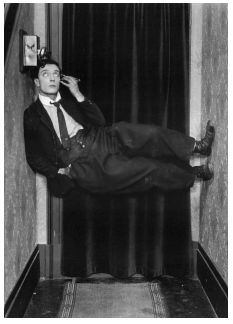
Cross-validation and early stopping: a Bayesian point of view?

Luca Porta Mana

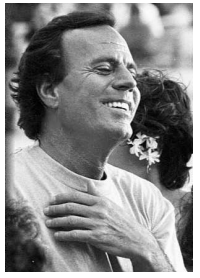
Roudi group SpinOr, Kavli Institute, Trondheim

In collaboration with

Saitama



Buster



Julio

The probability calculus

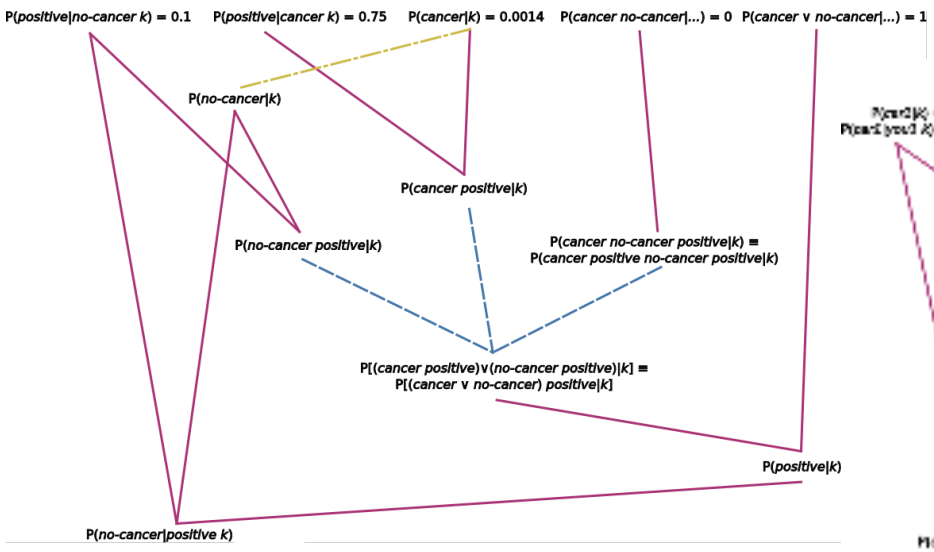
$$\mathrm{P}(\neg A \mid I) = 1 - \mathrm{P}(A \mid I)$$

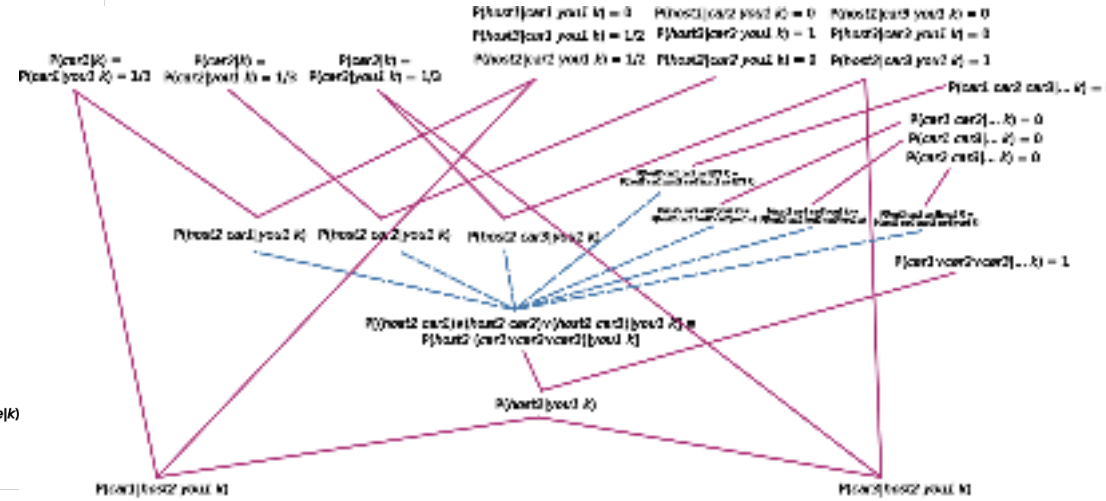
$$\mathrm{P}(A\ B \mid I) = \mathrm{P}(A \mid B\ I)\,\mathrm{P}(B \mid I)$$

$$\mathrm{P}(A \vee B \mid I) = \mathrm{P}(A \mid I) + \mathrm{P}(B \mid I) - \mathrm{P}(A\ B \mid I)$$

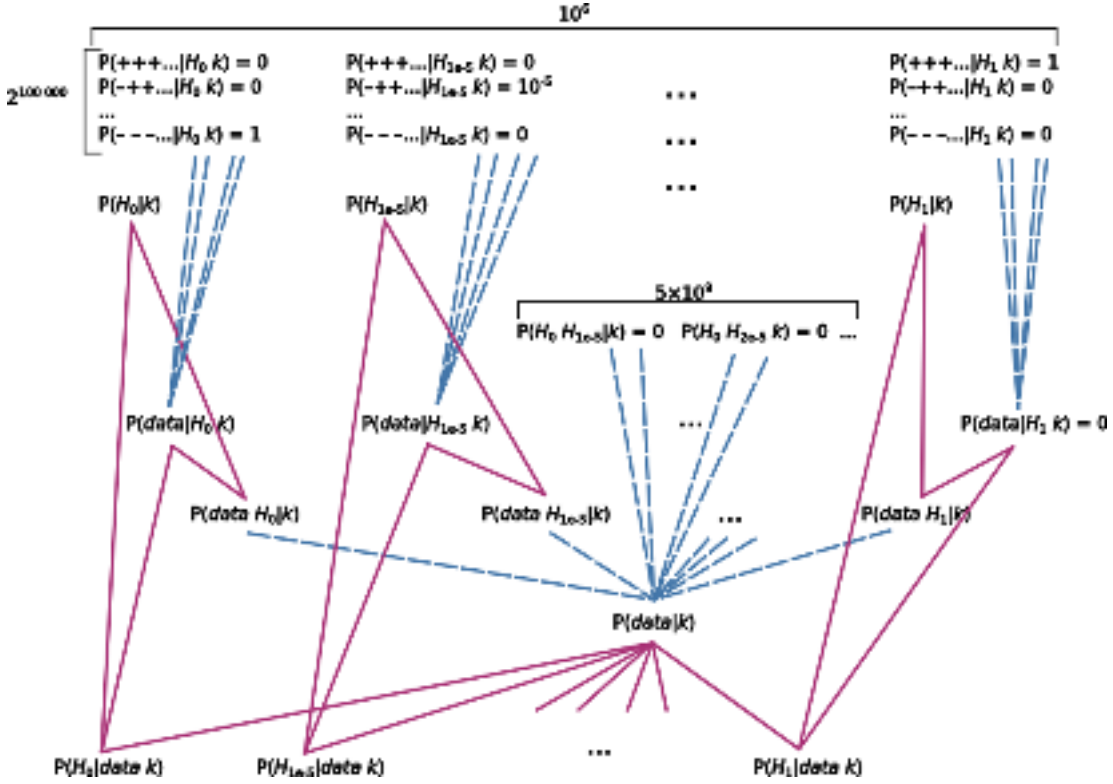
> They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of Logic is conversant at present only with things either certain, impossible, or *entirely* doubtful, none of which (fortunately) we have to reason on. Therefore the true Logic for this world is the Calculus of Probabilities, which takes account of the magnitude of the probability (which is, or which ought to be in a reasonable man's mind).

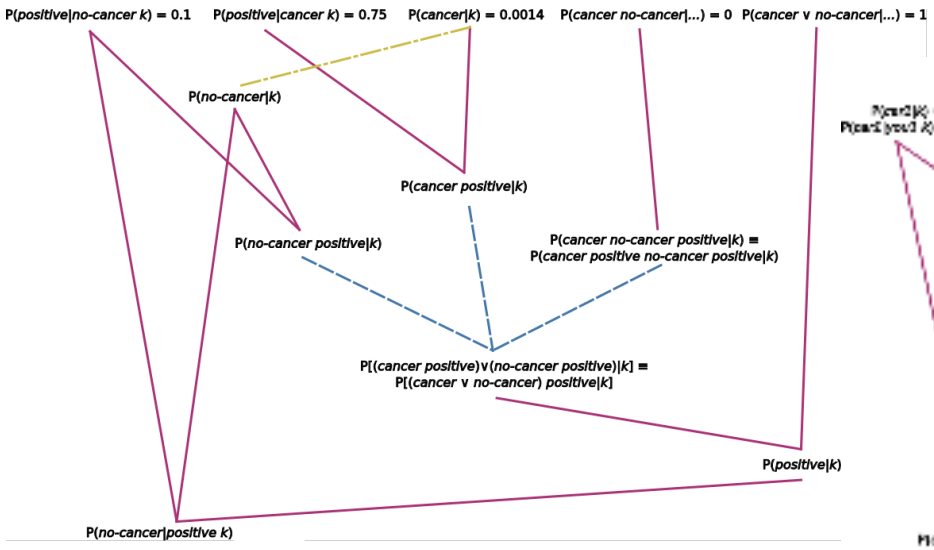
(J. C. Maxwell, 1850)

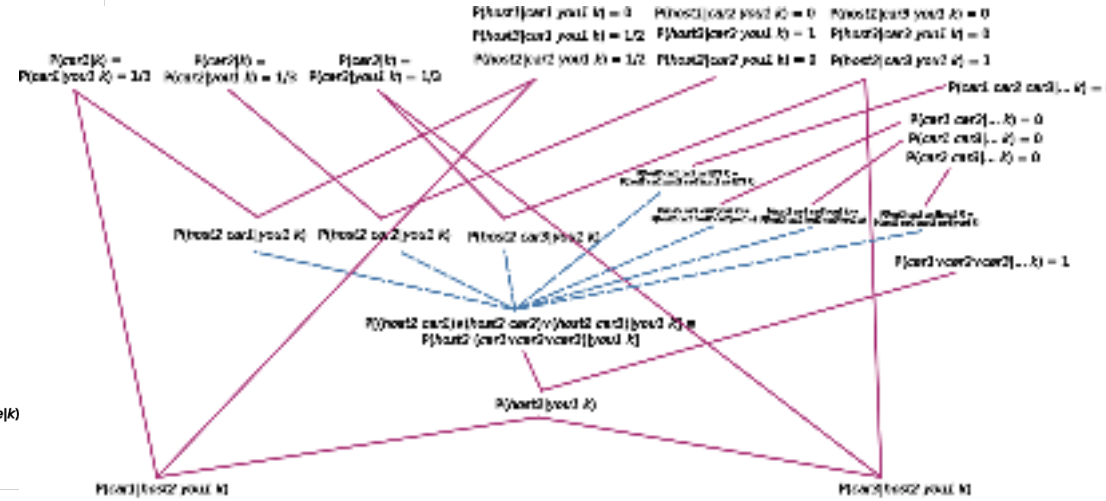

P(positive|no-cancer k) = 0.1
P(positive|cancer k) = 0.75
P(cancer|k) = 0.0014
P(cancer no-cancer|...) = 0
P(cancer v no-cancer|...) = 1
P(no-cancer|k)
P(cancer positive|k)
P(no-cancer positive|k)
P(cancer no-cancer positive|k) =
P(cancer positive no-cancer positive|k)
P[(cancer positive)v(no-cancer positive)|k] =
P[(cancer v no-cancer) positive|k]
P(positive|k)
P(no-cancer|positive k)



and
or
not


10^6
2^100 000
P(+++...|H0 k) = 0
P(-++...|H0 k) = 0
P(- - -...|H0 k) = 1
P(+++...|H1e-5 k) = 0
P(-++...|H1e-5 k) = 10^-5
P(- - -...|H1e-5 k) = 0
P(+++...|H1 k) = 1
P(-++...|H1 k) = 0
P(- - -...|H1 k) = 0
P(H0|k)
P(H1e-5|k)
P(H1|k)
5×10^4
P(H0 H1e-5|k) = 0
P(H1 H2e-5 k) = 0
P(data|H0 k)
P(data|H1e-5 k)
P(data|H1 k) = 0
P(data H0|k)
P(data H1e-5|k)
P(data H1|k)
P(data|k)
P(H0|data k)
P(H1e-5|data k)
P(H1|data k)





$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{B \vdash B}(I) \quad \dfrac{}{C \vdash C}(I)}{B \vee C \vdash B, C}(\vee L)}{B \vee C \vdash C, B}(PR)}{B \vee C, \neg C \vdash B}(\neg L) \quad \dfrac{}{\neg A \vdash \neg A}(I)}{(B \vee C), \neg C, (B \to \neg A) \vdash \neg A}(\to L)$$

$$\dfrac{(B \vee C), \neg C, (B \to \neg A) \vdash \neg A}{(B \vee C), \neg C, ((B \to \neg A) \wedge \neg C) \vdash \neg A}(\wedge L_1)$$

$$\dfrac{(B \vee C), \neg C, ((B \to \neg A) \wedge \neg C) \vdash \neg A}{(B \vee C), ((B \to \neg A) \wedge \neg C), \neg C \vdash \neg A}(PL)$$

$$\dfrac{\dfrac{\dfrac{}{A \vdash A}(I)}{\vdash \neg A, A}(\neg R)}{\vdash A, \neg A}(PR) \quad \dfrac{(B \vee C), ((B \to \neg A) \wedge \neg C), ((B \to \neg A) \wedge \neg C) \vdash \neg A}{(B \vee C), ((B \to \neg A) \wedge \neg C) \vdash \neg A}(CL)}{((B \to \neg A) \wedge \neg C), (B \vee C) \vdash \neg A}(PL)$$

$$\dfrac{((B \to \neg A) \wedge \neg C), (B \vee C) \vdash \neg A}{((B \to \neg A) \wedge \neg C), (A \to (B \vee C)) \vdash \neg A, \neg A}(\to L)$$

and
or
not

The probability calculus

$$\mathrm{P}(\neg A \mid I) = 1 - \mathrm{P}(A \mid I)$$

$$\mathrm{P}(A\ B \mid I) = \mathrm{P}(A \mid B\ I)\,\mathrm{P}(B \mid I)$$

$$\mathrm{P}(A \vee B \mid I) = \mathrm{P}(A \mid I) + \mathrm{P}(B \mid I) - \mathrm{P}(A\ B \mid I)$$

+ utility function → decision theory

‘Model averaging’

$$\mathrm{P}(x \mid D\ M_1\ I) \quad \textbf{or} \quad \mathrm{P}(x \mid D\ M_2\ I)\ \textbf{?}$$

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 45–97

Assessment and Propagation of Model Uncertainty

By DAVID DRAPER†

University of Bath, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 16th, 1994,* Professor V. S. Isham *in the Chair*]

SUMMARY

In most examples of inference and prediction, the expression of uncertainty about unknown quantities y on the basis of known quantities x is based on a *model* M that formalizes assumptions about how x and y are related. M will typically have two parts: *structural* assumptions S, such as the form of the link function and the choice of error distribution in a generalized linear model, and *parameters* θ whose meaning is specific to a given choice of S. It is common in statistical theory and practice to acknowledge parametric uncertainty about θ given a particular assumed structure S; it is less common to acknowledge structural uncertainty about S itself. A widely used approach involves enlisting the aid of x to specify a plausible single ‘best’ choice S^* for S, and then proceeding as if S^* were known to be correct. In general this approach fails to assess and propagate structural uncertainty fully

‘Model averaging’

$$\mathrm{P}[x \mid D\ (M_1 \vee M_2)\ I] =$$

$$\mathrm{P}(x \mid D\ M_1\ I) \times \frac{\mathrm{P}(M_1 \mid D\ I)}{\mathrm{P}(M_1|D\ I)+\mathrm{P}(M_2|D\ I)} +$$

$$\mathrm{P}(x \mid D\ M_2\ I) \times \frac{\mathrm{P}(M_2 \mid D\ I)}{\mathrm{P}(M_1|D\ I)+\mathrm{P}(M_2|D\ I)}$$

‘Model averaging’

$$\mathrm{P}[x \mid D\ (M_1 \vee M_2)\ I] =$$

$$\mathrm{P}(x \mid D\ M_1\ I) \times \frac{\mathrm{P}(M_1 \mid D\ I)}{\mathrm{P}(M_1|D\ I)+\mathrm{P}(M_2|D\ I)} +$$

$$\mathrm{P}(x \mid D\ M_2\ I) \times \frac{\mathrm{P}(M_2 \mid D\ I)}{\mathrm{P}(M_1|D\ I)+\mathrm{P}(M_2|D\ I)}$$

Typically the data appear as...

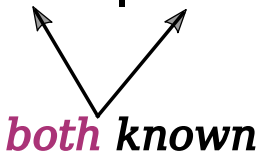
$$\mathrm{P}(M \mid D\ I) \qquad\qquad \mathrm{P}(D \mid M\ I)$$

Cross-validation

$$\mathrm{P}(D' \mid D'' \; M \; I)$$

Cross-validation

$$\mathrm{P}(D' \mid D'' \; M \; I)$$



Early stopping

$$\mathrm{P}(x \mid D' \; D'' \; M \; I)$$

Cross-validation

$$\mathrm{P}(D' \mid D'' \; M \; I)$$

both known

Early stopping

$$\mathrm{P}(x \mid D' \; D'' \; M \; I)$$

Can we make Bayesian sense of these?

Or do they lead to inconsistencies?

What is a ‘model’?

What is a 'model'?

Statements of interest $S_1, S_2, \ldots, S_N$

typically $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$

What is a 'model'?

