

The Structure Function and Distinguishable Models of Data

J. Rissanen
University of London, Royal Holloway, CLRC, UK
Helsinki Institute for Information Technology
and
Technical Universities of Tampere and Helsinki

2/3/2006

Models

data:

$$x^n = x_1, \dots, x_n \quad \text{or} \quad (y^n, x^n) = (y_1, x_1), \dots, (y_n, x_n)$$

class of models (structure γ):

$$\mathcal{M}_\gamma = \{p(x^n; \theta, \gamma) : \theta \in \Omega \subseteq R^k, \gamma \in \Gamma\}$$

Example: Normal family of density functions for curve fitting; mean $\mu(\beta) = \hat{x}_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \dots$, and variance σ^2 . Par's $\theta = \beta, \sigma^2$; structure = subset of indices $\{\beta_i : i \in \gamma\}$ for parameters

model:

- finitely describable distribution to be fitted to data
- traditional ‘nonparametric’ models often idealized targets, which cannot be fitted to data; class \mathcal{M} includes ‘nonparametric’ models like histograms
- name ‘class’ more appropriate than ‘likelihood’ function
- **NO** assumptions made about data generating mechanism; no model ‘true’ or ‘right’ nor ‘false’ or ‘wrong’ not even approximation; just models with varying performance
- fundamental difference from traditional statistics!

Modeling Problem: fit parameters θ and their number k (more generally *structure* γ) to data

Main idea: Find parameters such that **all** information from data extracted with given model class

Needed to quantify and measure *intuitive* notions:

- complexity
- information
- noise

all as shortest code length. Want to be able to say ” *When using this model the data have x bits of complexity, y bits of information, leaving z bits as unexplained noise*”

Justification: For shortest code length must take advantage of all regular features the models permit \Rightarrow best model is the one with which shortest code length for data and the model achieved

Two formalizations:

1. Kolmogorov’s structure function in algorithmic complexity theory
2. A suitable generalization of coding theory

Kolmogorov Structure Function

(original unpublished; I learned it from Vereshagin and Vitanyi)

Kolmogorov-complexity $K(x)$ = length of shortest (self-limiting) program in a universal language to generate $x = x^n$ (program = codeword of x in a 'prefix' code)

Model of data x : Finite set $S \ni x$ (properties of x)

Intuition:

- all strings in S share a common property
- size $|S|$ inverse measure of amount of properties extracted from string with S :
- $|S|$ large \Leftrightarrow few properties (restrictions) of x extracted
- $S = \{x\}$ ($|S| = 1$) $\Leftrightarrow S$ captures all conceivable properties of x

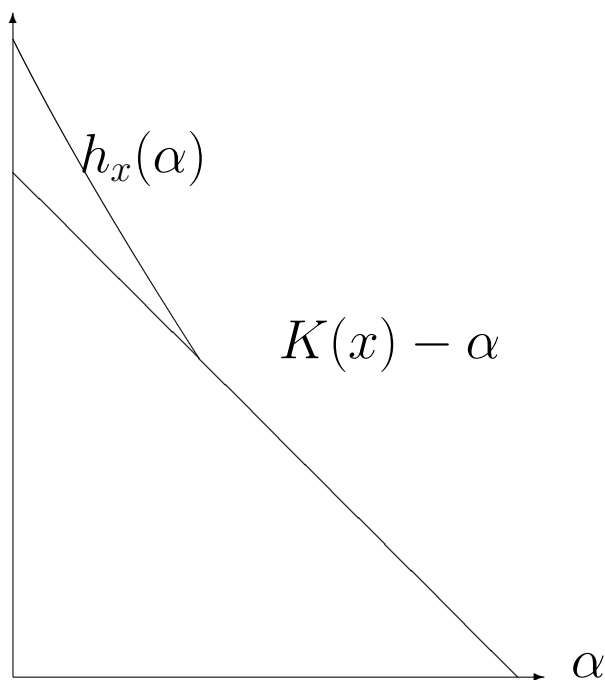
Note: pointless to claim that one model (set $S \ni x$) 'true' and others 'false'! However, can define *optimal* model

