已知$\mathbf{AX}=\mathbf{Y}$，其中$\mathbf{A}$为$m\times n$矩阵，$\mathbf{X}$为n维列向量，$\mathbf{Y}$为m维列向量

因为是过定问题，有$m>n$，设$rank\mathbf{A}=n$（列满秩）。

用最小二乘求解本问题，就是求解一个$\hat{\mathbf{X}}$，使得

$\mathbf{J}=\left[\mathbf{A}\hat{\mathbf{X}}-\mathbf{Y}\right]^{\mathrm{H}}\left[\mathbf{A}\hat{\mathbf{X}}-\mathbf{Y}\right]$最小化。

Under-deterministic problem

已知$\mathbf{AX}=\mathbf{Y}$，其中$\mathbf{A}$为$m\times n$矩阵，$\mathbf{X}$为n维列向量，$\mathbf{Y}$为m维列向量

因为是欠定问题，有$m<n$。

对于这类问题，一般的解决方法是求方程$\mathbf{AX}=\mathbf{0}$的解，一般有$n-rankA$个，然后考虑$\mathbf{Y}$再求原方程的解。

问题

- ✓ 在所有可行解中是否还有倾向？
- ✓ 如何给出所有可行解的最准确估计？
- ✓ …

2 Maximum Entropy Principle

- Under-deterministic problem -> more than one solution
- Which one is most “reasonable”?
- E.T.Jayne proposed ME Principle in 1957

Edwin Thompson Jaynes
July 5, 1922 - April 30, 1998

Intuition

- ✓ Max Entropy -> “uniform” distribution
- ✓ LLN -> uniform distribution

Formal Problem

- Suppose we have a discrete RV X with unknown p.m.f. $p(x)$. Given its expectation of some functions

$$\sum_{x \in \mathrm{X}} p(x) f_m(x) = C_m \quad m = 1, 2, \cdots, M$$

determine the p.m.f. $\hat{p}(x)$

- Convert the problem into a constraint optimization problem
  - Objective function: $H(X) = -\sum_{x \in \mathrm{X}} p(x) \log p(x)$
  - Constraints:

$$\sum_{x \in \mathrm{X}} p(x) = 1,$$

$$\sum_{x \in \mathrm{X}} p(x) f_m(x) = C_m, m = 1, 2, \ldots, M$$

  - Solution:

$$\hat{p}(x) = \underset{p(x)}{Arg \max} H(X)$$

Maximum Entropy Theorem

- **Theorem 5.1:** The p.m.f. that achieve maximum entropy is

$$\hat{p}(x) = \exp\left[-\lambda_0 - \sum_{m=1}^{M} \lambda_m f_m(x)\right]$$

where $\lambda_0, \ldots, \lambda_M$ satisfy $\hat{p}(x)$

$$\sum_{x \in X} p(x) = 1,$$

$$\sum_{x \in X} p(x) f_m(x) = C_m, m = 1, 2, \ldots, M$$

Proof: Apply Lagrange multiplier.

Proof of Theorem 5.1

Let auxiliary function be

$$F = H(X) - \beta\left(\sum_{x\in X} p(x) - 1\right) - \sum_{m=1}^{M}\lambda_m\left(\sum_{x\in X} p(x) f_m(x) - C_m\right)$$

and by taking

$$\hat{p}(x) = \exp\left[-\lambda_0 - \sum_{m=1}^{M}\lambda_m f_m(x)\right]$$

we have

$$\frac{\partial F}{\partial p(x)} = -1 - \log p(x) - \beta - \sum_{m=1}^{M}\lambda_m f_m(x) = 0$$

where $\lambda_0=\beta+1$, and $\lambda_m (m=0,1,\dots,M)$ can be solved by $M+1$ constraints.

Continuous R.V.s

- Substitute entropy with differential entropy。

$$p(x) \geq 0, 且 p(x) = 0, 当 x \notin \mathrm{S}$$

$$\int_S p(x)dx = 1$$

$$\int_S p(x) f_m(x)dx = C_m, m = 1,2,3,\ldots,M$$

最大熵分布定理：满足约束条件且使微分熵达到最大值的分布为$\hat{p}(x) = \exp[\lambda_0 + \sum_{k=1}^{K} \lambda_m f_m(x)]$。

3 Minimum K-L Divergence

- ME principle: No *prior* knowledge on p.m.f.
- What if there is *prior* on p.m.f.
- K-L Divergence measure the difference between two p.m.f.
- Minimum K-L Divergence: under the given constraints, find a p.m.f. that is as close to the *prior* as possible.