Statements of interest $S_1, S_2, \ldots, S_N$

typically $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$

Background knowledge I

What is a 'model'?

Statements of interest $S_1, S_2, \ldots, S_N$

typically $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$

Background knowledge I

Model M **:**

owledge/assumptions that determine all wanted probabiliti

$$\mathrm{P}(X_1 = x_1, \ldots, X_N = x_N \mid M\ I)$$

What is a 'model'?

Statements of interest $S_1, S_2, \ldots, S_N$

typically $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$

Background knowledge I

Model M **:**

owledge/assumptions that determine all wanted probabiliti

$$\mathrm{P}(X_1 = x_1, \ldots, X_N = x_N \mid M\ I)$$

('model' ≠ 'hypothesis')

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $x_1, x_2, \ldots$ **very long sequence**

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $x_1, x_2, \ldots$ **very long sequence**

$\boldsymbol{F} := (F_{-1}, F_0, F_1)$ **long-run relative frequencies**

"But this long run *is a misleading guide to current affairs.*
In the long run *we are all dead"* (J. M. Keynes 1923)

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $x_1, x_2, \ldots$ **very long sequence**

$\boldsymbol{F} := (F_{-1}, F_0, F_1)$ **long-run relative frequencies**

"But this long run *is a misleading guide to current affairs.*
In the long run *we are all dead"* (J. M. Keynes 1923)

If we assume all sequences with same long-run frequency equally probable:

$$\mathrm{p}(X_1 = -1 \mid \boldsymbol{F}\ M\ I) = F_{-1}$$

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $x_1, x_2, \ldots$ **very long sequence**

$\boldsymbol{F} := (F_{-1}, F_0, F_1)$ **long-run relative frequencies**

"But this long run *is a misleading guide to current affairs.*
In the long run *we are all dead"* (J. M. Keynes 1923)

If we assume all sequences with same long-run frequency equally probable:

$$p(X_1 = -1 \mid \boldsymbol{F}\ M\ I) = F_{-1} \qquad p(X_1 = -1, X_2 = 0 \mid \boldsymbol{F}\ M\ I) \approx F_{-1}\, F_0$$

The Annals of Probability
1980, Vol. 8, No. 4, 745–764

FINITE EXCHANGEABLE SEQUENCES

BY P. DIACONIS[1] AND D. FREEDMAN[2]

Bell Laboratories and Stanford University, and University of California at Berkeley

Let $X_1, X_2, \cdots, X_k, X_{k+1}, \cdots, X_n$ be exchangeable random variables taking values in the set S. The variation distance between the distribution of $X_1, X_2, \cdots, X_k$ and the closest mixture of independent, identically distributed random variables is shown to be at most $2ck/n$, where c is the cardinality of S. If c is infinite, the bound $k(k-1)/n$ is obtained. These results imply the most general known forms of de Finetti's theorem. Examples are given to show that the rates k/n and $k(k-1)/n$ cannot be improved.

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $x_1, x_2, \ldots$ **very long sequence**

$\boldsymbol{F} := (F_{-1}, F_0, F_1)$ **long-run relative frequencies**

"But this long run *is a misleading guide to current affairs.*
In the long run *we are all dead"* (J. M. Keynes 1923)

If we assume all sequences with same long-run frequency equally probable:

$$\mathrm{p}(X_1 = -1 \mid \boldsymbol{F}\ M\ I) = F_{-1} \qquad \mathrm{p}(X_1 = -1, X_2 = 0 \mid \boldsymbol{F}\ M\ I) \approx F_{-1}\, F_0$$

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $x_1, x_2, \ldots$ **very long sequence**

$\boldsymbol{F} := (F_{-1}, F_0, F_1)$ **long-run relative frequencies**

"But this long run *is a misleading guide to current affairs.*
In the long run *we are all dead"* (J. M. Keynes 1923)

If we assume all sequences with same long-run frequency equally probable:

$$\mathrm{p}(X_1 = -1 \mid \boldsymbol{F}\ M\ I) = F_{-1} \qquad \mathrm{p}(X_1 = -1, X_2 = 0 \mid \boldsymbol{F}\ M\ I) \approx F_{-1}\, F_0$$

$$\mathrm{p}(x_1, \ldots, x_n \mid \boldsymbol{F}\ M\ I) = \prod_{t=1}^{n} F_{x_t}$$

A visual intro to exchangeable models

$x \in \{-1, 0, 1\}$ $\quad x_1, x_2, \ldots$ **very long sequence**

$\boldsymbol{F} := (F_{-1}, F_0, F_1)$ **long-run relative frequencies**

"But this long run *is a misleading guide to current affairs.*
In the long run *we are all dead"* (J. M. Keynes 1923)

If we assume all sequences with same long-run frequency equally probable:

$$\mathrm{p}(X_1 = -1 \mid \boldsymbol{F}\ M\ I) = F_{-1} \qquad \mathrm{p}(X_1 = -1, X_2 = 0 \mid \boldsymbol{F}\ M\ I) \approx F_{-1}\, F_0$$

$$\mathrm{p}(\underbrace{x_1, \ldots, x_n}_{\{-1,\,0,\,1\} \text{ appear with frequencies } \boldsymbol{f} := (f_{-1}, f_0, f_1)} \mid \boldsymbol{F}\ M\ I) = \prod_{t=1}^{n} F_{x_t} = \prod_{x=-1}^{1} (F_x)^{n f_x}$$

$$\mathrm{p}(\underbrace{x_1, \ldots, x_n}_{\{-1,\,0,\,1\}\text{ appear with frequencies } \boldsymbol{f} := (f_{-1}, f_0, f_1)} \mid \boldsymbol{F}\ M\ I) = \prod_{t=1}^{n} F_{x_t} = \prod_{x=-1}^{1} (F_x)^{n f_x}$$

i.i.d.

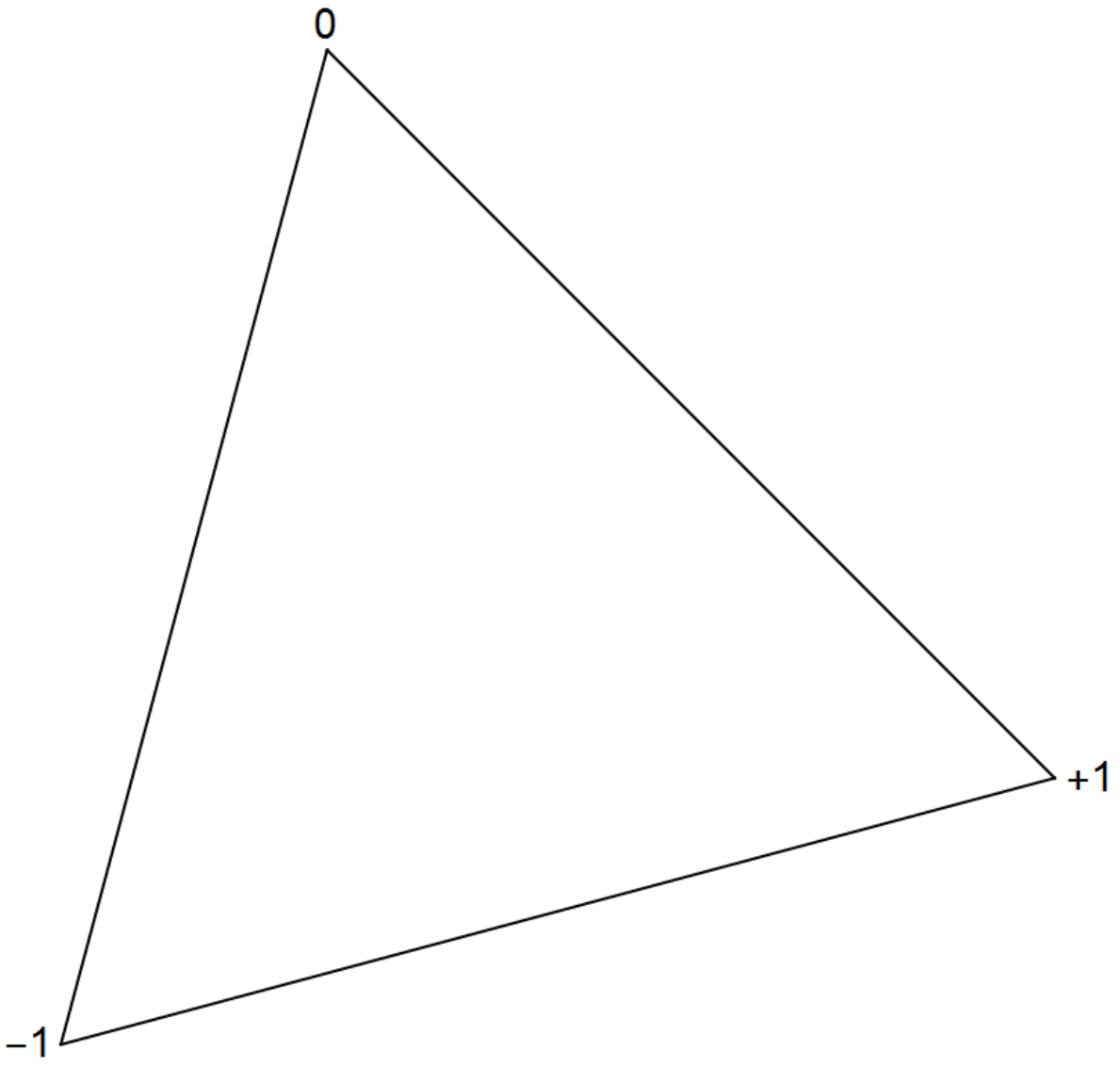
$$\mathrm{p}(\underbrace{x_1, \ldots, x_n}_{\{-1,\,0,\,1\}\text{ appear with frequencies }\boldsymbol{f} := (f_{-1}, f_0, f_1)} \mid \boldsymbol{F}\ M\ I) = \prod_{t=1}^{n} F_{x_t} = \prod_{x=-1}^{1} (F_x)^{n f_x}$$

i.i.d., or:

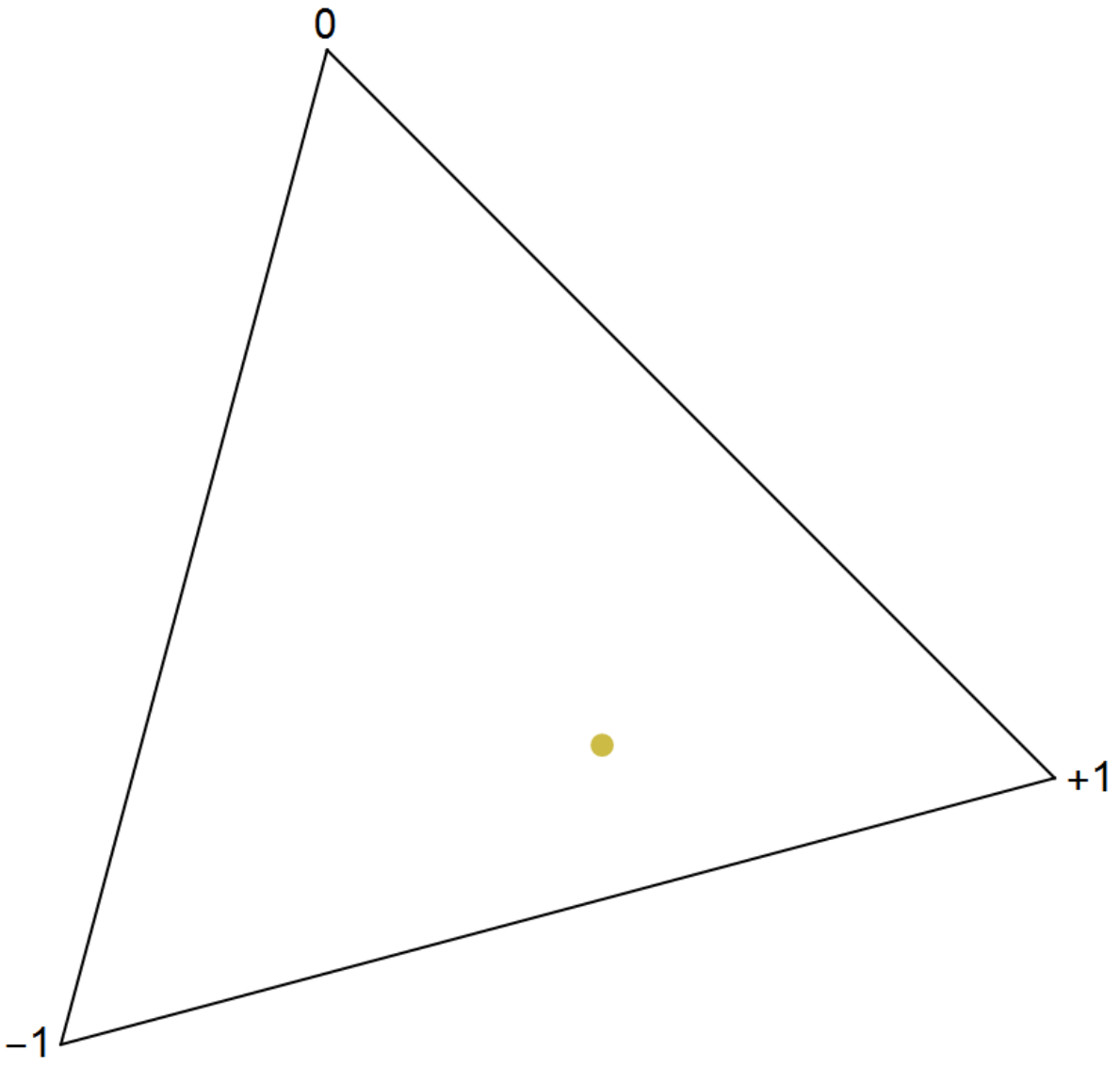
non-learning model

$$\mathrm{p}(x_1, \ldots, x_n \mid \cancel{x_{1'}, \ldots, x_{n'}}\ \boldsymbol{F}\ M\ I) = \mathrm{p}(x_1, \ldots, x_n \mid \boldsymbol{F}\ M\ I)$$