Idea: Want the shortest code length of whatever remains when properties of maximum amount α extracted from x

$$h_x(\alpha) = \min_{S \ni x} \{\log |S| : K(S) \leq \alpha\}$$

- Find smallest set S_α on level α that includes x : all properties that can be extracted from x with code length needed to describe S not exceeding α
- $\log |S| \doteq \max_{y \in S} K(y|S)$ (\doteq, \gtrsim) mean (equality, inequality) up to a constant not dependent on length of y)
- $\alpha > \alpha' \Rightarrow h_x(\alpha) \dot{<} h_x(\alpha')$
- *structure line*: $K(x) - \alpha$, (least possible amount of unexplained 'noise' on level α)
- Two-part code length (on level α)

$$L(x, \alpha) = L(x|S_\alpha) + L(S_\alpha) = h_x(\alpha) + \alpha$$



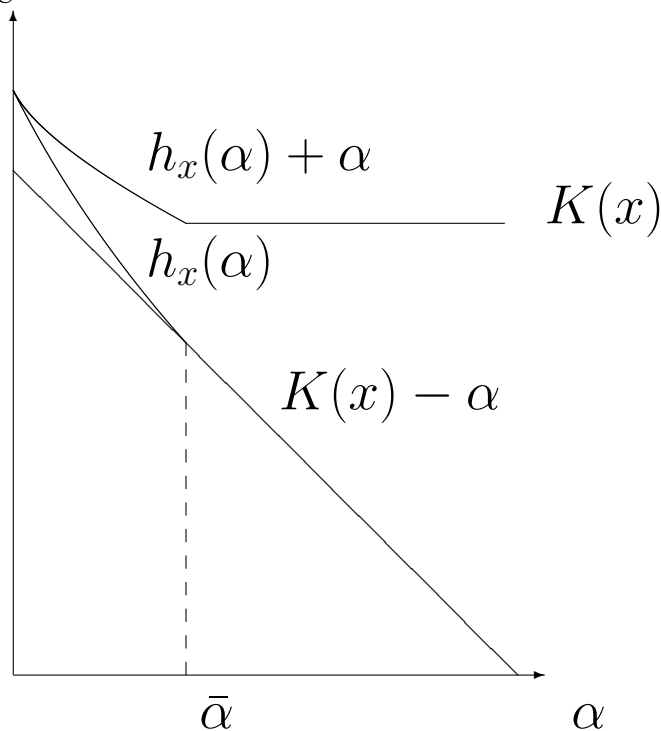
Optimal model

$$h_x(\bar{\alpha}) + \bar{\alpha}$$

where $\bar{\alpha}$ is optimal level:

$$\min\{\alpha : h_x(\alpha) + \alpha = K(x)\}$$

- *MDL* principle for 2-part code!
- $S_{\bar{\alpha}}$ represents all learnable properties of x that can be captured by finite sets
- leaving $h_x(\bar{\alpha})$ as the code length for noninformative ‘noise’



for $\alpha < \bar{\alpha}$
 some properties
 captured leaving
 a lot as noise

for $\alpha \geq \bar{\alpha}$
 all properties and even
 some noise modeled

Coding Prelude

A code is a one-to-one map $C : A \rightarrow B^*$,

- *alphabet* $A = \{a_i\}$ is a finite set of *symbols* of any kind, and $B^* = \{0, 1\}^*$, set of finite binary strings
- extend C to sequences s by concatenation: $C(sa_i) = C(s)C(a_i)$
- C is a *prefix* code, if no *codeword* $C(a_i)$ is a prefix of another.

Codeword lengths $|C(a_i)|$ of prefix codes satisfy **Kraft** inequality

$$\sum_i 2^{-|C(a_i)|} \leq 1,$$

which gives a probability distribution

- $P(a_i) = 2^{-|C(a_i)|} / \sum_{a \in A} 2^{-|C(a)|}$
- Conversely, for any P on A the numbers $\lceil \log 1/P(a_i) \rceil$ define a prefix code (non-unique).
- For large alphabets, like data strings, regard $\log 1/P(a_i)$ as *ideal* codeword lengths

- **Conclusion**

$$\mathbf{C} \Leftrightarrow \mathbf{P}.$$

- same with *arithmetic codes* by their construction

Theorem 1 (McMillan, Doob) *Let P be a distribution on A . The solution to*

$$\min_Q E_P \log \frac{P(X)}{Q(X)}$$

is given by $Q(x) = P(x)$, all x . Hence for all prefix codes C the mean code lengths $|C(x)|$ satisfies

$$\sum_x P(x)|C(x)| \geq \sum_x P(x) \log 1/P(x) = H(P) \text{ entropy.}$$

Equality holds iff $|C(x)| = \log 1/P(x)$, all x .