S. Kullback
1903–1994

最小鉴别信息原理的问题描述

- 问题：某随机变量X，概率分布$q(x)$未知，已知其先验概率密度$p(x)$及其若干函数的期望

$$\int_S q(x) f_m(x) dx = C_m, m = 1,2,\ldots,M$$

求在上述条件下对$q(x)$的最佳估计。

- 按照最小鉴别信息原理，上述问题的求解可以表述为以下受限优化问题。
  - 取先验分布与目标分布之间的鉴别信息作为目标函数

$$D(\mathbf{q} \| \mathbf{p}) = \int_S q(x) \log \frac{q(x)}{p(x)} dx$$

  - 求在约束条件：

$$\int_s q(x) dx = 1 \qquad \int_s q(x) f_m(x) dx = C_m, m = 1,2,\ldots,M$$

  - 下的解

$$\hat{q}(x) = Arg \min_{q(x)} D(\mathbf{q} \| \mathbf{p})$$

Minimum K-L Divergence Principle

- **Theorem 5.2:** Given a prior p.m.f. p(x) and constraints, the minimum K-L divergence p.m.f is

$$\hat{q}(x) = p(x)\exp\left[\lambda_0 + \sum_{m=1}^{M} \lambda_m f_m(x)\right]$$

where $\lambda_0, \ldots, \lambda_M$ are taken such that $\hat{q}(x)$ satisfies

$$\sum_{x \in X} q(x) = 1,$$

$$\sum_{x \in X} q(x) f_m(x) = C_m, m = 1, 2, \ldots, M$$

Min KL Principle is a generalization of ME principle

Assume the *prior* is $p(x)=\dfrac{1}{K}$

则 $$\mathrm{D}(q(x)\,\|\,p(x)) = D(q(x)\,\|\,\frac{1}{K})$$

$$= \sum_{x\in \mathrm{X}} q(x)\log\frac{q(x)}{\frac{1}{K}} = \sum_{x\in \mathrm{X}} q(a_k)\log q(a_k) + \log K$$

$$= -H(X) + \log K$$

则 $\mathrm{D}(q(x)\,\|\,\frac{1}{K})$最小 $\Rightarrow H(X)$最大。

5 Intuition of ME Principle

- Physicality
- Second Law of Thermodynamics
- States far from equilibrium states is possible, but very rarely seen
  - A flock of money jumping on typewrite and type out the British Encyclopedia.
  - The second type of perpetual motion machine
  - Room temperature deviated from balanced state
  - Half glass of water jumping up into air

Jaynes对最大熵原理的解释

随机试验$X=\{a_1,a_2,\ldots,a_K\}$，连续进行N次试验，得到独立同分布随机序列X^N的一个实现，即$x^n=x_1x_2\ldots x_N$。它共有K^N种可能。设在K^N种可能序列中，第k个事件出现$N_k=Nf_k(k=1,2,\ldots,K)$次的序列共有$W(f_1,f_2,\ldots,f_K)$个。

$$W(f_1,f_2,\ldots,f_K)=\frac{N!}{(Nf_1)!(Nf_2)!\ldots(Nf_K)!}$$

使用*Stirling*阶乘近似公式，当n足够大时，$n!\approx\left(\frac{n}{e}\right)^n$

$$\lim_{N\to\infty}W(f_1,f_2,\ldots,f_K)=\lim_{N\to\infty}\frac{N!}{(Nf_1)!(Nf_2)!\ldots(Nf_K)!}=\frac{\left(\frac{N}{e}\right)^N}{\left(\frac{Nf_1}{e}\right)^{Nf_1}\left(\frac{Nf_2}{e}\right)^{Nf_2}\cdots\left(\frac{Nf_K}{e}\right)^{Nf_K}}=\prod_{i=1}^{K}\left(\frac{1}{f_i}\right)^{Nf_i}$$