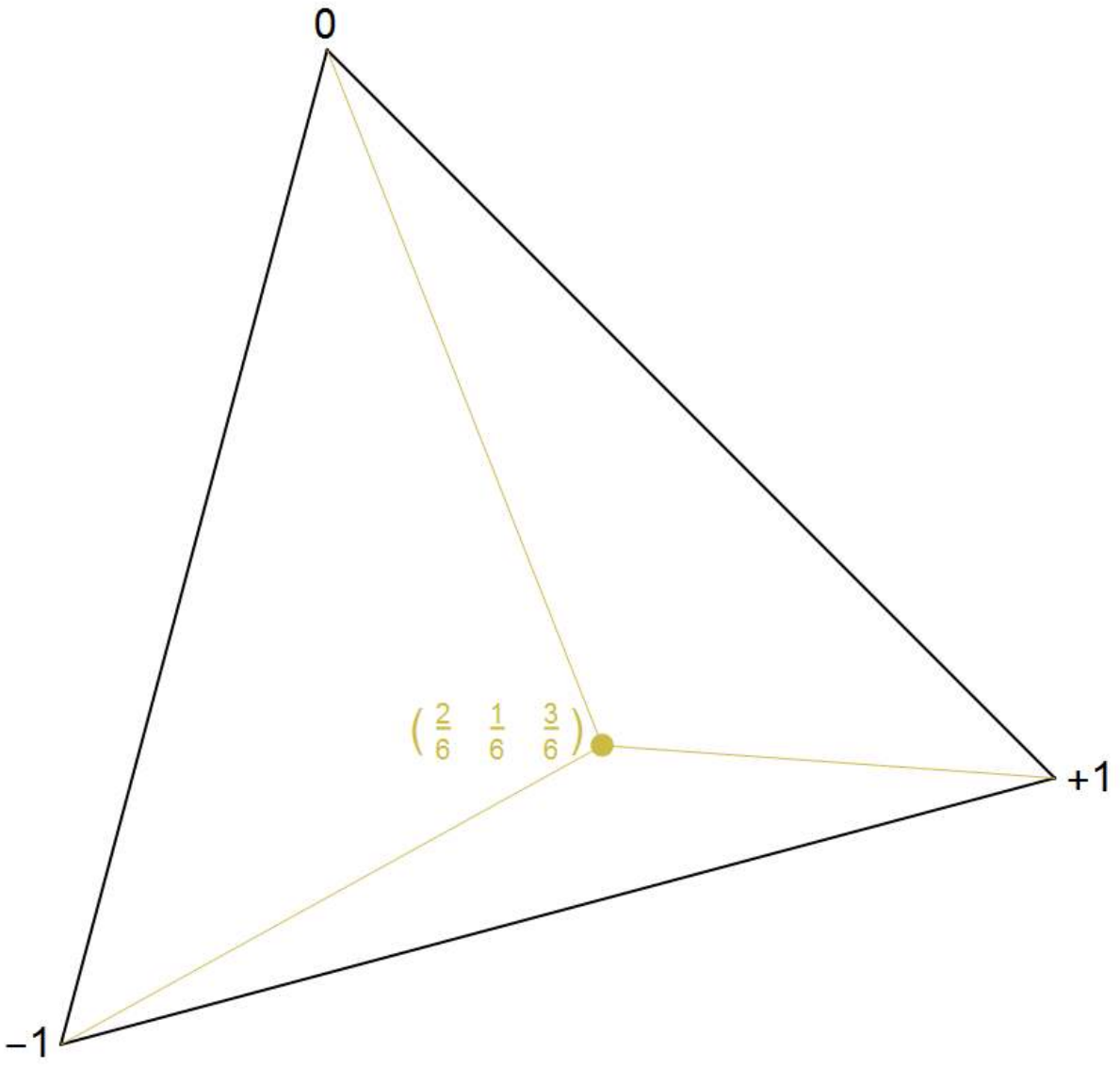
Geometric representation



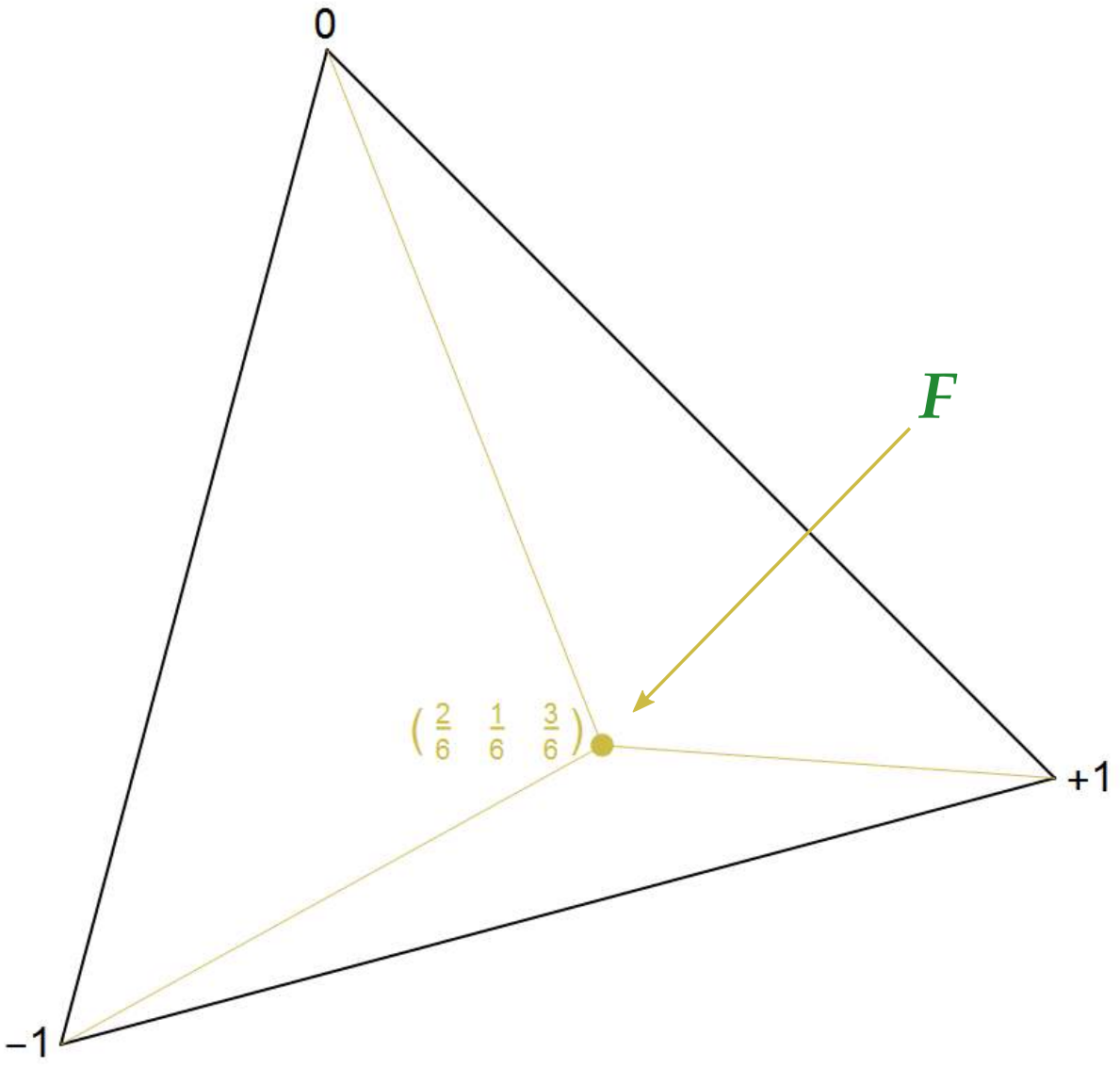
Geometric representation



Geometric representation

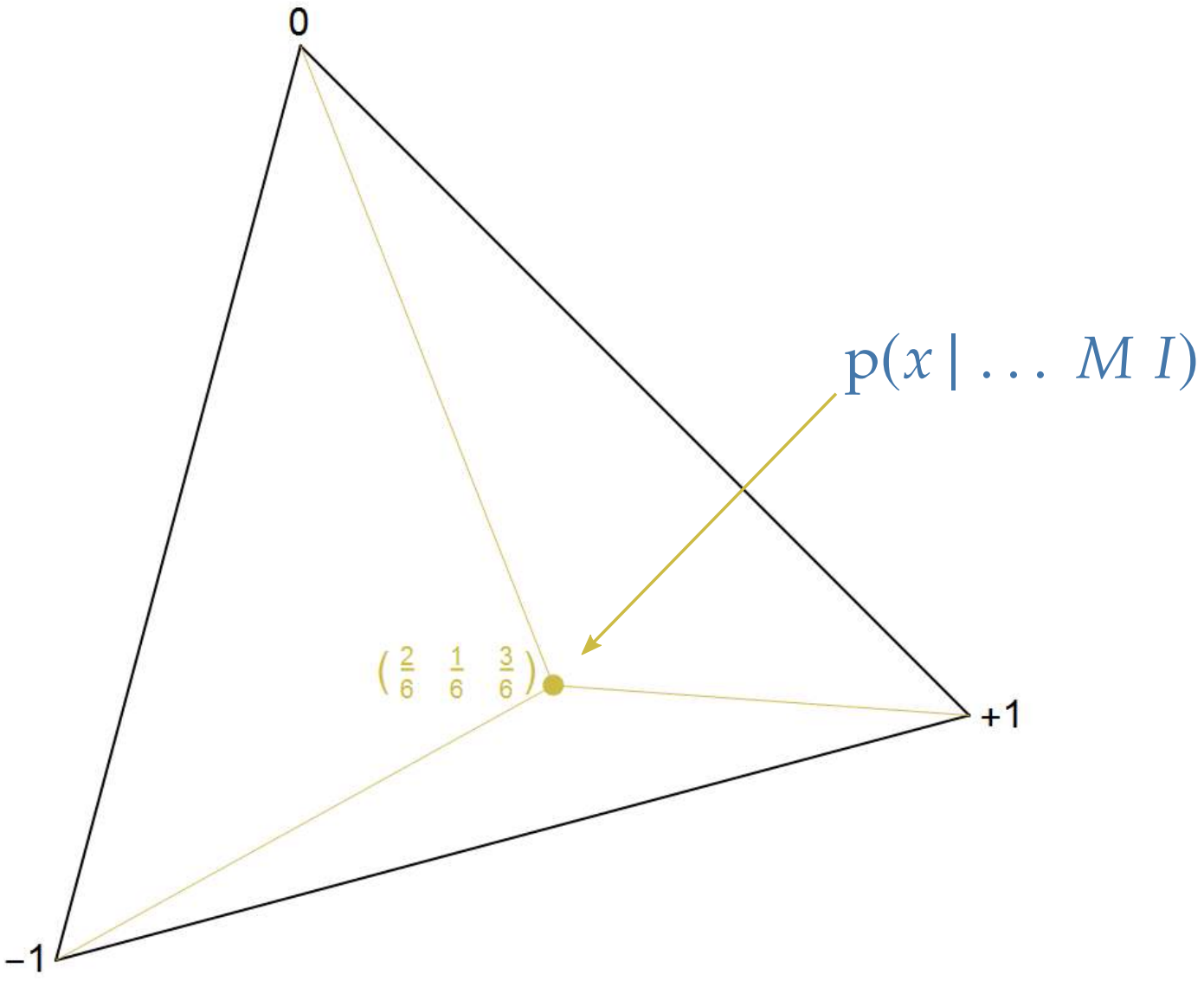


Geometric representation



6.4

Geometric representation



$$\mathrm{p}(x_1, \ldots, x_n \mid \boldsymbol{F}\ M\ I) = \prod_{x=-1}^{1} (F_x)^{n f_x} \equiv \frac{\exp[-n\, D(\boldsymbol{f}, \boldsymbol{F})]}{\exp[n\, H(\boldsymbol{f})]}$$

$D(\boldsymbol{f}, \boldsymbol{F}) := \sum_x f_x \ln(f_x / F_x) \geqslant 0$ **relative entropy (Kullback-Leibler)**

$H(\boldsymbol{f}) := -\sum_x f_x \ln f_x \geqslant 0$ **Shannon entropy**

we don't know the long-run frequencies

$$p(F \mid M\ I)\ \mathrm{d}F$$

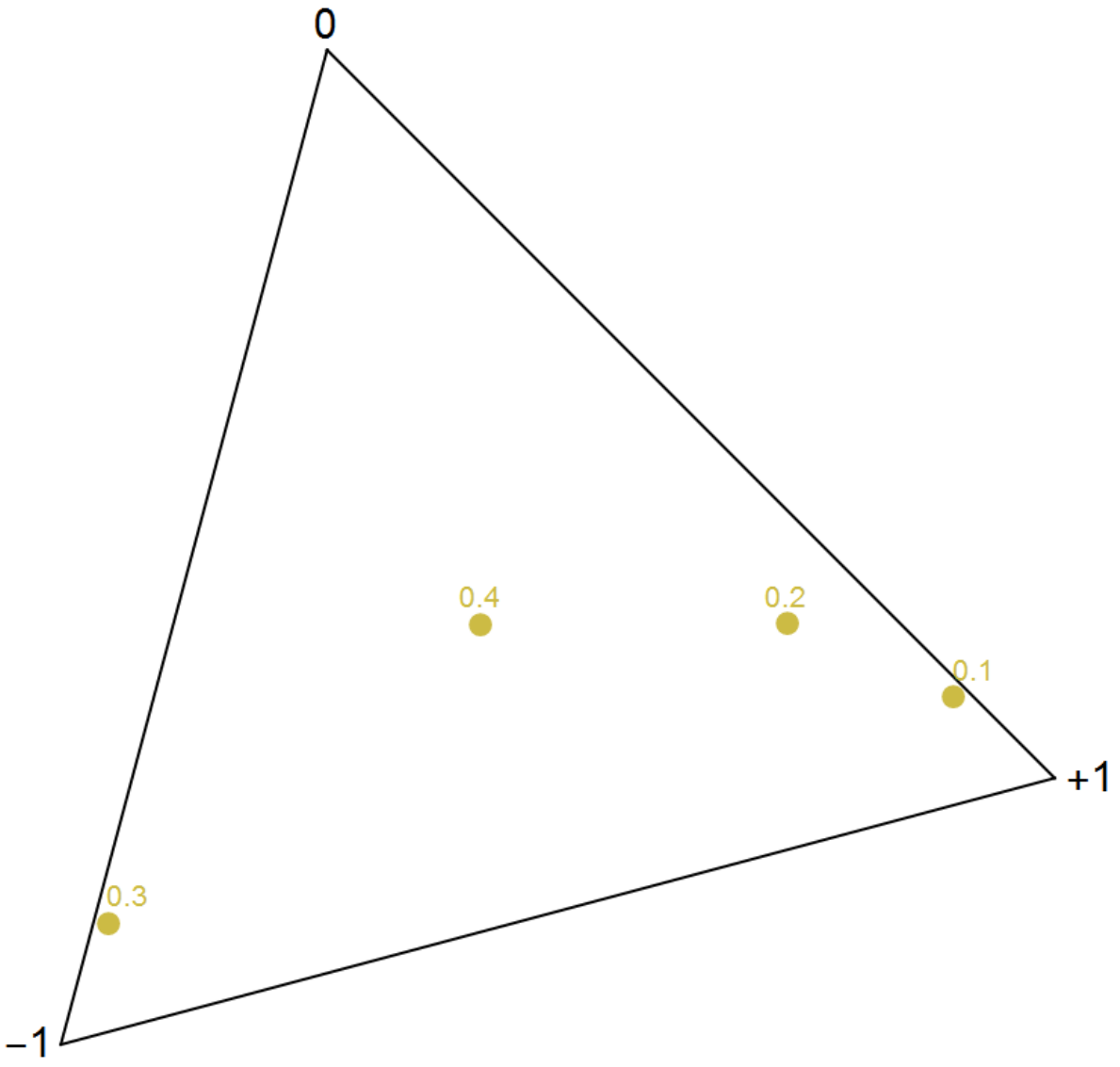
we don't know the long-run frequencies $\mathrm{p}(\boldsymbol{F} \mid M\ I)\ \mathrm{d}\boldsymbol{F}$

$$\mathrm{p}(x_1, \ldots, x_n \mid M\ I) = \int \mathrm{d}\boldsymbol{F}\ \mathrm{p}(x_1, \ldots, x_n \mid \boldsymbol{F}\ M\ I)\ \mathrm{p}(\boldsymbol{F} \mid M\ I)$$

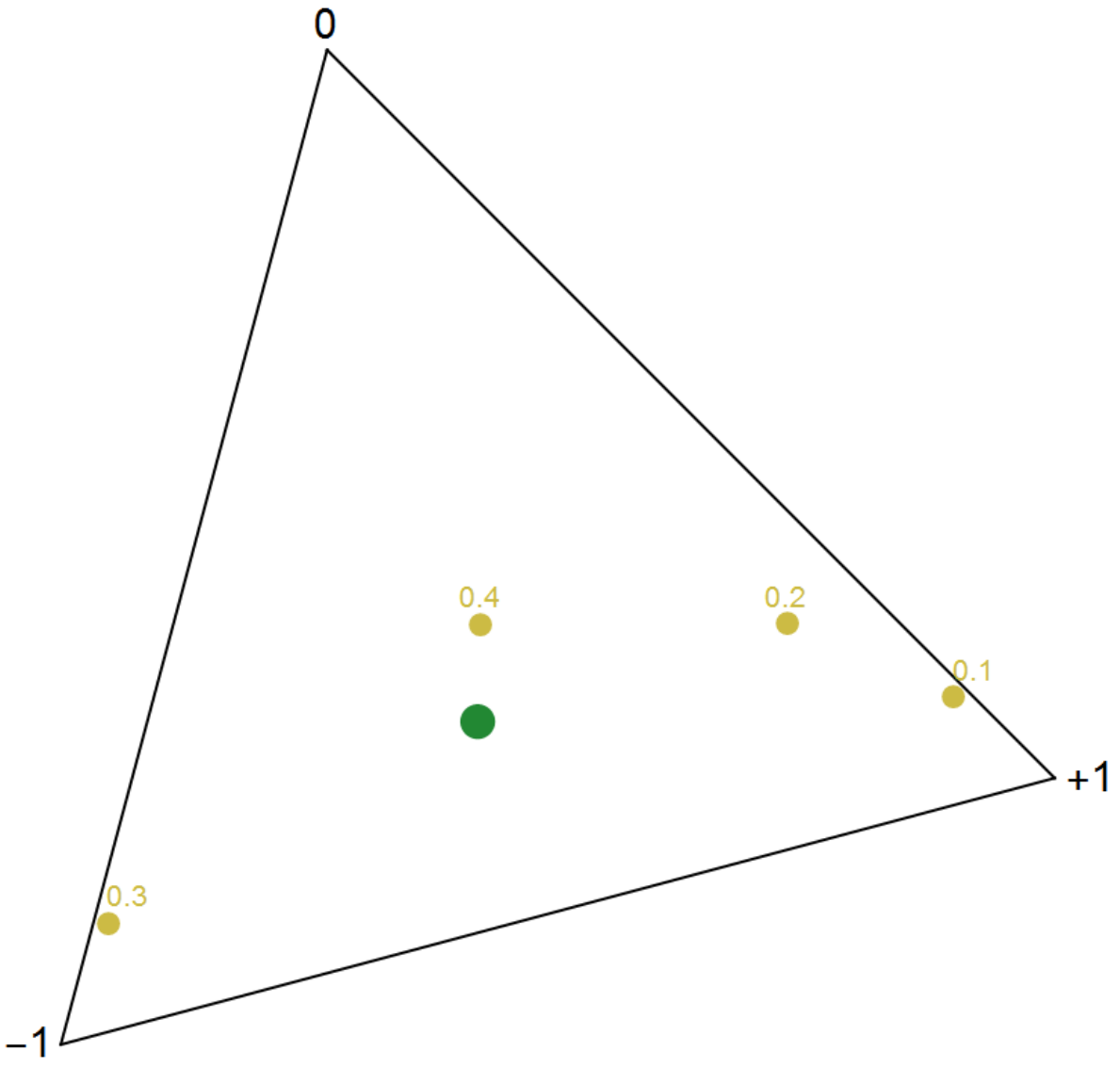
$$= \int \mathrm{d}\boldsymbol{F}\ \frac{\exp[-n\, D(\boldsymbol{f}, \boldsymbol{F})]}{\exp[n\, H(\boldsymbol{f})]}\ \mathrm{p}(\boldsymbol{F} \mid M\ I)$$