Conclusions:

- Optimal code design when P is given
- Best prefix code mimics data generating distribution
- Meaning of entropy: tight lower bound of all prefix codes
- Important corollary: *Kullback-Leibler* distance

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Notes:

- Entropy can also be defined as the *per symbol limit of the logarithm of number of ‘typical’ strings generated by P*

- What to do when distribution $f(x^n)$ over alphabet $\{x^n\}$ not given?
- Instead, model classes $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$,

$$\mathcal{M}_\gamma = \{f(y^n; \theta) : \theta \in \Omega^\gamma\}$$
 selected; $\theta = \theta_1, \dots, \theta_k$ in structure γ
- First problem: How to estimate γ or just k ?
- Cannot do it by $\max_k f(x^n; \hat{\theta}(x^n))$
- What play the roles of shortest code length $\log 1/f(x^n)$ and entropy $H(f)$ now?

Generalization of McMillan-Doob Theorem

Idea: Represent each model class \mathcal{M}_k by a *universal* model $\hat{f}(x^n; \mathcal{M}_\gamma)$

- For each γ the role of shortest code length is played by

$$\log 1/\hat{f}(x^n; \mathcal{M}_\gamma)$$

for the *best* universal model

- Optimality of entropy $H(f)$ in previous coding theorem gets generalized to inequality

$$\begin{aligned} E_\theta \log 1/q(x^n) &\geq E_\theta \log 1/\hat{f}(x^n; \mathcal{M}_\gamma) \cong \\ &\cong E_\theta \log 1/f(x^n; \theta) + \frac{k}{2} \log n + \dots \end{aligned}$$

for all q and all θ except in a set whose size vanishes as n grows.

Two Universal Models

1. Mixture Model

$$f_w(x^n; \mathcal{M}_\gamma) = \int f(x^n; \theta) w(\theta) d\theta$$

With special prior w , $\hat{q} = f_w$ solves the minmax problem

$$\min_q \max_\theta D(f(X^n; \theta) \| q(X^n))$$

Asymptotically, for models satisfying certain smoothness conditions

$$E_\theta \log \frac{f(X^n; \theta)}{f_w(X^n)} \cong \frac{k}{2} \log \frac{n}{2\pi} + \log \int_\Omega |J(\theta)|^{1/2} d\theta$$

Fisher information matrix

$$J(\theta) = \lim n^{-1} E_\theta \left\{ \frac{\partial^2 \log 1/f(y^n; \theta)}{\partial \theta_i \partial \theta_j} \right\}$$

The excess over the entropy $E_\theta \log 1/f(X^n; \theta)$ called *regret*,
(**bad** name!)

2. Normalized ML Model

Normalized ML universal model

$$\hat{f}(x^n; \mathcal{M}_\gamma) = \frac{f(x^n; \hat{\theta}(x^n))}{C_{\gamma,n}}$$

$$C_{\gamma,n} = \int_{y^n: \hat{\theta}(y^n) \in \Omega^\circ} f(y^n; \hat{\theta}(y^n)) dy^n$$

Since

$$f(x^n; \theta) = f(x^n, \hat{\theta}(x^n); \theta) = f(x^n | \hat{\theta}(x^n); \theta) g(\hat{\theta}(x^n); \theta)$$

and

$$g(\hat{\theta}(x^n); \theta) = \int_{y^n: \hat{\theta}(y^n) = \hat{\theta}(x^n)} f(y^n; \theta) dy^n$$

we get

$$C_{\gamma,n} = \int_{\hat{\theta} \in \Omega^\circ} g(\hat{\theta}; \hat{\theta}) d\hat{\theta}$$

Ω° interior of Ω^γ , taken here as a compact set; $d\theta$ and dy^n differential volumes