而$H(f_1,f_2,\ldots,f_K)=\sum_{i=1}^{K}f_i\ln\frac{1}{f_i}$

故$W(f_1,f_2,\ldots,f_K)=\exp\left[NH(f_1,f_2,\ldots,f_K)\right]$

Jaynes对最大熵原理的解释（续）

于是，频率$\{f_k, k=1,2,...,K\}$可以看作是K维空间中的一个点P，它构成一个凸集$S=\{P: f_k \ge 0, \sum_{k=1}^{K} f_k = 1\}$。

- 在S的顶点：　$H(f_1, f_2, ..., f_K) = 0$
- 在S的内部：　$\max H(f_1, f_2, ..., f_K) = \log K$

$M+1$个线性约束下，存在$L=K-M-1$维凸约束集S_M，所有的可行解被限制在集合$S' = S_M \cap S$中，其维数为$K-M-1$。S'具有这样的性质：

- 满足约束条件的解在S'上取到
- 熵是S'上的凸函数，存在唯一最大值点

在$L=K-M-1$空间中进行座标的线性变换，使熵函数在原点处取得最大值。

Jaynes对最大熵原理的解释（续）

将熵函数在原点附近进行级数展开

$$H(P) = H_{\max} - ar^2 + \ldots, (a > 0)$$

r是P到原点的距离 $r = (\sum_{k=1}^{L} x_k)^{\frac{1}{2}}$

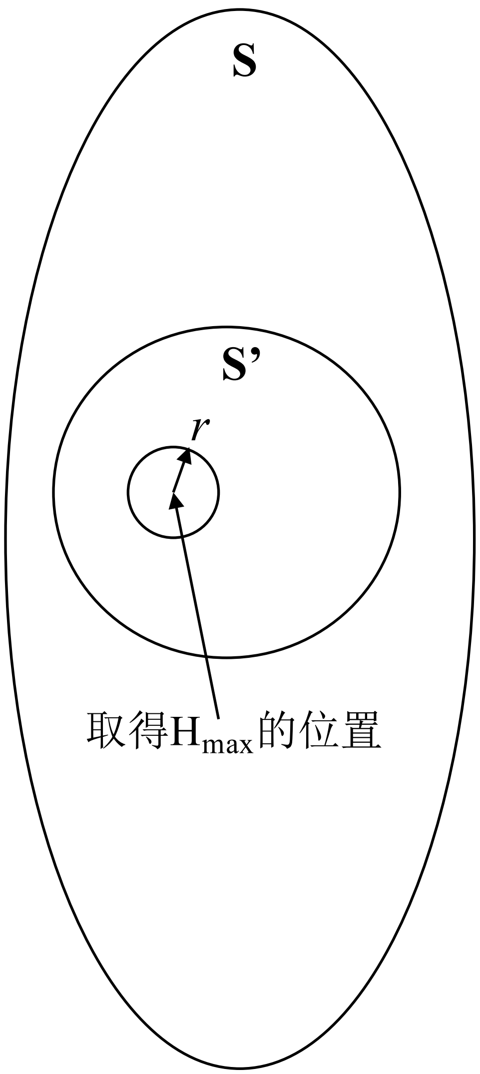
设集合$S_R = \{P : \|H_{\max} - H(P)\|^2 \le aR^2\}$

集合S_R中的分布的熵与最大熵的距离小于$\Delta H = H_{\max} - H(P) = aR^2$，而根据熵函数的连续性，这些分布与最大熵分布也相差不多。