convex mixture of non-learning models

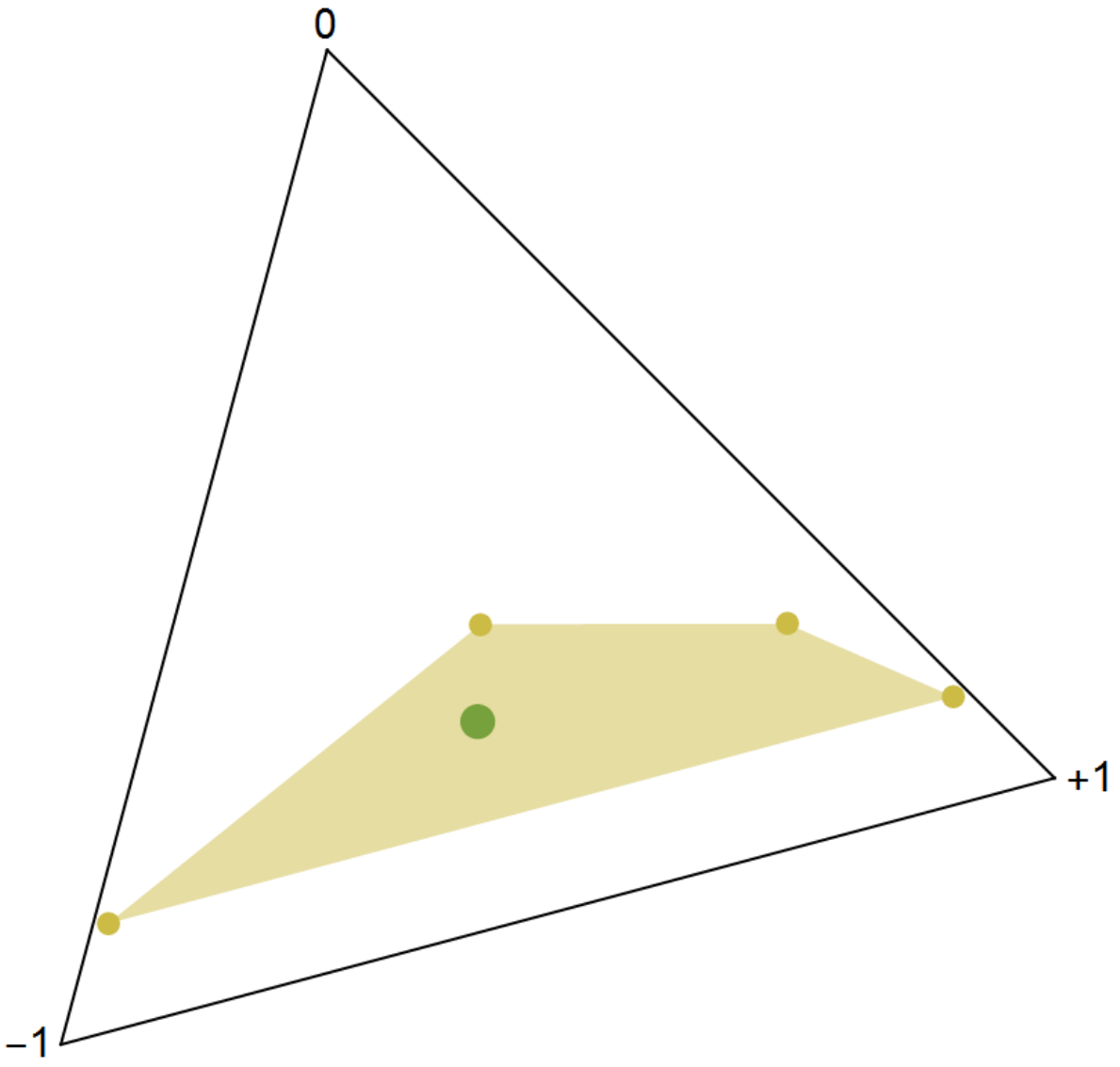
Convex mixture



Convex mixture



Convex mixture



The result is a **learning** model

$$\mathrm{p}(x_1, \ldots, x_n \mid x_{1'}, \ldots, x_{n'}\ M\ I) \neq \mathrm{p}(x_1, \ldots, x_n \mid M\ I)$$

The result is a **learning** model

$$\mathrm{p}(x_1, \ldots, x_n \mid x_{1'}, \ldots, x_{n'}\ M\ I) \neq \mathrm{p}(x_1, \ldots, x_n \mid M\ I)$$

$$\mathrm{p}(\overline{x} \mid \overline{x'}\ M\ I) = \int \mathrm{d}\boldsymbol{F}\ \frac{\exp[-n\, D(\boldsymbol{f}, \boldsymbol{F})]}{\exp[n\, H(\boldsymbol{f})]}\ \mathrm{p}(\boldsymbol{F} \mid \overline{x'}\ M\ I)$$

The result is a **learning** model

$$\mathrm{p}(x_1, \ldots, x_n \mid x_{1'}, \ldots, x_{n'}\ M\ I) \neq \mathrm{p}(x_1, \ldots, x_n \mid M\ I)$$

$$\mathrm{p}(\overline{x} \mid \overline{x'}\ M\ I) = \int \mathrm{d}\boldsymbol{F}\ \frac{\exp[-n\, D(\boldsymbol{f}, \boldsymbol{F})]}{\exp[n\, H(\boldsymbol{f})]}\ \mathrm{p}(\boldsymbol{F} \mid \overline{x'}\ M\ I)$$

$$\mathrm{p}(\boldsymbol{F} \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\, D(\boldsymbol{f}', \boldsymbol{F})]\ \mathrm{p}(\boldsymbol{F} \mid M\ I)}{\int \mathrm{d}\check{\boldsymbol{F}}\ \exp[-n'\, D(\boldsymbol{f}', \check{\boldsymbol{F}})]\ \mathrm{p}(\check{\boldsymbol{F}} \mid M\ I)}$$

Parametric model

support of $p(F \mid M\ I)$ has lower dimension $F(\theta)$ $\theta \in \mathbf{R}$

Parametric model

support of $p(F \mid M\ I)$ has lower dimension $F(\theta)$ $\theta \in \mathbf{R}$

$$p(\overline{x} \mid \overline{x'}\ M\ I) = \int dF\ \frac{\exp[-n\ D(f, F)]}{\exp[n\ H(f)]}\ p(F \mid \overline{x'}\ M\ I)$$

$$p(F \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\ D(f', F)]\ p(F \mid M\ I)}{\int d\check{F}\ \exp[-n'\ D(f', \check{F})]\ p(\check{F} \mid M\ I)}$$

Parametric model

support of $p(F \mid M\ I)$ has lower dimension $F(\theta)$ $\theta \in \mathbf{R}$

$$p(\overline{x} \mid \overline{x'}\ M\ I) = \int dF\ \frac{\exp[-n\ D(f, F)]}{\exp[n\ H(f)]}\ p(F \mid \overline{x'}\ M\ I)$$

$$p(F \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\ D(f', F)]\ p(F \mid M\ I)}{\int d\check{F}\ \exp[-n'\ D(f', \check{F})]\ p(\check{F} \mid M\ I)}$$

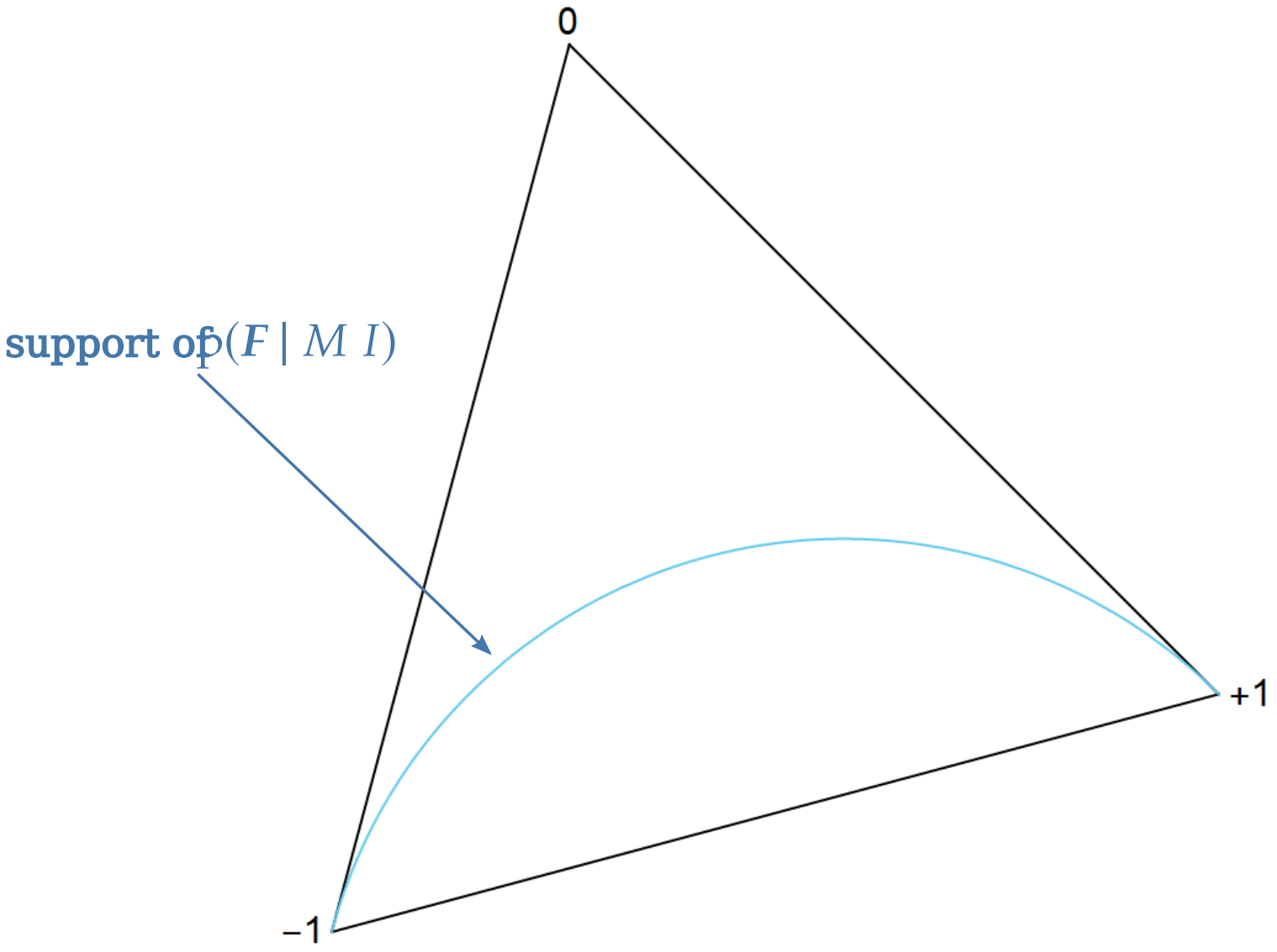
Parametric model

support of $p(F \mid M\ I)$ has lower dimension $F(\theta)$ $\theta \in \mathbf{R}$

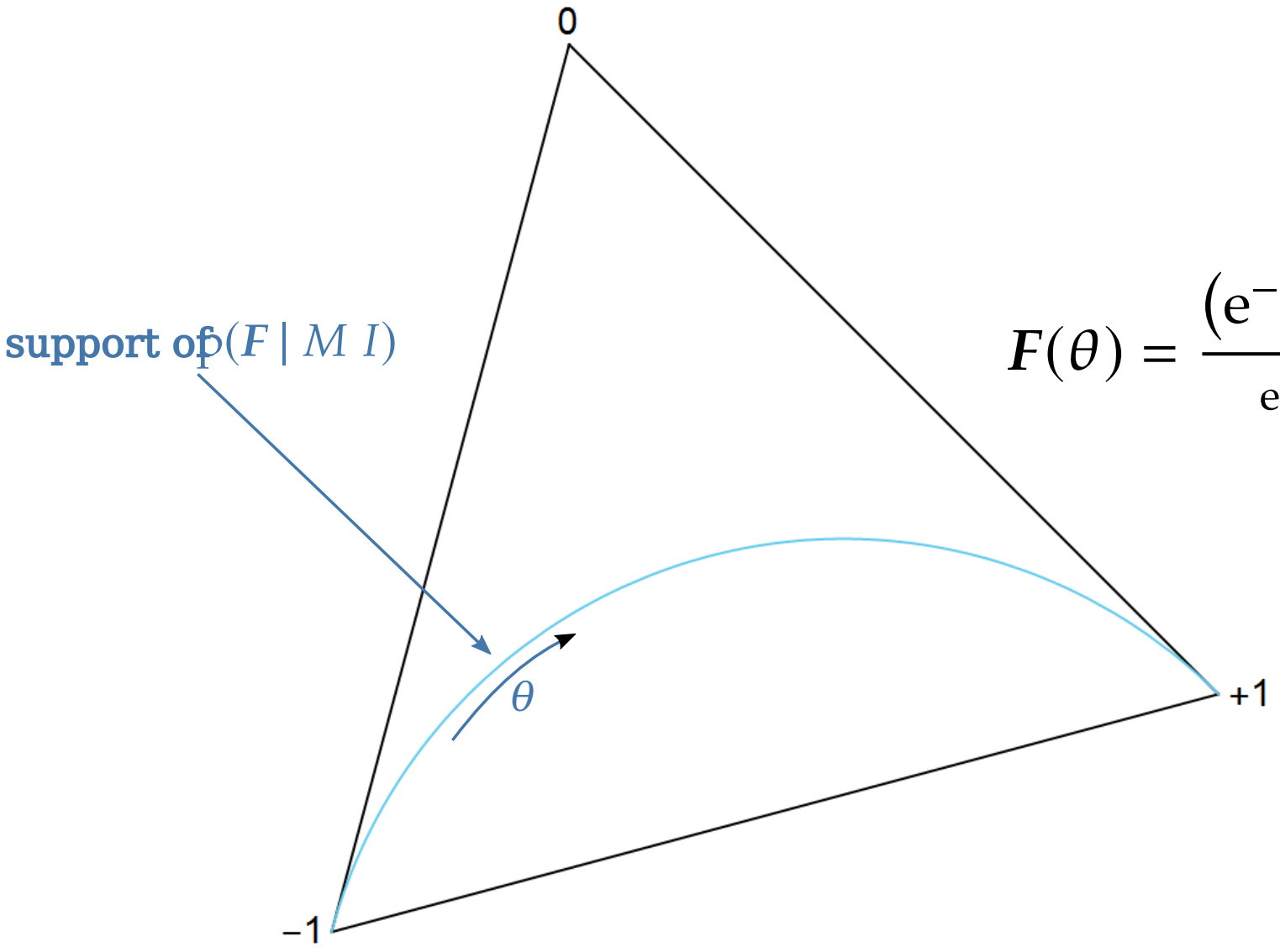
$$p(\overline{x} \mid \overline{x'}\ M\ I) = \int d\theta\ \frac{\exp\left[-n\, D\left(f, F(\theta)\right)\right]}{\exp\left[n\, H(f)\right]}\ p(\theta \mid \overline{x'}\ M\ I)$$

$$p(\theta \mid \overline{x'}\ M\ I) = \frac{\exp\left[-n'\, D\left(f', F(\theta)\right)\right] p(\theta \mid M\ I)}{\int d\breve{\theta}\, \exp\left[-n'\, D\left(f', F(\breve{\theta})\right)\right] p(\breve{\theta} \mid M\ I)}$$

Parametric model



Parametric model



$$\boldsymbol{F}(\theta) = \frac{\left(\mathrm{e}^{-\theta},\ 1,\ \mathrm{e}^{\theta}\right)}{\mathrm{e}^{-\theta}+1+\mathrm{e}^{\theta}}$$

$$\mathrm{p}(\theta \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\, D(\boldsymbol{f}', \boldsymbol{F}(\theta))]\ \mathrm{p}(\theta|M\ I)}{\int \mathrm{d}\breve{\theta}\, \exp[-n'\, D(\boldsymbol{f}', \boldsymbol{F}(\breve{\theta}))]\ \mathrm{p}(\theta|M\ I)}$$

$$\mathrm{p}(\theta \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\ D(f', \boldsymbol{F}(\theta))]\ \mathrm{p}(\theta|M\ I)}{\int \mathrm{d}\breve{\theta}\ \exp[-n'\ D(f', \boldsymbol{F}(\breve{\theta}))]\ \mathrm{p}(\theta|M\ I)} \xrightarrow{n'\to\infty} \delta(\theta - \theta^*)$$

$$\theta^* = \arg\inf_{\theta} D\big(f' - \boldsymbol{F}(\theta)\big)$$

$$\mathrm{p}(\theta \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\ D(f', \boldsymbol{F}(\theta))]\ \mathrm{p}(\theta|M\ I)}{\int \mathrm{d}\breve{\theta}\ \exp[-n'\ D(f', \boldsymbol{F}(\breve{\theta}))]\ \mathrm{p}(\theta|M\ I)} \xrightarrow{n'\to\infty} \delta(\theta - \theta^*)$$

$$\theta^* = \arg\inf_{\theta} D\big(f' - \boldsymbol{F}(\theta)\big)$$

$$\mathrm{p}(\overline{x} \mid \overline{x'}\ M\ I) \xrightarrow{n'\to\infty} \mathrm{p}\big(\overline{x} \mid \boldsymbol{F}(\theta^*),\ M\ I\big)$$