Properties

1. $\hat{f}(x^n; \mathcal{M}_\gamma)$ is the unique solution $\hat{g} = \hat{q}$ to the minmax problem (Shtarkov)

$$\min_q \max_{x^n} \log \frac{f(x^n; \hat{\theta}(x^n))}{q(x^n)} = \log C_{\gamma,n},$$

2. as well as to the maxmin problem

$$\max_g \min_q E_g \log \frac{f(X^n; \hat{\theta}(X^n))}{q(X^n)} = \log C_{\gamma,n}.$$

3. When g restricted to \mathcal{M}_γ , the maxmin and minmax value $\log C_{\gamma,n}$ cannot be beaten except for g in a set whose volume $\rightarrow 0$ as $n \rightarrow \infty$

Because of these results define **stochastic complexity** of x^n , given \mathcal{M}_γ :

$$-\log \hat{f}(x^n; \mathcal{M}_\gamma) = -\log f(x^n; \hat{\theta}(x^n)) + \log C_{k,n}$$

With Central Limit Theorem *stochastic complexity* has the form

$$\begin{aligned} \log 1/\hat{f}(x^n; \mathcal{M}_\gamma) &= \log 1/f(x^n; \hat{\theta}(x^n)) + \\ &+ \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Omega} |J(\theta)|^{1/2} d\theta + o(1) \end{aligned}$$

Estimation of γ by MDL Principle

For model classes $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$, construct ‘prior’ $\mu(\gamma)$ from Γ , and joint

$$F_\mu(x^n, \gamma) = \hat{f}(x^n; \mathcal{M}_\gamma)\mu(\gamma)$$

such that (MDL principle):

- $\min_{\mu, \gamma} \{\log 1/\hat{f}(x^n; \mathcal{M}_\gamma) + \log 1/\mu(\gamma)\}$
- usually not many μ 's available and $L(\mu)$ short; can be ignored

A deeper theory required to represent the optimal model in the optimal class $\mathcal{M}_{\hat{\gamma}}$ because parameter space $\Omega^{\hat{\gamma}}$ not countable. (The code length for properties (model) of data should not exceed that of the data!)

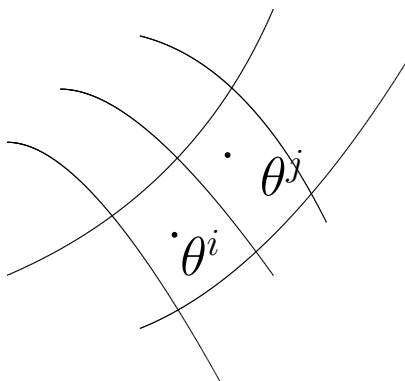
Structure function

Analogies with algorithmic notions:

- Set of programs replaced by model class \mathcal{M}_γ (taken as fixed)
- Set A replaced by quantized model $f(x^n; \theta^i)$
- Kolmogorov complexity replaced by stochastic complexity
- $K(A)$ replaced by code length $L(\theta^i)$
- $\log |A|$, max code length of $y \in A$, replaced by the maximum or mean code length of typical strings of $f(x^n; \theta^i)$

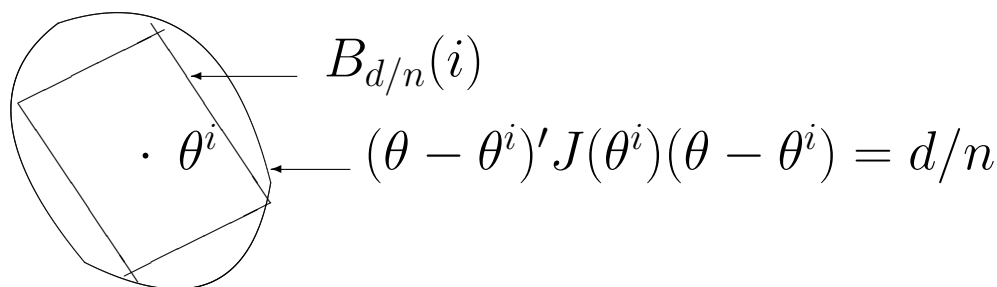
Quantization of parameters

- Partition compact parameter space Ω by curvilinear rectangles $B_{d/n}(i)$ such that $D(f(X^n; \theta^i) \| f(X^n; \theta^j))$ between adjacent models constant