则$$\frac{W(H)}{W(H_{\max})} \cong \exp[N(H - H_{\max})] = \exp(-NaR^2)$$

所以，在半径为R的球中对应的序列数目在K^N种可能序列中所占的比例F_R为。

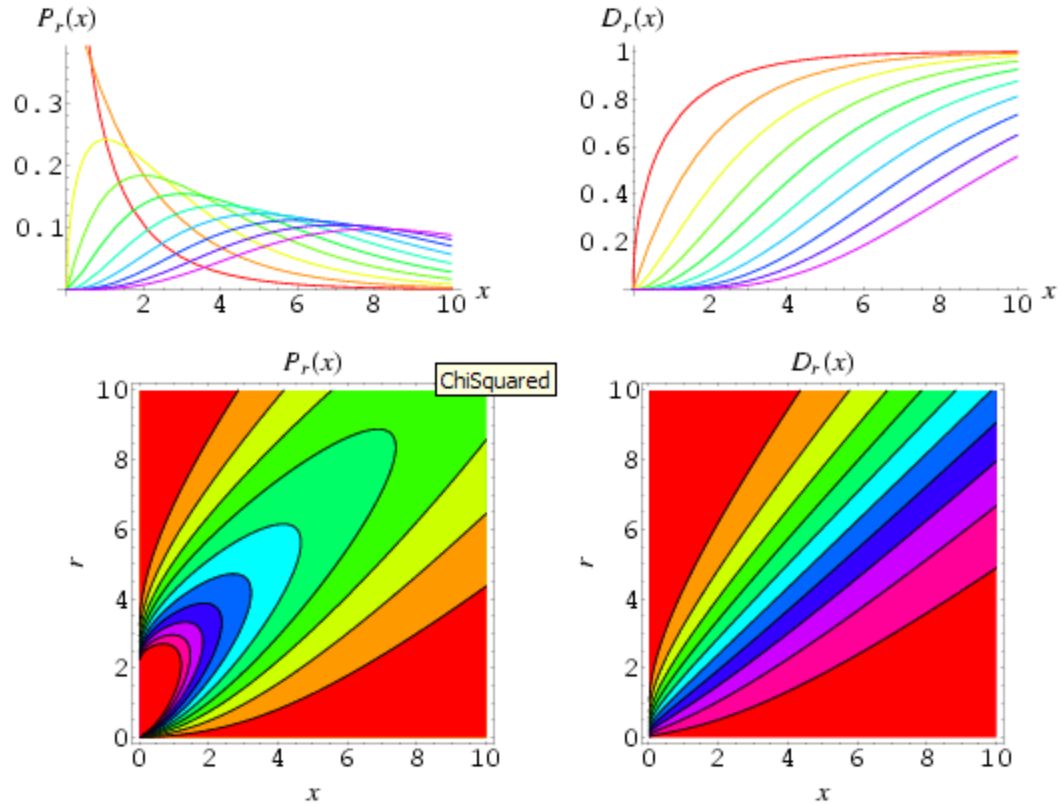
$$F_R \propto \frac{\int_0^R e^{-Nar^2} r^{L-1} dr}{K^N}$$



Jaynes对最大熵原理的解释（续）

$\int_0^R e^{-Nar^2} r^{L-1} dr$为自由度为$L$的$\chi^2$分布的分布函数。

$$2N\Delta H \cong \chi_L^{\ 2}\left(1 - F_R\right)$$



Chi方分布

最大熵分布的例子

例6. 1： 投骰子试验。

投1000次，已知平均点数为：$\sum_{k=1}^{6} kf_k = 4.5.$

其最大熵分布为：$(f_1, f_2, \ldots, f_6) =$（$0.0543, 0.0788, 0.1142, 0.1654, 0.2398.0.3475$）

此时$H_{\max} = 1.61358$

则按$\chi_L^{\ 2}$分布可得：$2N\Delta H \cong \chi_L^{\ 2}(1-F_R)$　　$\Delta H = \dfrac{\chi_L^{\ 2}(1-F_R)}{2N}$

F_R	$\chi_L^{\ 2}(1-F_R)$	ΔH	$H_{\max} - \Delta H \le H \le H_{\max}$
0.95	9.488	0.004744	$1.609 \le H \le 1.61358$
0.99	13.277	0.0066385	$1.602 \le H \le 1.61358$

6 Application of ME Principle

- **Under the following constraints, solve ME p.m.f.**
  - **S**=[a,b]
  - **S**=[0, ∞)， EX＝μ
  - **S**=(－∞,∞) ， EX＝μ
  - **S**=(－∞,∞) ， EX＝α_1， EX^2＝α_2

Spectrum Estimation

- Zero-mean stationary stochastic process $\{X_i\}$
- Autocorrelation $R(k)=EX_iX_{i+k}$
- Fourier transformation of autocorrelation function is the spectrum density of the process

$$S(\lambda) = \sum_{m=-\infty}^{\infty} R(m)e^{-im\lambda}, -\pi < \lambda < \pi$$