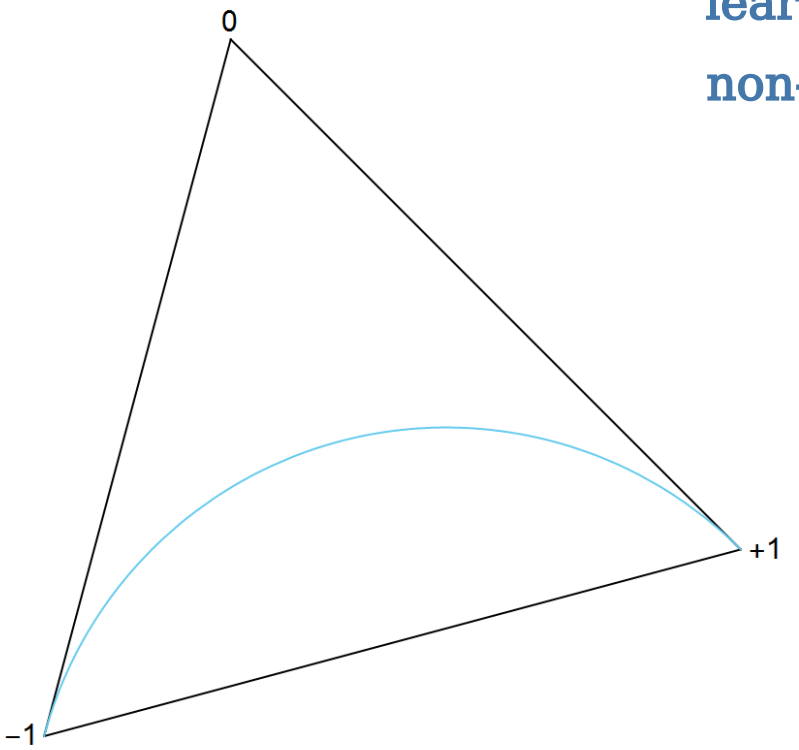
$$\mathrm{p}(\theta \mid \overline{x'}\ M\ I) = \frac{\exp[-n'\, D(f', \boldsymbol{F}(\theta))]\ \mathrm{p}(\theta|M\ I)}{\int \mathrm{d}\breve{\theta}\, \exp[-n'\, D(f', \boldsymbol{F}(\breve{\theta}))]\ \mathrm{p}(\theta|M\ I)} \xrightarrow{n'\to\infty} \delta(\theta - \theta^*)$$

$$\theta^* = \arg\inf_{\theta} D\big(f' - \boldsymbol{F}(\theta)\big)$$

$$\mathrm{p}(\overline{x} \mid \overline{x'}\ M\ I) \xrightarrow{n'\to\infty} \mathrm{p}\big(\overline{x} \mid \boldsymbol{F}(\theta^*),\ M\ I\big)$$

learning model —increasing data size→ **non-learning** model

(Summary)



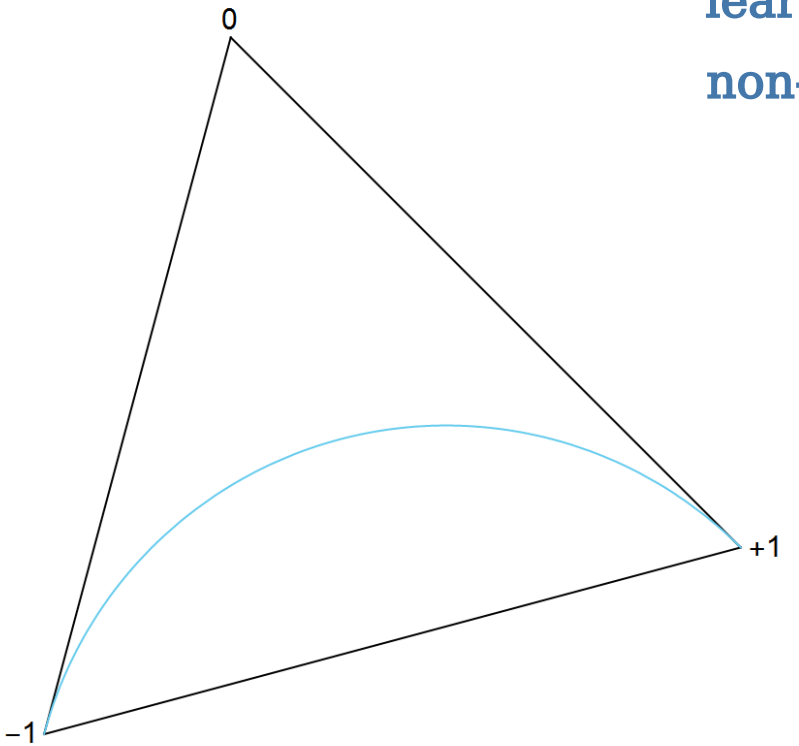
learning model = convex mixture of non-learning mo

non-learning models = extreme points (sufficiency m

(Summary)

learning model = convex mixture of non-learning mo

non-learning models = extreme points (sufficiency m



Scand J Statist 1: 128–134, 1974

Sufficiency, Prediction and Extreme Models

STEFFEN L. LAURITZEN

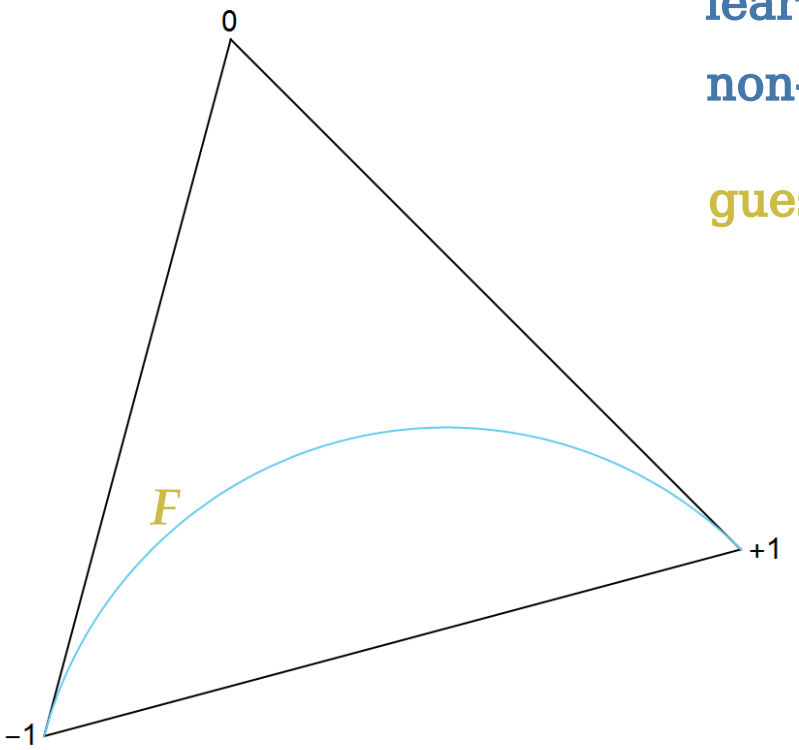
University of Copenhagen

Received August 1973, revised October 1973

ABSTRACT. A modified concept of sufficiency, relevant in connection with statistical analysis of stochastic processes, is defined and its basic properties investigated. A method of prediction that applies when the probability structure is partly unknown is introduced and the method is shown to possess certain important invariance properties. The concept of an extreme model is defined and its probabilistic and statistical properties discussed. Existence of maximum likelihood estimators and predictors is established under weak regularity assumptions. For technical convenience, only discrete-valued stochastic processes are considered throughout the paper.

ed total sufficiency
mentary properties
In section 3 a p
(based on the meth
applies to the situat
ture of the process
However, it appears
able results if appl
tistical models.

(Summary)

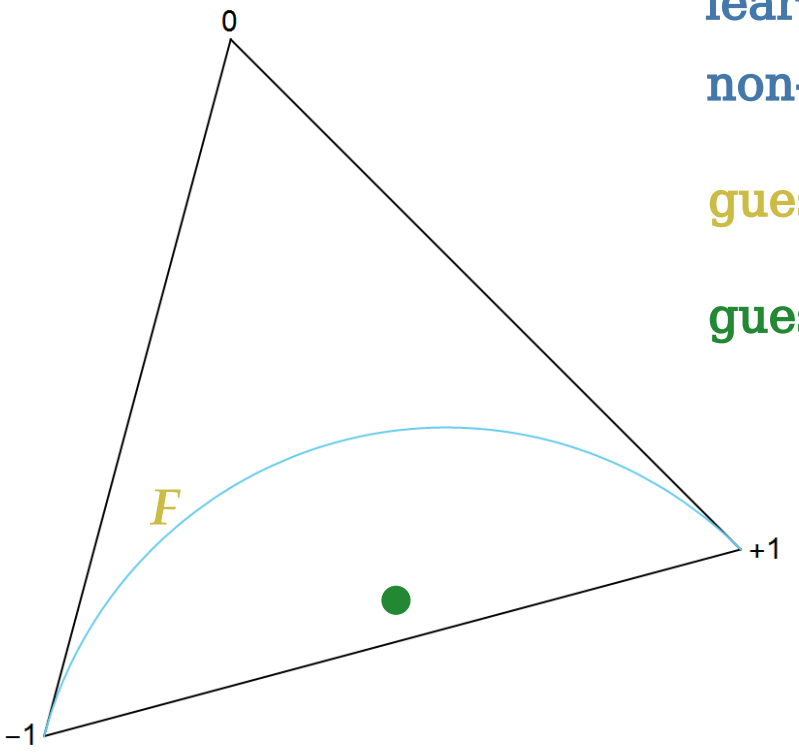


learning model = convex mixture of non-learning mo

non-learning models = extreme points (sufficiency m

guess for long-run frequency restricted to extreme p

(Summary)



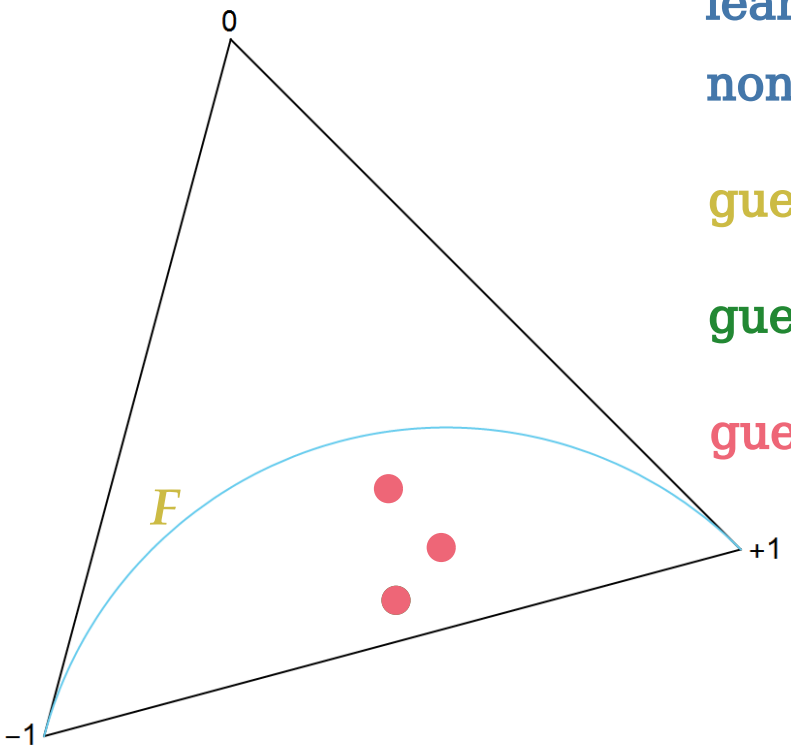
learning model = convex mixture of non-learning mo

non-learning models = extreme points (sufficiency m

guess for long-run frequency restricted to extreme p

guess for *next* observation = point in convex hull

(Summary)



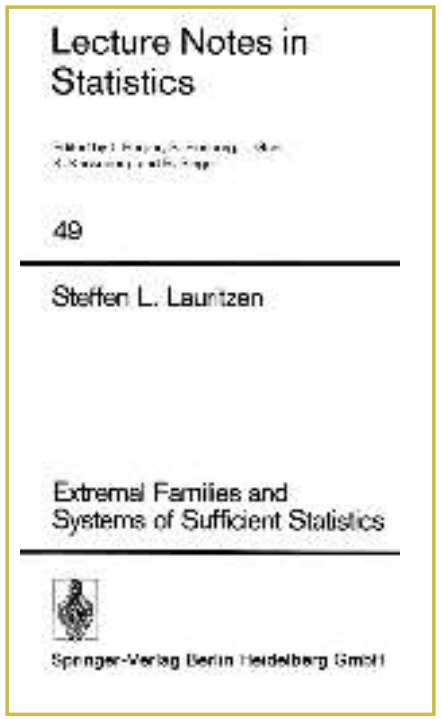
learning model = convex mixture of non-learning mo

non-learning models = extreme points (sufficiency m

guess for long-run frequency restricted to extreme p

guess for *next* observation = point in convex hull

guess for *several* observations = several points in co

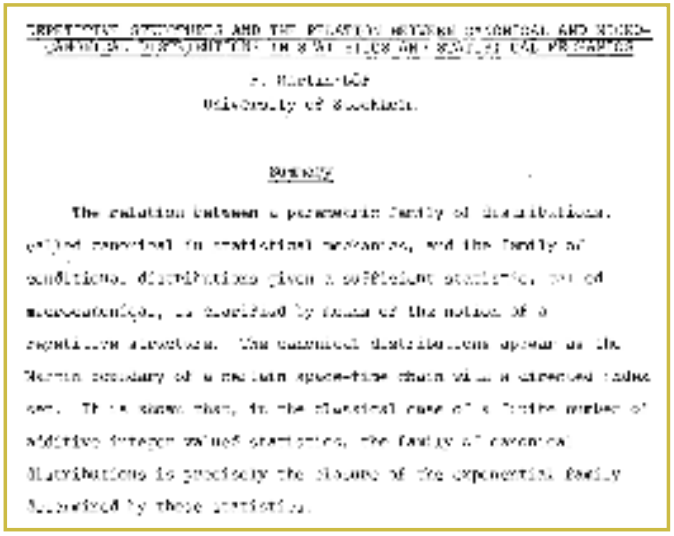


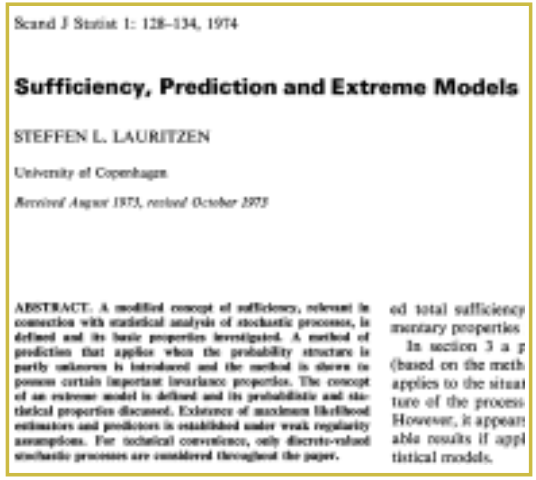
Notes on the Theory of Statistical Symbol Recognition