- Achieved (in the limit) with $B_{d/n}(i)$ as maximal rectangle within hyper ellipsoid $(\theta - \theta^i)' J(\theta^i) (\theta - \theta^i) = d/n$
- $J(\theta)$ Fisher information matrix satisfying

$$0 < J(\theta) = \lim n^{-1} E \left\{ \frac{\partial^2 \log 1/f(y^n; \theta)}{\partial \theta_i \partial \theta_j} \right\} < \infty$$



- volume of $B_{d/n}(i)$

$$V = \left(\frac{4d}{kn}\right)^{k/2} |J(\theta^i)|^{-1/2}.$$

- number of rectangles in compact space $|\Omega|/V = O(n^{k/2})$; more accurately as follows
- define with $g(\hat{\theta}; \theta)$, d.f. for statistic $\hat{\theta}$,

$$Q_{d/n}(i) = \int_{\hat{\theta} \in B_{d/n}(i)} g(\hat{\theta}; \theta) d\hat{\theta} \rightarrow \left(\frac{2d}{\pi k}\right)^{k/2}$$

- nearly uniform prior $W(\theta^i) = Q_{d/n}(i)/C_{k,n}$; number of rectangles $1/W(\theta^i)$
- no need to construct partition; enough to estimate rectangle $B_{d/n}(i)$ that includes $\hat{\theta}(x^n)$ with its code length

$$L_d(\theta^i) = \ln 1/W(\theta^i) \cong \frac{k}{2} \ln \frac{\pi k}{2d} + \ln C_{k,n}.$$

Two structure functions

$$h_{x^n}^1(\alpha) = \min_d \{ \ln 1/f(x^n; \hat{\theta}) + d/2 : L_d(\theta^i) \leq \alpha \}$$

$$h_{x^n}^2(\alpha) = \min_d \left\{ \frac{1}{|B_{d/n}(i)|} \int_{\hat{\theta}(y^n) \in B_{d/n}(i)} \ln 1/f(y^n; \theta^i) dy^n : L_d(\theta^i) \leq \alpha \right\}$$

Can show

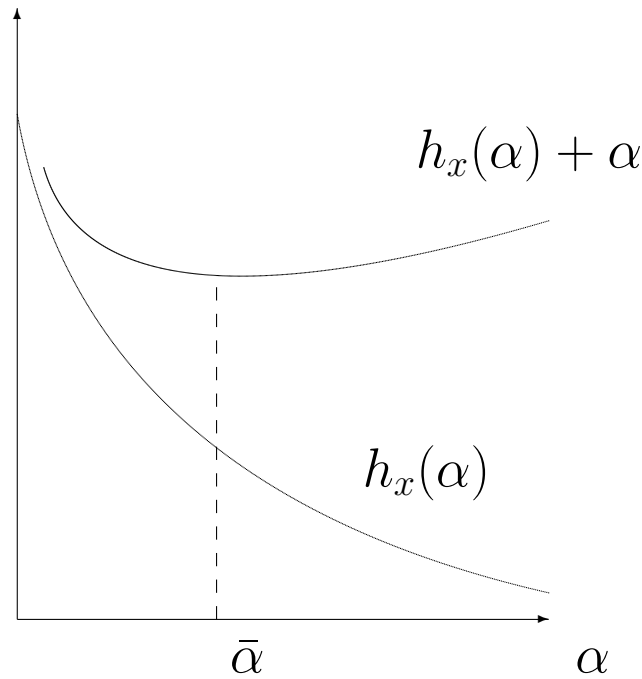
$$h_{x^n}^2(\alpha) = \min_d \{ \ln 1/f(x^n; \hat{\theta}) + d/6 : L_d(\theta^i) \leq \alpha \}$$

Difference between the two: code length for maximum typical strings

$$\ln 1/f(x^n; \hat{\theta}) + d/2$$

and average code length of typical strings

$$\ln 1/f(x^n; \hat{\theta}) + d/6$$



for $\alpha < \bar{\alpha}$
 some properties
 captured leaving
 a lot as noise

for $\alpha \geq \bar{\alpha}$
 all properties and even
 some noise modeled

With second structure function ($\hat{d} = 3k$)