- Typically, we have finite length sample trajectory of the process to estimate the spectrum

$$\hat{R}(k) = \frac{1}{n-k}\sum_{i=1}^{n-k} X_i X_{i+k}$$

- The problems are
  - Bigger k, R(k) could not be estimated reliably
  - Small k, coarse grain spectrum estimation

ME Principle applied to spectrum estimation

- J.P.Burg, 1967
- Formulation
  - Given $p+1$ autocorrelation function values
  - Based on $p+1$ constraints, solve the maximal ME stochastic process
- The solution is a Gaussian-Markov process

$$X_i = -\sum_{k=1}^{p} \alpha_k X_{i-k} + Z_i$$

whose parameters are given by Yule-Walker Equations

$$r_0 = -\sum_{k=1}^{p} \alpha_k r_{-k} + \sigma^2$$

$$r_l = -\sum_{k=1}^{p} \alpha_k r_{l-k}, l = 1, 2, \ldots, p$$

$$\Rightarrow \quad S(l) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} \alpha_k e^{-ikl}\right|^2}$$

Yule-Walker Equations

Burg最大熵率定理

Theorem 5.3: 设有随机过程$\{X_i\}$满足约束条件：

$$EX_iX_{i+k} = r_k, k = 0,1,\ldots,p,\ \text{对所有}i$$

使之满足最大熵率的随机过程为具有以下形式的p阶高斯马尔可夫过程：

$$X_i = -\sum_{k=1}^{p} \alpha_k X_{i-k} + Z_i$$

式中，Z_i是独立同分布的高斯随机变量$N(0,\sigma^2)$，α_i称为自回归系数，由$p+1$个约束确定。

证明本定理的诀窍：

证明随机过程的有限长样本序列的熵受限于与其具有同样协方差矩阵的高斯过程，而后者受限与高斯马尔可夫过程的熵率。

定理6.3的证明概略

- 设$X_1,X_2,\ldots,X_n$是任意满足约束$EX_iX_{i+k}=r_k$的随机过程，设$Z_1,Z_2,\ldots Z_n$是具有与$X_1,X_2,\ldots X_n$相同协方差矩阵的高斯过程。

$$h(X_1,X_2,\ldots,X_n)\le \quad h(Z_1,Z_2,\ldots,Z_n)=h(Z_1,\ldots,Z_p)+\sum_{i=p+1}^{n}h(Z_i\mid Z_{i-1},Z_{i-2},\ldots,Z_1)$$

$$\le h(Z_1,\ldots,Z_p)+\sum_{i=p+1}^{n}h(Z_i\mid Z_{i-1},Z_{i-2},\ldots,Z_{i-p})$$

- 定义$Z_1',Z_2',\ldots,Z_n'$为p阶高斯马尔可夫过程，使之具有与$Z_1,Z_2,\ldots,Z_n$相同的1,2,…,p阶分布。于是

$$h(X_1,X_2,\ldots,X_n)\le h(Z_1,\ldots,Z_p)+\sum_{i=p+1}^{n}h(Z_i\mid Z_{i-1},Z_{i-2},\ldots,Z_{i-p})$$

$$=h(Z'_1,\ldots,Z'_p)+\sum_{i=p+1}^{n}h(Z'_i\mid Z'_{i-1},Z'_{i-2},\ldots,Z'_{i-p})=h(Z'_1,\ldots,Z'_n)$$

Burg定理的应用

- 求解**定理6. 3**中的$p+1$个约束方程是关键

$$r_0 = -\sum_{k=1}^{p} \alpha_k r_{-k} + \sigma^2$$

$$r_l = -\sum_{k=1}^{p} \alpha_k r_{l-k}, l = 1, 2, \ldots, p$$

- 上述方程称为Yule-Walker方程组，其解的形式为

$$S(l) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} \alpha_k e^{-ikl}\right|^2}$$

- 在实际的问题中，获得长度为n的样本序列后，计算p个自相关，外推到最大熵分布。

Conclusion

- Direct and reverse problem
- Under-deterministic problem
- ME and Min KL Divergence Princeiples
- Intuition to the ME Principle
- Application of ME in Spectrum Estimation