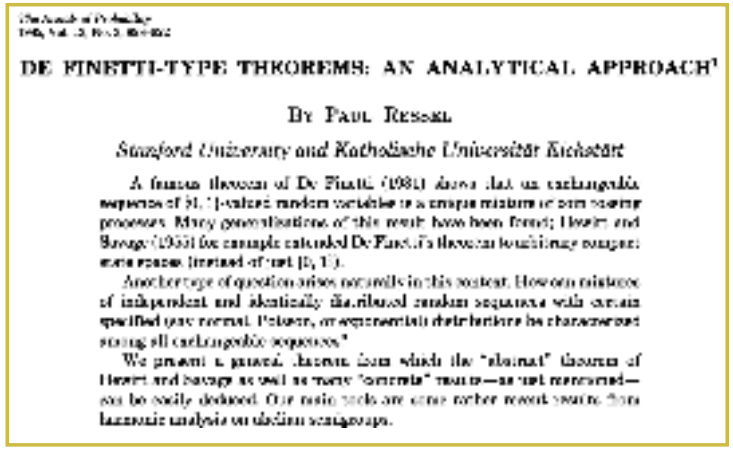
Report no. 778/1986

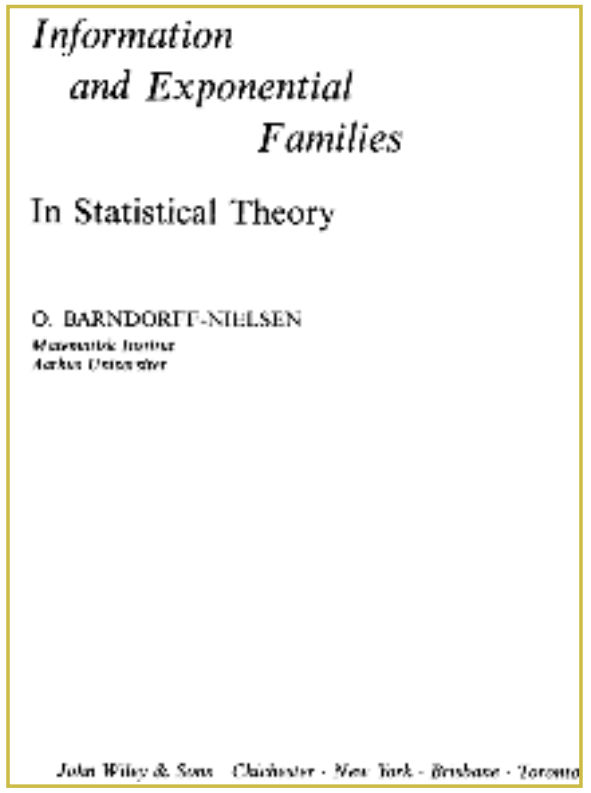
By Nils Lid Hjort

Norwegian Computing Center





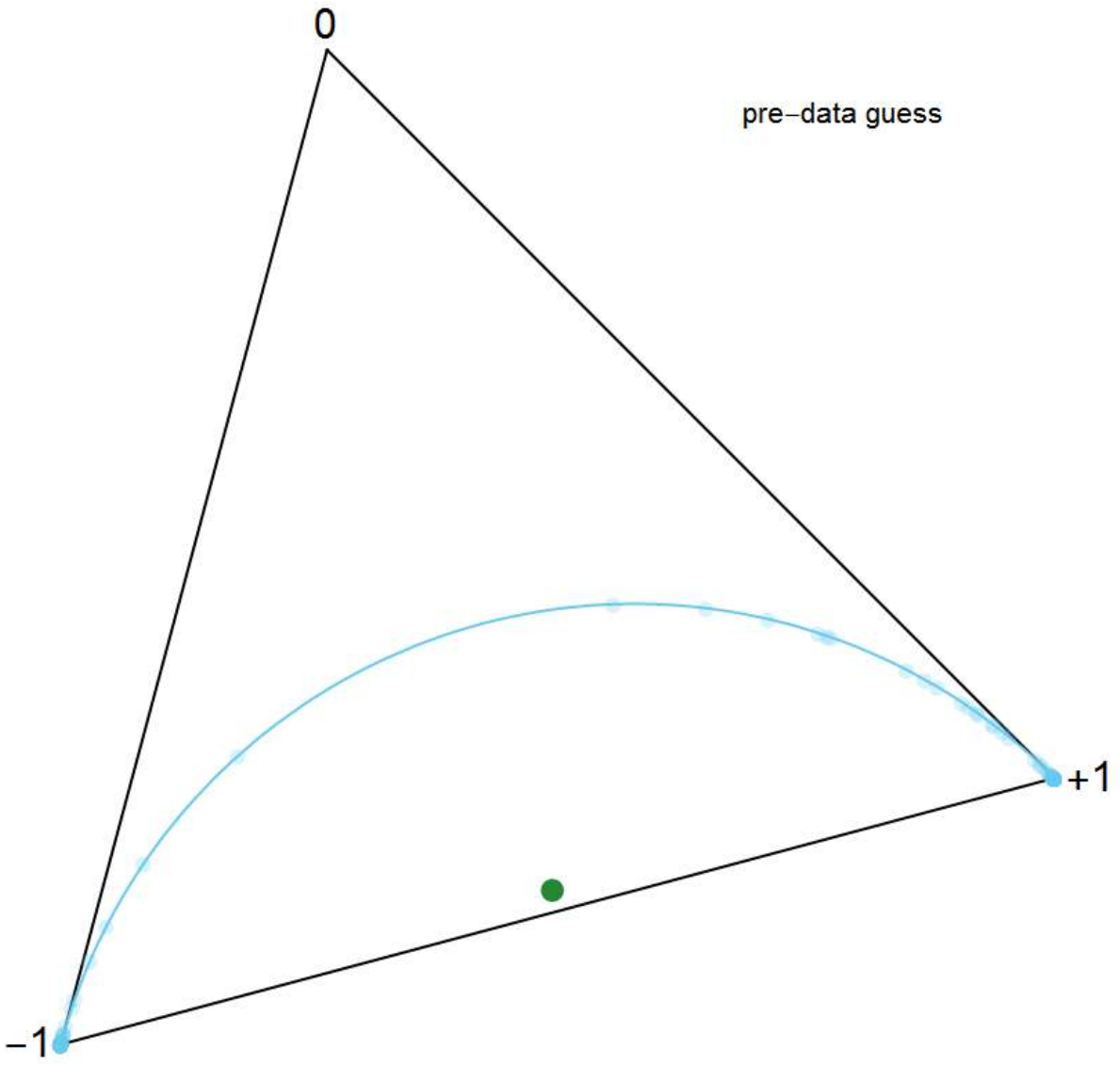




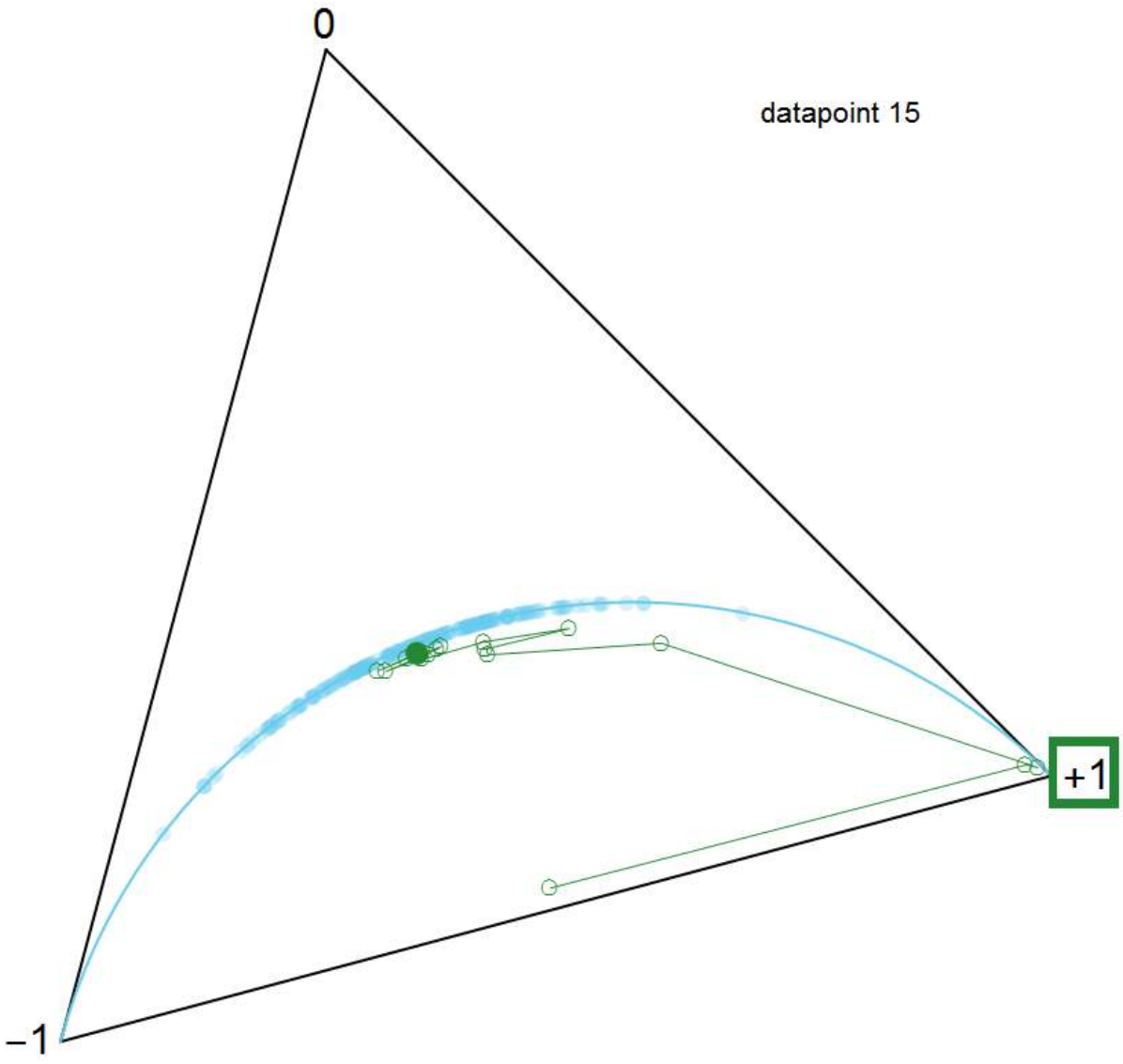
2 Exchangeability and its ramifications

A. PHILIP DAWID

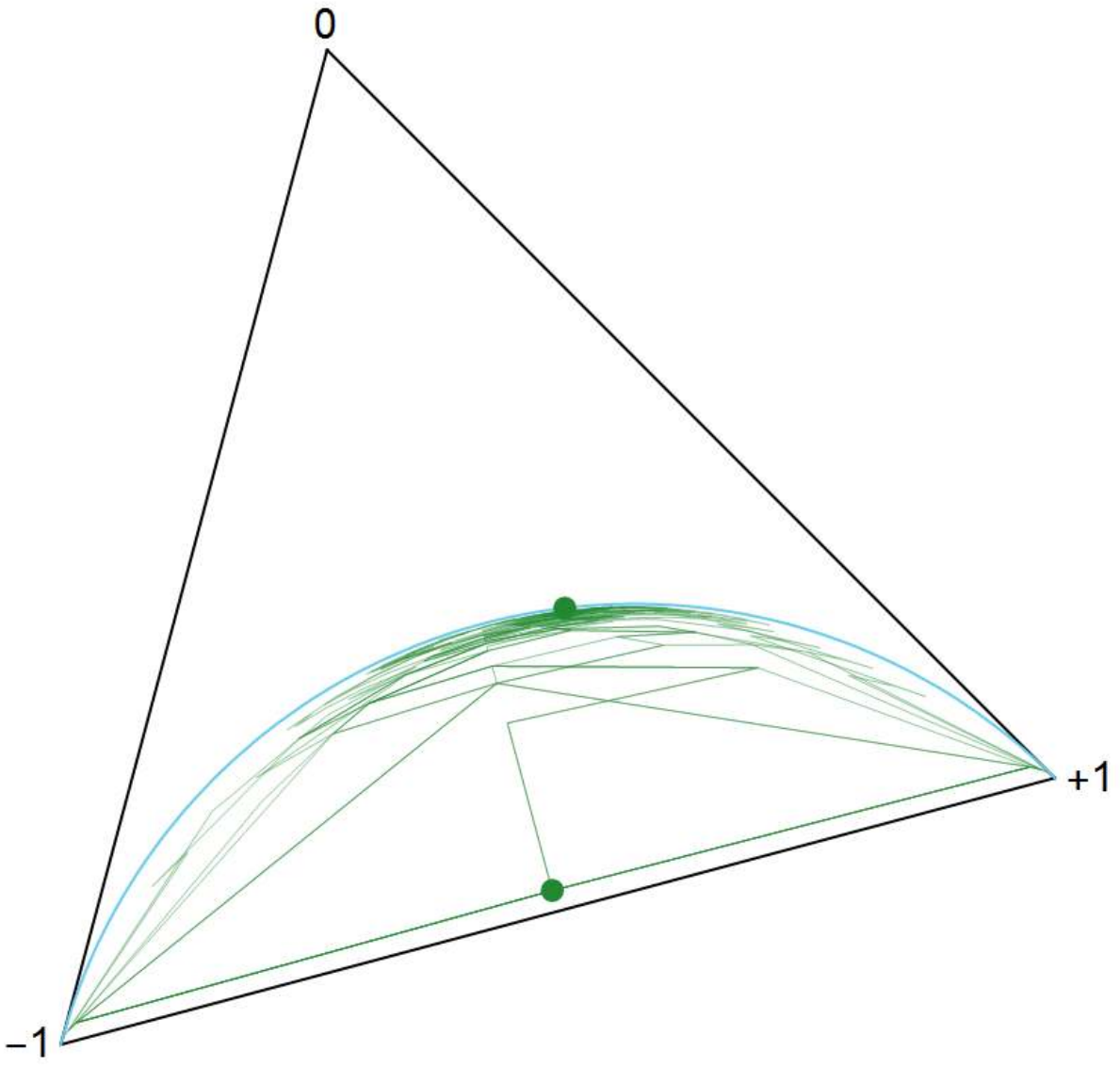
Learning



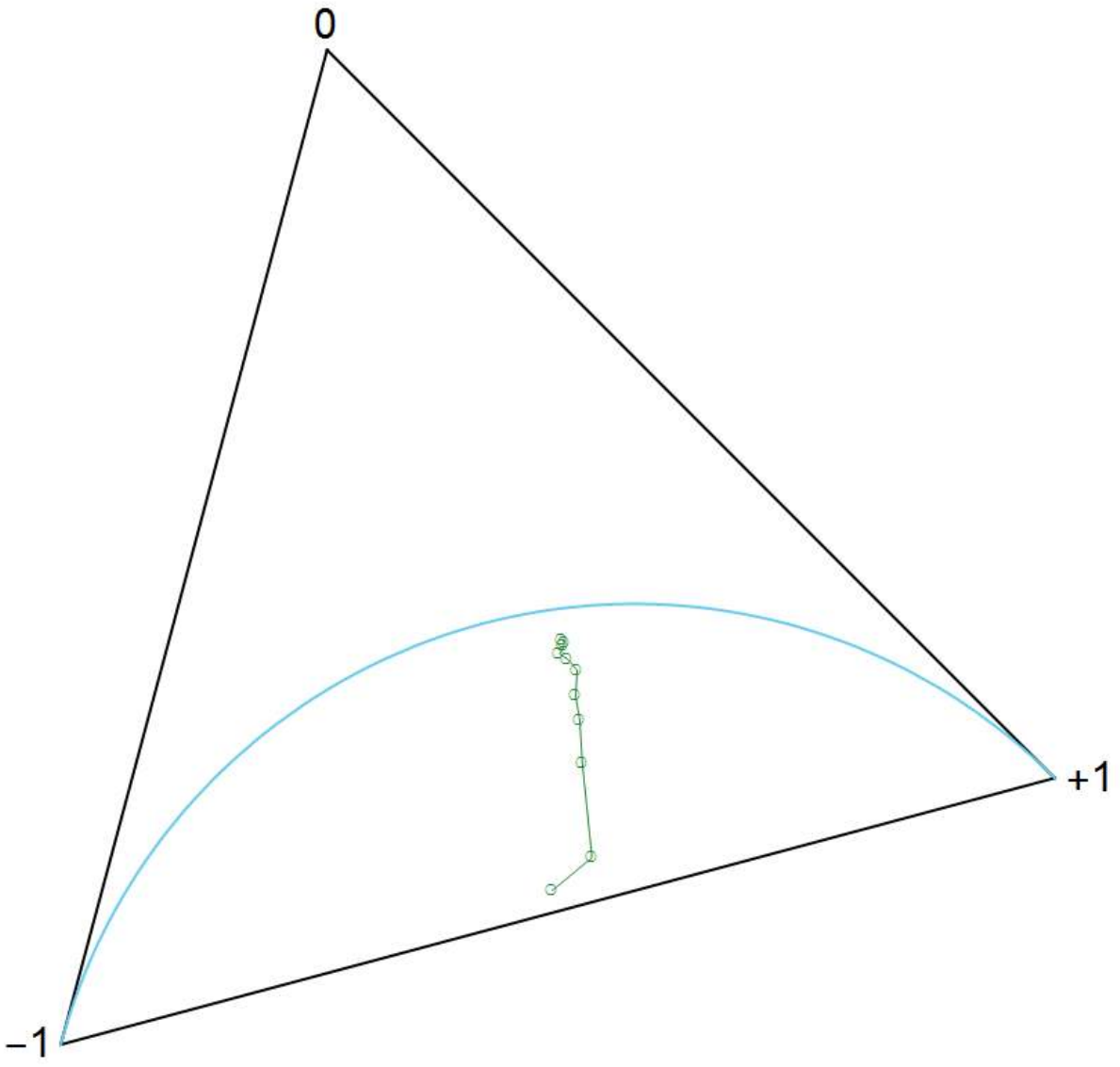
Learning



Average learning path



Average learning path



Average learning path

$$1, 2, \ldots, n' \xrightarrow{\pi^k} \pi_1^k, \pi_2^k, \ldots, \pi_k^k$$

Average learning path

$$1, 2, \ldots, n' \xrightarrow{\pi^k} \pi_1^k, \pi_2^k, \ldots, \pi_k^k$$

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \xrightarrow{\pi^4} 2, 5, 8, 9$$

Average learning path

$$1, 2, \ldots, n' \xrightarrow{\pi^k} \pi^k_1, \pi^k_2, \ldots, \pi^k_k$$

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \xrightarrow{\pi^4} 2, 5, 8, 9$$

$$\overline{x'} := x'_1, x'_2, \ldots, x'_{n'} \longrightarrow \pi^k(\overline{x'}) := x'_{\pi^k_1}, x'_{\pi^k_2}, \ldots, x'_{\pi^k_k}$$

Average learning path

$$1, 2, \ldots, n' \xrightarrow{\pi^k} \pi^k_1, \pi^k_2, \ldots, \pi^k_k$$

$$1,2,3,4,5,6,7,8,9,10 \xrightarrow{\pi^4} 2,5,8,9$$

$$\overline{x'} := x'_1, x'_2, \ldots, x'_{n'} \longrightarrow \pi^k(\overline{x'}) := x'_{\pi^k_1}, x'_{\pi^k_2}, \ldots, x'_{\pi^k_k}$$

$$\mathrm{p}(x \mid \overline{x'}^{(k)}\ M\ I) := \frac{1}{\binom{n'}{k}} \sum_{\{\pi^k\}} \mathrm{p}(x_i \mid \pi^k(\overline{x'})\ M\ I)$$

average guess after *k* data points from a set of *n*' data points

Average learning path

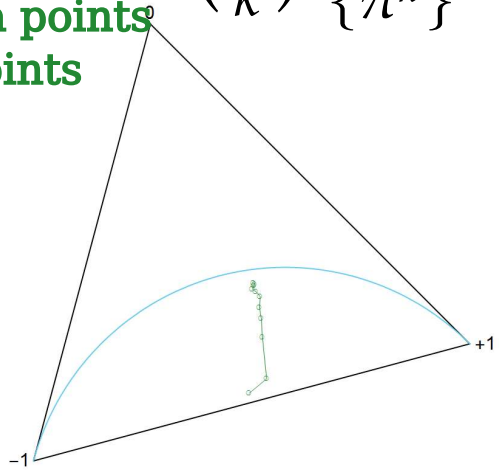
$$1, 2, \ldots, n' \xrightarrow{\pi^k} \pi^k_1, \pi^k_2, \ldots, \pi^k_k$$

$$1,2,3,4,5,6,7,8,9,10 \xrightarrow{\pi^4} 2,5,8,9$$

$$\overline{x'} := x'_1, x'_2, \ldots, x'_{n'} \longrightarrow \pi^k(\overline{x'}) := x'_{\pi^k_1}, x'_{\pi^k_2}, \ldots, x'_{\pi^k_k}$$

$$\mathrm{p}(x \mid \overline{x'}^{(k)}\ M\ I) := \frac{1}{\binom{n'}{k}} \sum_{\{\pi^k\}} \mathrm{p}(x_i \mid \pi^k(\overline{x'})\ M\ I)$$