- The amount of learnable information in data

$$\bar{\alpha} = \ln C_{k,n} + (k/2) \ln \pi/6$$

- The amount of unexplained noise

$$h_{x^n}^2(\bar{\alpha}) = -\ln f(x^n; \hat{\theta}(x^n)) + \frac{k}{2}$$

- The complexity of data (stochastic complexity)

$$-\ln \hat{f}(x^n; \mathcal{M}_k) = -\ln f(x^n; \hat{\theta}(x^n)) + \ln C_{k,n}$$

Distinguishable distributions

Balasubramanian:

- The normalizing coefficient $C_{k,n}$ in the *NML* model is the number of ‘distinguishable’ models $f_i = f(x^n; \theta^i)$

An intriguing result - abstract

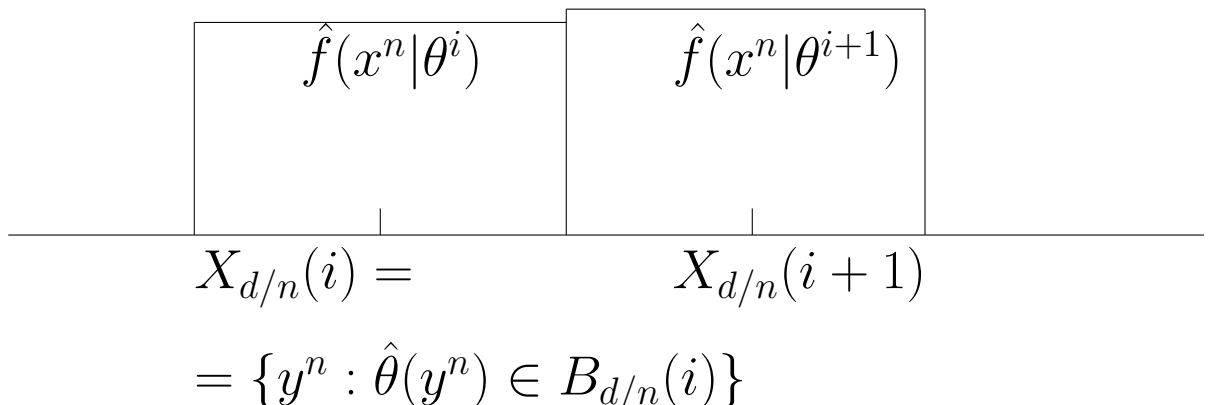
No partition of parameter space

Support of all models the entire space; perfect distinguishability for f_i not possible

Perfectly distinguishable (but impractical) models:

$$\hat{f}(y^n | \theta^i) = \begin{cases} f(y^n; \hat{\theta}(y^n)) / Q_{d/n}(i) & \text{if } \hat{\theta}(y^n) \in B_{d/n}(i) \\ 0 & \text{otherwise} \end{cases}$$