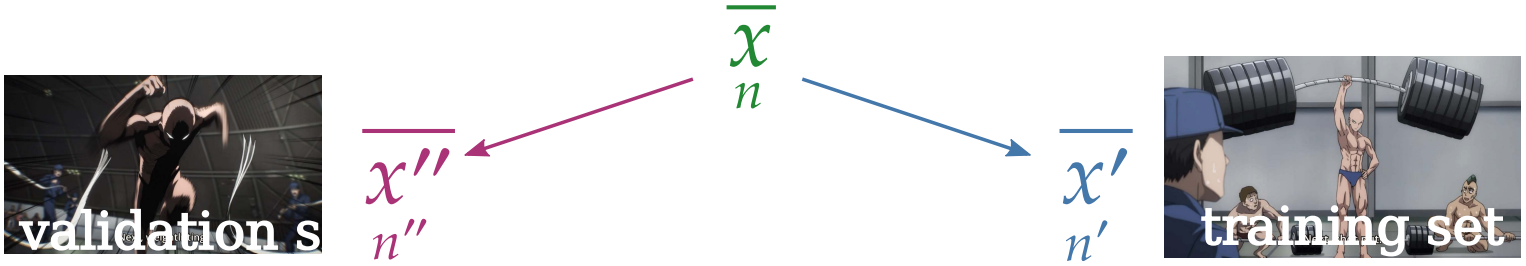
average guess after *k* data points from a set of *n*' data points



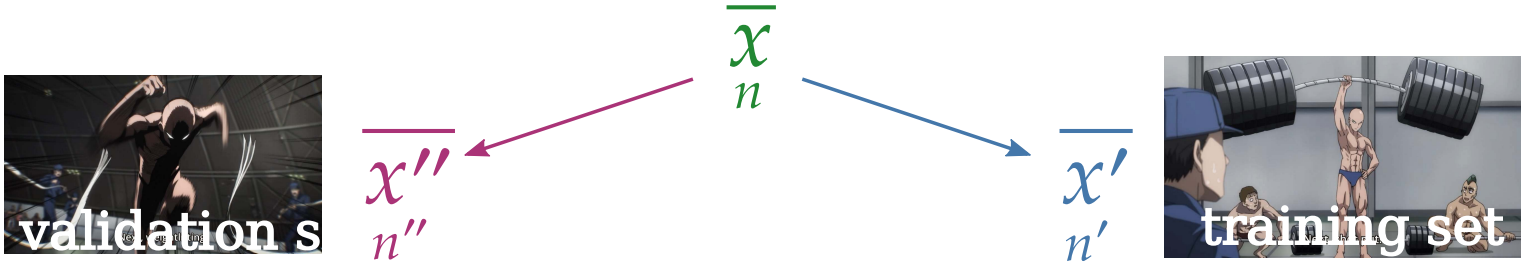
Let's cross-validate!

$$\overline{x}_n$$

Let's cross-validate!



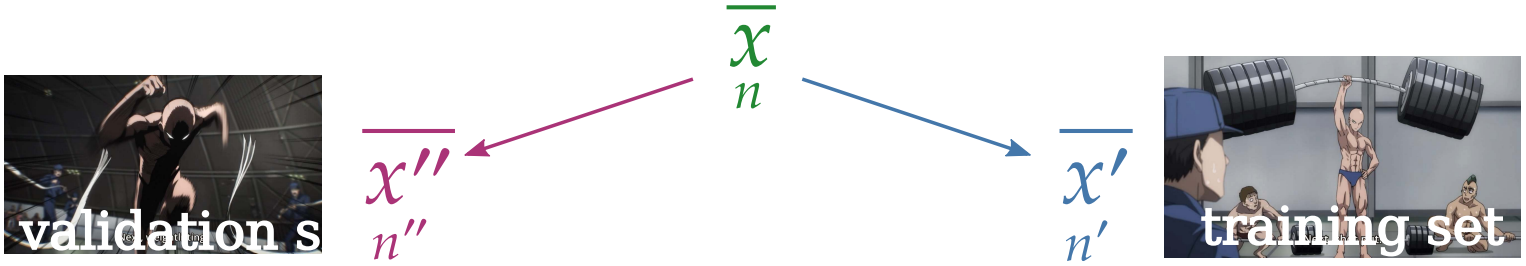
Let's cross-validate!



$$\overline{x}_{n} \rightarrow \overline{x''}_{n''}, \quad \overline{x}_{n} \rightarrow \overline{x'}_{n'}$$

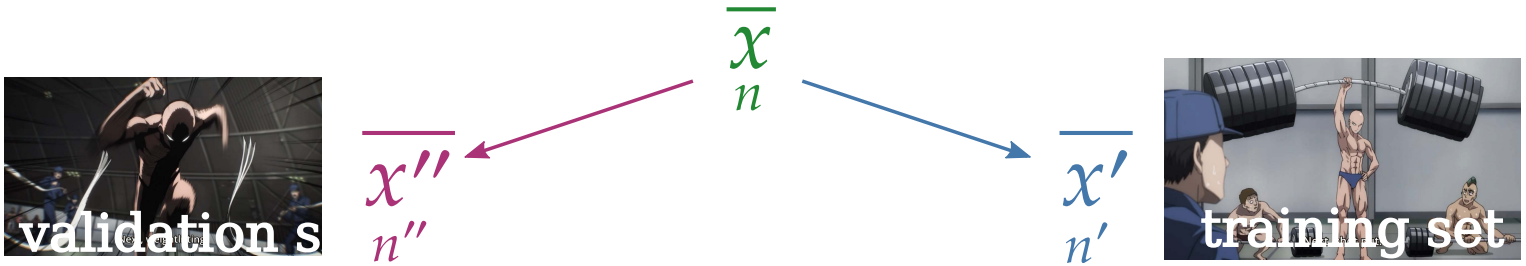
$$\mathrm{p}(x''_i \mid \overline{x'}^{(k)}\ M\ I)$$

Let's cross-validate!



$$\ln \mathrm{p}(x_i'' \mid \overline{x'}^{(k)}\ M\ I)$$

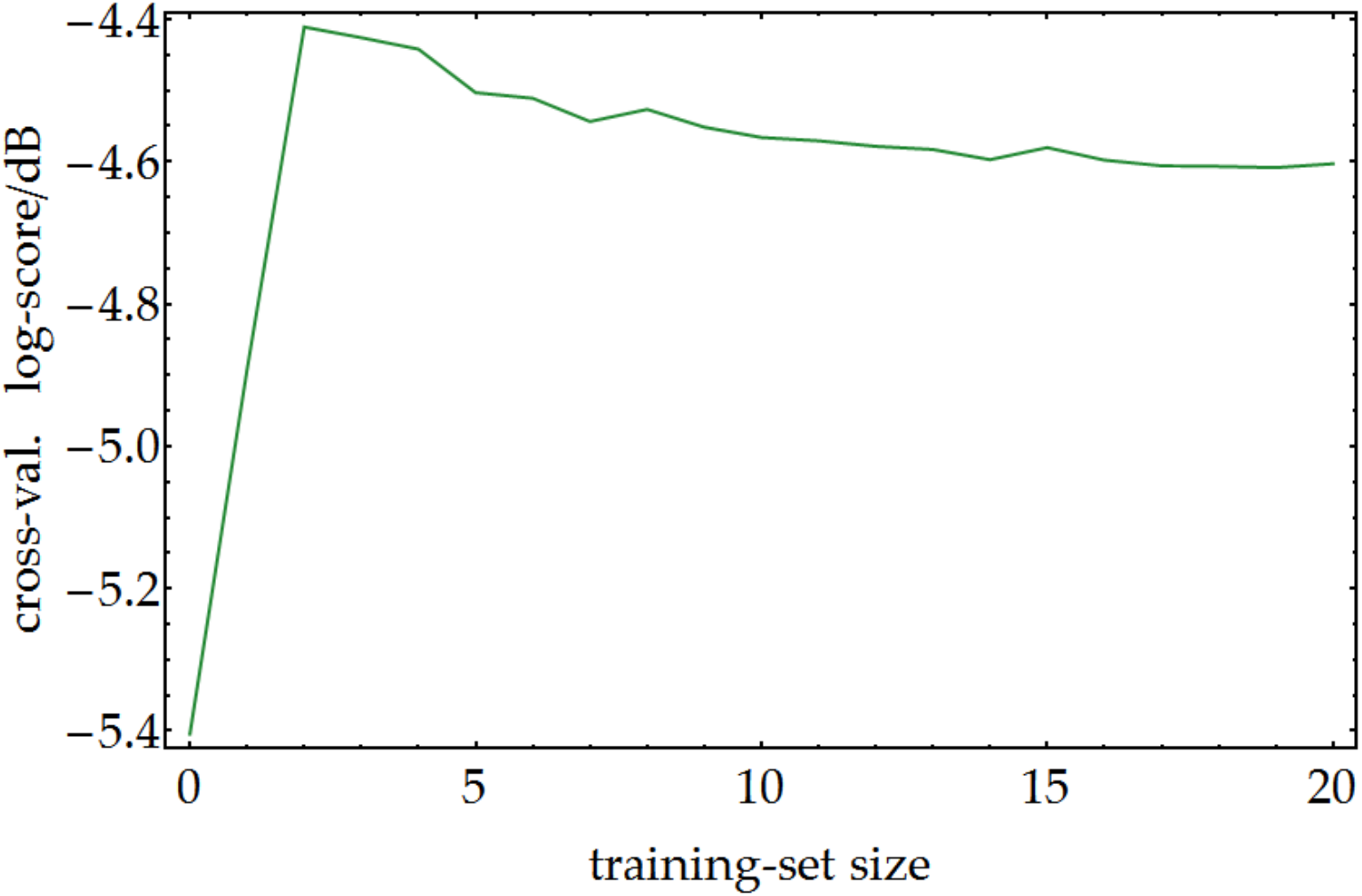
Let's cross-validate!



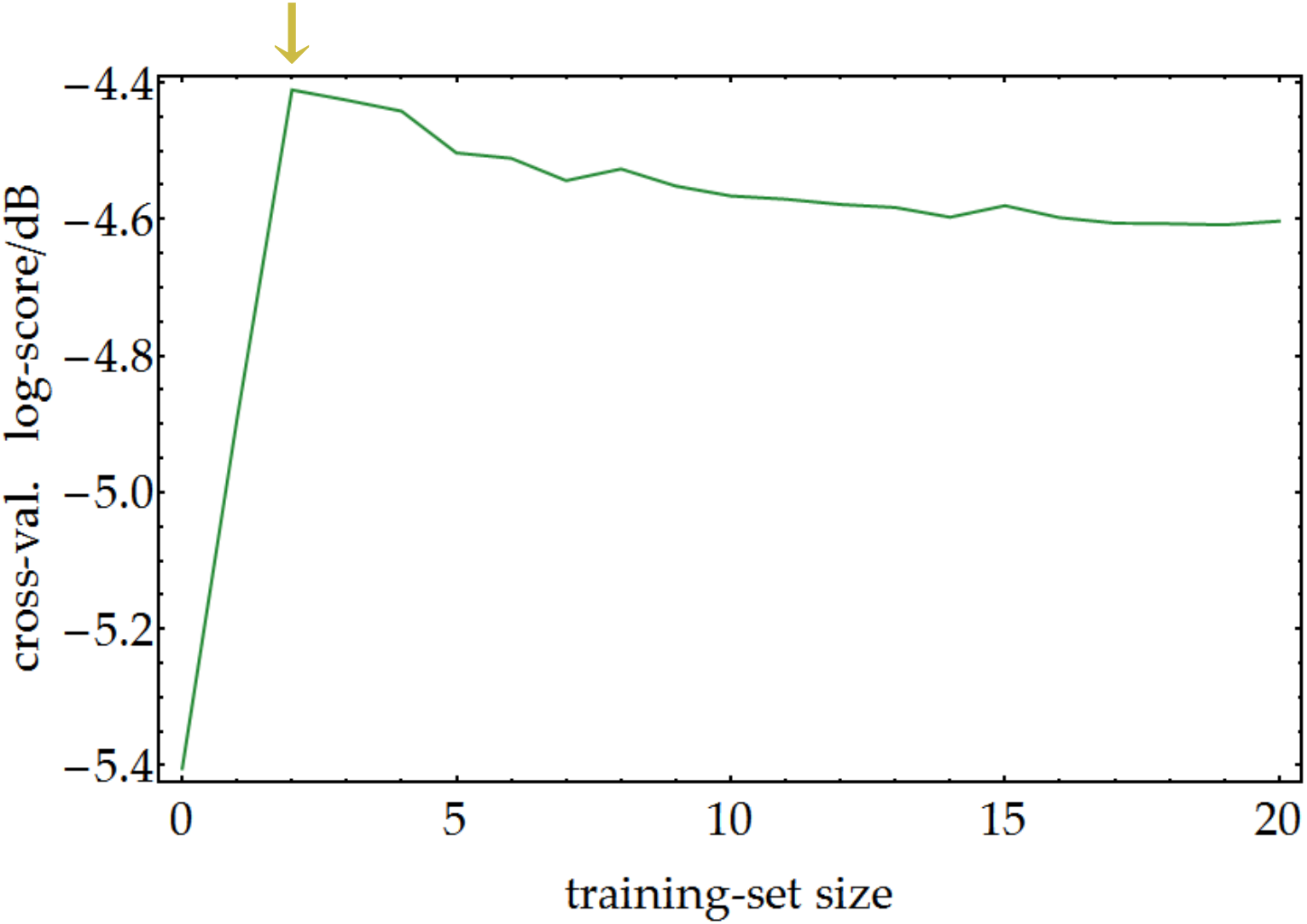
$$\frac{1}{n''} \sum_i \ln \mathrm{p}(x''_i \mid \overline{x'}^{(k)}\ M\ I)$$

cross-validation log-score

Let's cross-validate!



Let's cross-validate!



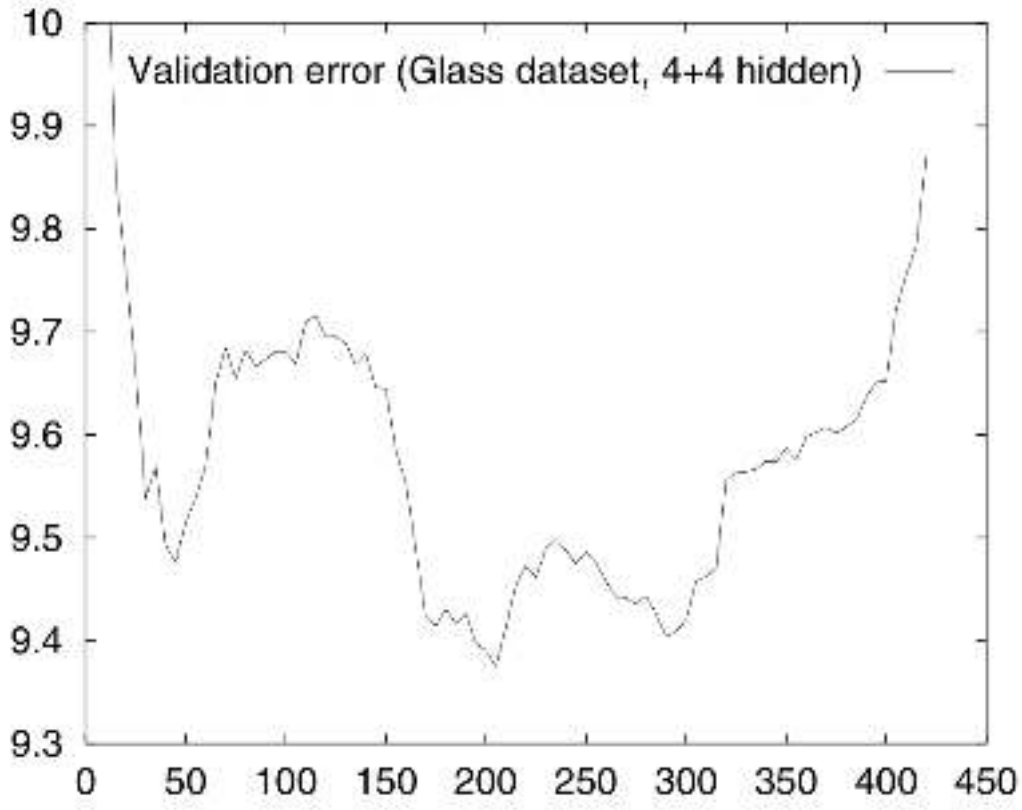


Fig. 2.2. A real validation error curve. Vertical: validation set error; horizontal: time (in training epochs).