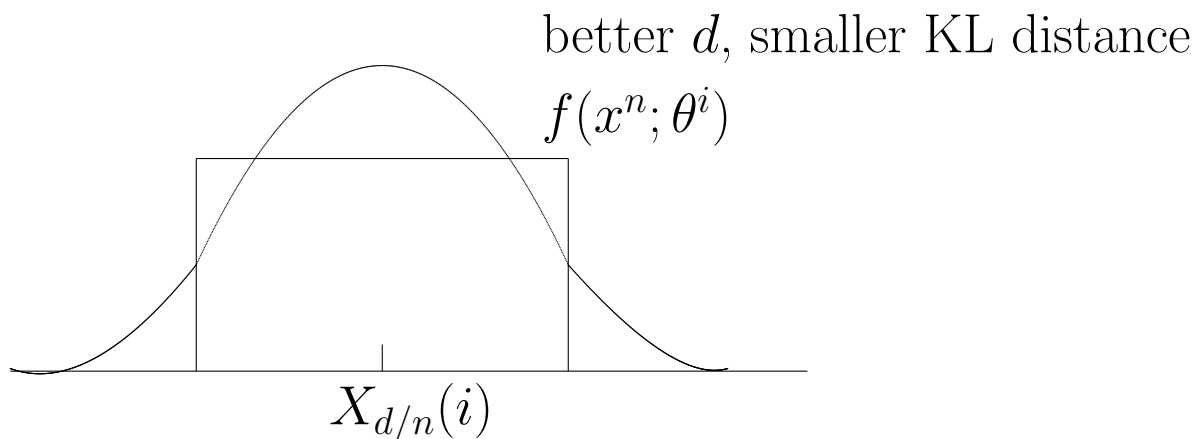
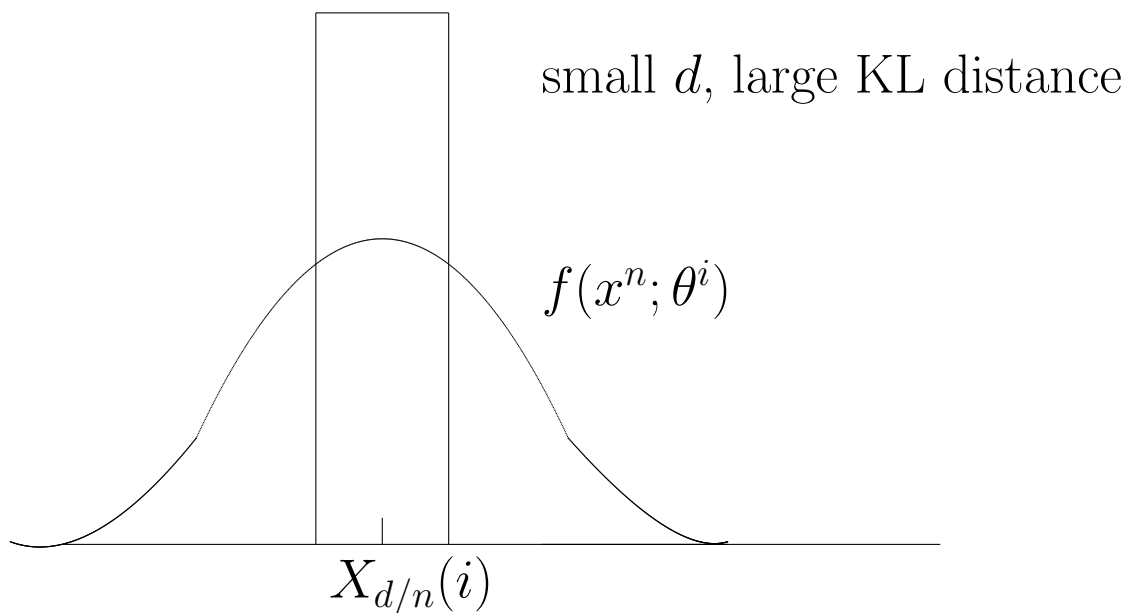
$$Q_{d/n}(i) = \int_{\hat{\theta} \in B_{d/n}(i)} g(\hat{\theta}; \hat{\theta}) d\hat{\theta} \rightarrow \left(\frac{2d}{\pi k} \right)^{k/2} .$$



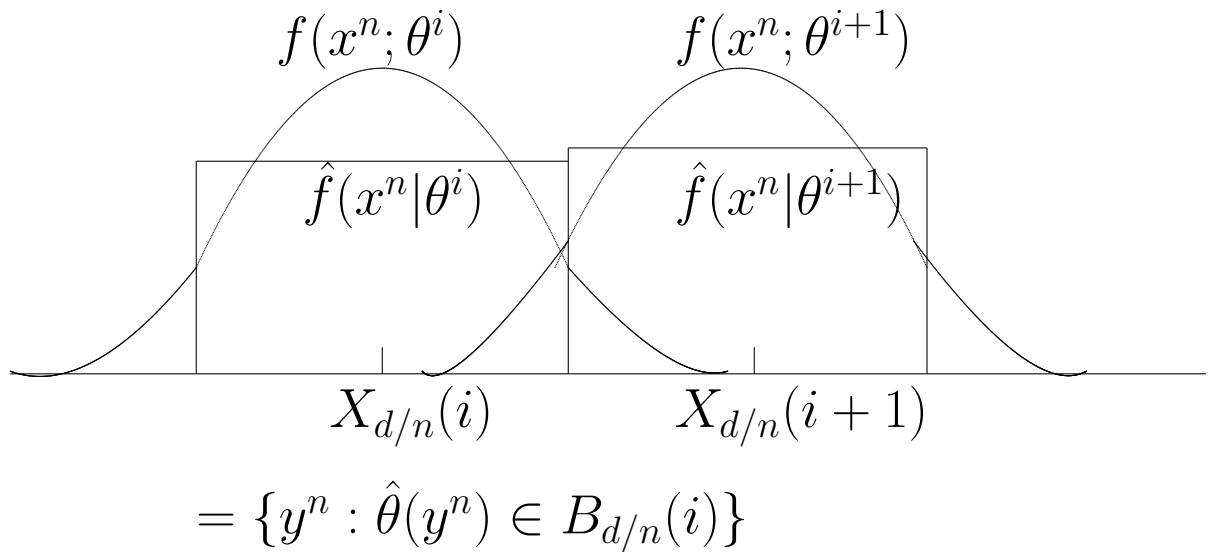
Find models $f(x^n; \theta^i)$ closest to the perfectly distinguishable ones

$$\min_d D(\hat{f}(X^n | \hat{\theta}(X^n)) \| f(X^n; \theta^i)).$$

- The minimizing value \hat{d}_n converges to $3k$
- same as with structure function $h_{x^n}^2(\alpha)$!
- note: curves only in support $X_{d/n}(i)$ count!



Optimally distinguishable models



- Can compute probabilities $P(X_{d/n}(i))$ under all $\hat{f}(x^n | \theta^j)$
- Needed for confidence assessment in hypothesis testing

Hypothesis testing

- Two main problems with Neyman-Pearson theory:
 1. hypotheses tested not models, hence the only uncertainty due to sampling
 2. no way to assess confidence, because don't know which hypothesis opposing the null hypothesis to pick

Testing in theory of distinguishable models:

Pick the null hypothesis as the center of one of the equivalence classes for the optimal $\hat{d} = 3k$, say θ^i .

The opposing composite hypothesis $\{f(x^n; \theta^j) : \theta^j \neq \theta^i\}$.

Test: Accept null hypothesis if $\hat{\theta}(x^n) \in B_{\hat{d}/n}(i)$; else reject it.

Confidence in test:

$$\frac{P(B_{\hat{d}/n}(j))}{1 - P(B_{\hat{d}/n}(i))}$$

where j such that $\hat{\theta}(x^n) \in B_{\hat{d}/n}(j)$.

(Confidence in accepting null hypothesis if $j = i$; else in rejecting it.)

Optimal balance between

- uncertainty due to sampling

$$\frac{1}{|B_{d/n}(i)|} \int_{\hat{\theta} \in B_{d/n}(i)} (\hat{\theta} - \theta^i)' J(\theta^i) (\hat{\theta} - \theta^i) dy^n \rightarrow d/6$$

(grows with increasing $|B_{d/n}(i)|$)

- uncertainty due to model fitting (grows with decreasing $|B_{d/n}(i)|$ because of increasing number of models)

$$L_d(\theta^i) = \ln(C_{k,n}/Q_{d/n}(i)).$$

- The confidence increases rapidly with the increasing distance from the null hypothesis; the adjacent models to it are hardest to distinguish.
- Don't worry about other than the adjacent hypotheses

Example: Bernoulli class.

Null hypothesis: $P(x = 0) = p = i/n; \hat{d} = 3.$

$$|B_{3/n}(i/n)| = \left(\frac{12}{n}\right)^{1/2}((i/n)(1 - i/n))^{1/2} \quad (1)$$

Its probability

$$P(B_{3/n}(i/n)) = 2\phi(\sqrt{3}(i/n)^{1/2}(1 - i/n)^{1/2}) - 1.$$

For $i/n = 1/3$ the probability is about 0.6. One half of the width of the interval is about 0.82, a little smaller than standard deviation 1.

If null hypothesis accepted, confidence is $\frac{P(B_{3/n}(i/n))}{2[1 - P(B_{3/n}(i/n))]}$, and if $\hat{f}(x^n)$ falls in one of the two adjacent intervals, the confidence in rejecting is about twice as great.