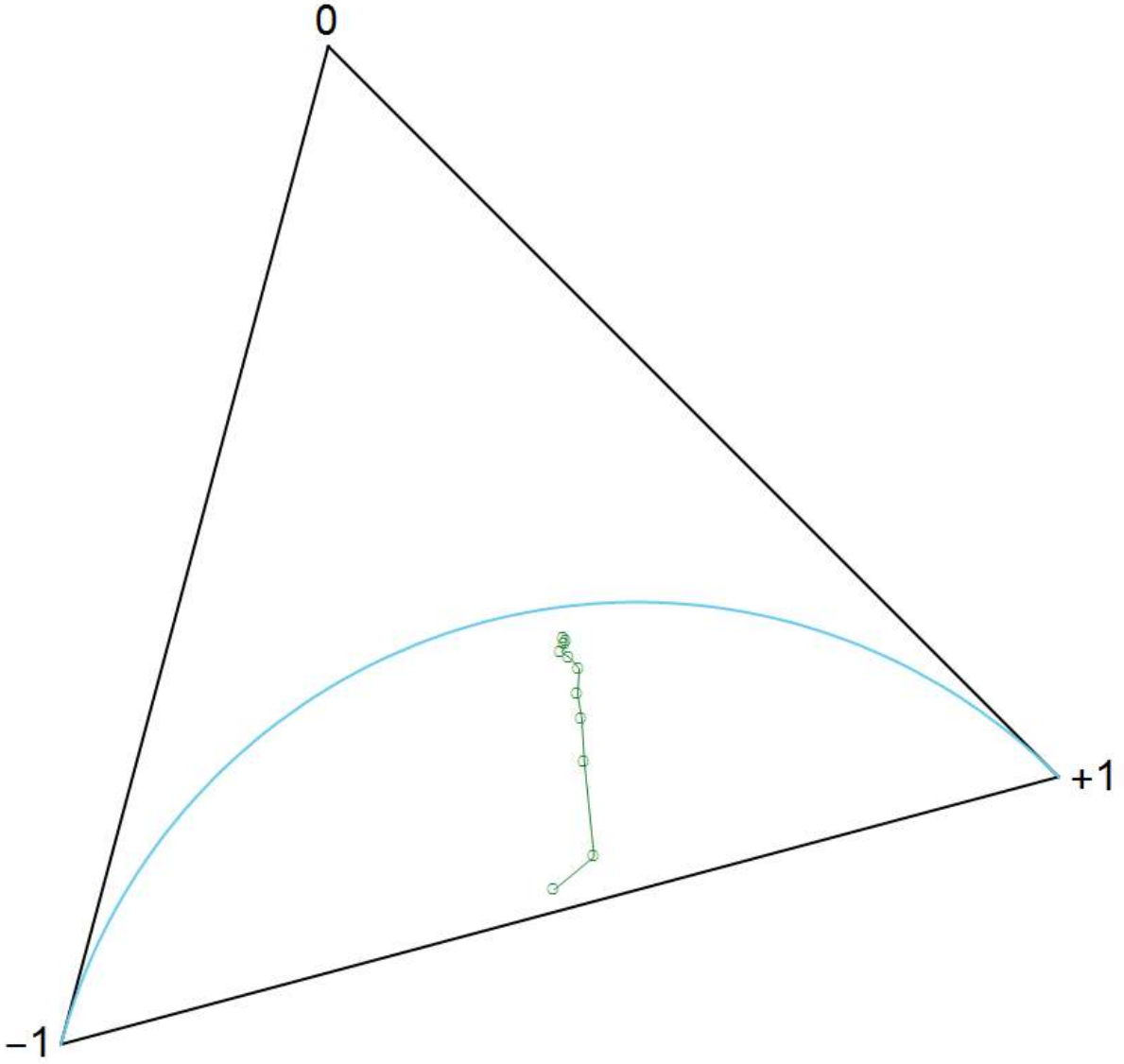
Grégoire Montavon
Geneviève B. Orr
Klaus-Robert Müller (Eds.)

Neural Networks:
Tricks of the Trade

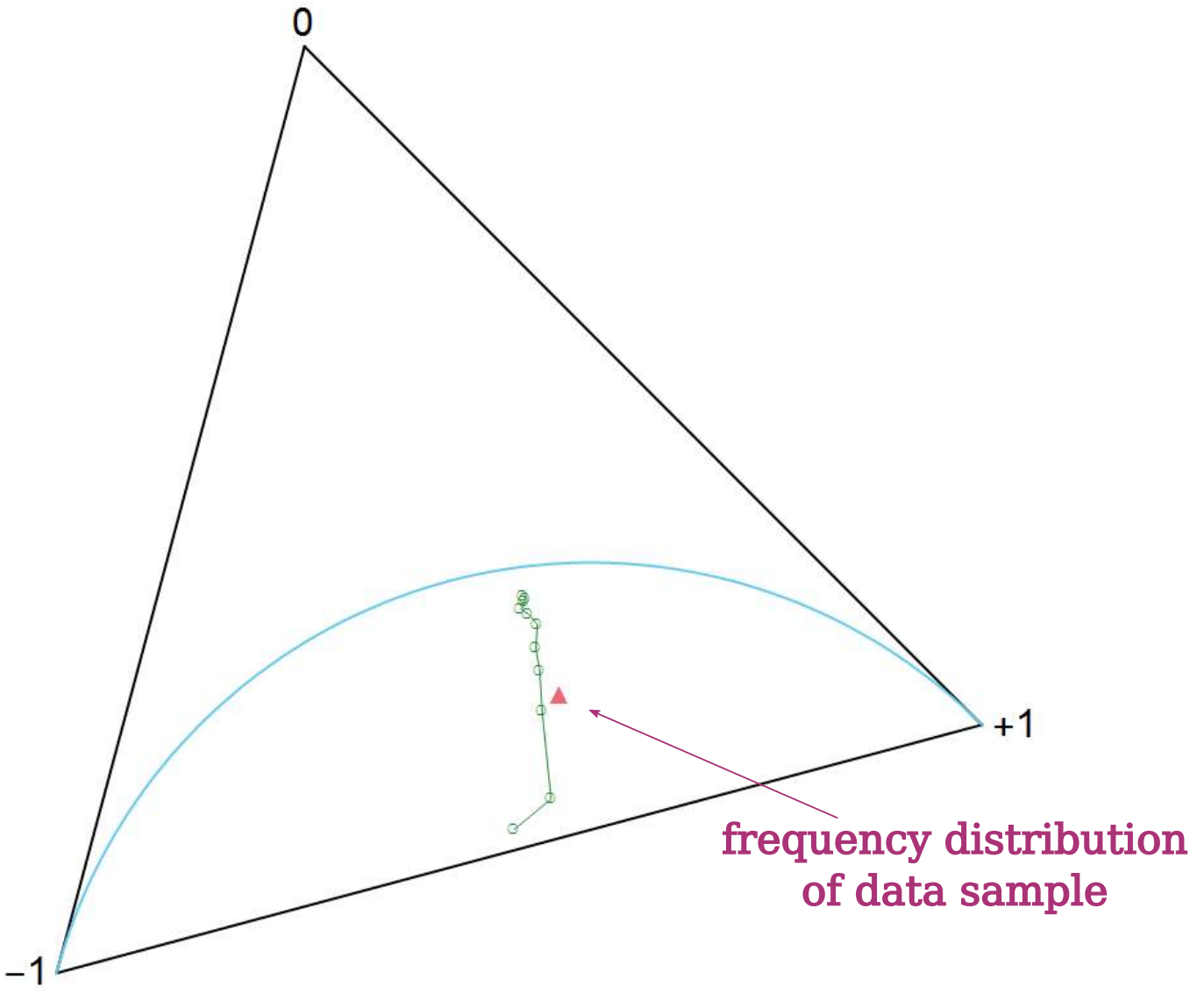
Second Edition

Why best log-score at 2 training points?

Why best log-score at 2 training points?



Why best log-score at 2 training points?



What does the cross-validation log-score actually measure?

$$\frac{1}{n''} \sum_i \ln \mathrm{p}(x''_i \mid \overline{x'}^{(k)}\ M\ I)$$

What does the cross-validation log-score actually measure?

$$\frac{1}{n''} \sum_i \ln \mathrm{p}(x''_i \mid \overline{x'}^{(k)}\ M\ I)$$

$$\mathrm{p}(x''_1 \mid \overline{x'}^{(k)}\ M\ I)\ \mathrm{p}(x''_2 \mid \overline{x'}^{(k)}\ M\ I) \cdots \mathrm{p}(x''_{n''} \mid \overline{x'}^{(k)}\ M\ I)$$

What does the cross-validation log-score actually measure?

$$\frac{1}{n''} \sum_i \ln \mathrm{p}(x''_i \mid \overline{x'}^{(k)}\ M\ I)$$

? $\mathrm{p}(x''_1 \mid \overline{x'}^{(k)}\ M\ I)\ \mathrm{p}(x''_2 \mid \overline{x'}^{(k)}\ M\ I) \cdots \mathrm{p}(x''_{n''} \mid \overline{x'}^{(k)}\ M\ I)$?

this can't be derived by the probability rules in our context

What does the cross-validation log-score actually measure?

$$\frac{1}{n''} \sum_i \ln \mathrm{p}(x''_i \mid \overline{x'}^{(k)}\ M\ I)$$

? $\mathrm{p}(x''_1 \mid \overline{x'}^{(k)}\ M\ I)\ \mathrm{p}(x''_2 \mid \overline{x'}^{(k)}\ M\ I) \cdots \mathrm{p}(x''_{n''} \mid \overline{x'}^{(k)}\ M\ I)$?

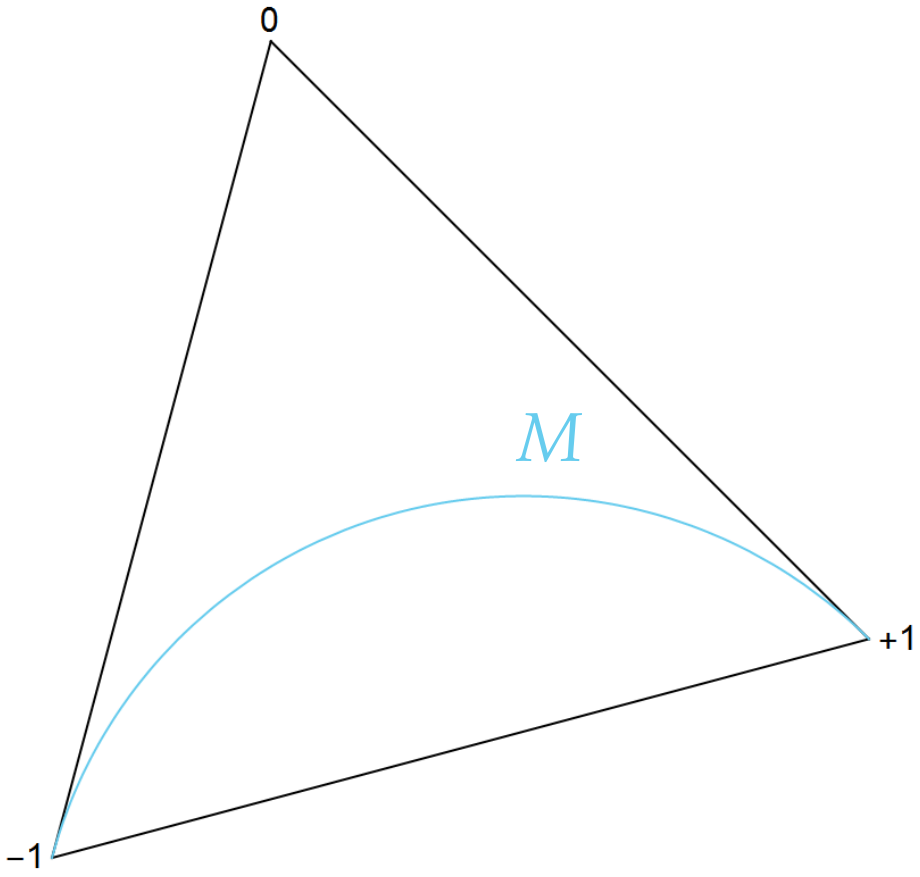
this can't be derived by the probability rules in our context

$$\mathrm{p}(\overline{x''} \mid \overline{x'}^{(k)}\ M\ I) =$$

$$\mathrm{p}(x''_1 \mid \overline{x'}^{(k)}\ M\ I)\ \mathrm{p}(x''_2 \mid x''_1, \overline{x'}^{(k)}\ M\ I) \cdots \mathrm{p}(x''_{n''} \mid x''_1, \ldots, x''_{n''-1}, \overline{x'}^{(k)}\ M\ I)$$

What does the cross-validation log-score actually measure?

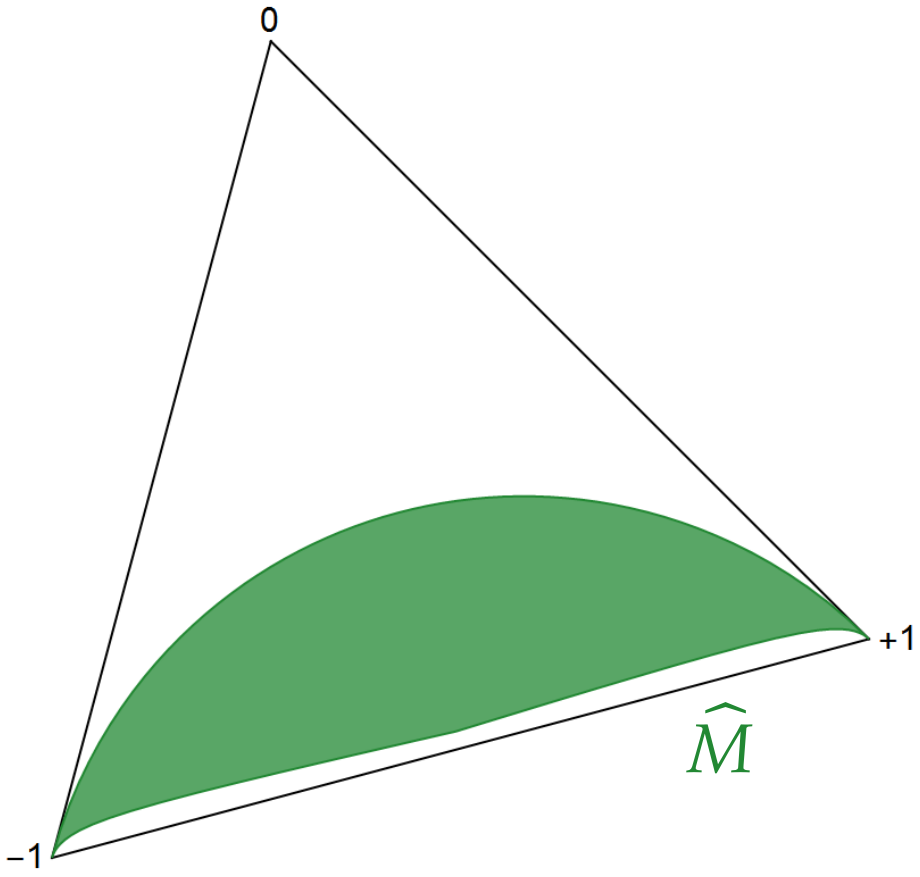
$$?\prod_i \mathrm{p}(x_i'' \mid \overline{x'}^{(k)}\ M\ I)$$



What does the cross-validation log-score actually measure?

$$?\prod_i \mathrm{p}(x_i'' \mid \overline{x'}^{(k)}\ M\ I)$$

$$\prod_i \mathrm{p}(x_i'' \mid \boldsymbol{F}^{(k)}\ \widehat{M}\ I)$$

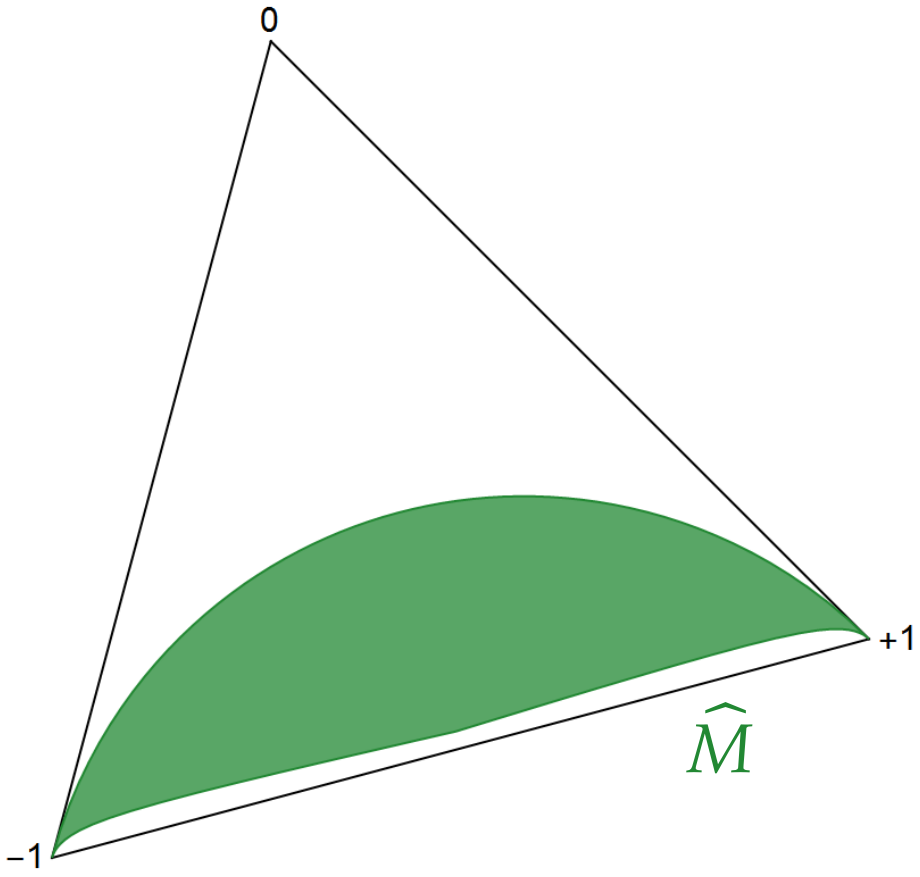


What does the cross-validation log-score actually measure?

$$?\prod_i \mathrm{p}(x_i'' \mid \overline{x'}^{(k)}\ M\ I)$$

$$\prod_i \mathrm{p}(x_i'' \mid \boldsymbol{F}^{(k)}\ \widehat{M}\ I)$$

$$= \mathrm{p}(\overline{x''} \mid \boldsymbol{F}^{(k)}\ \widehat{M}\ I)$$

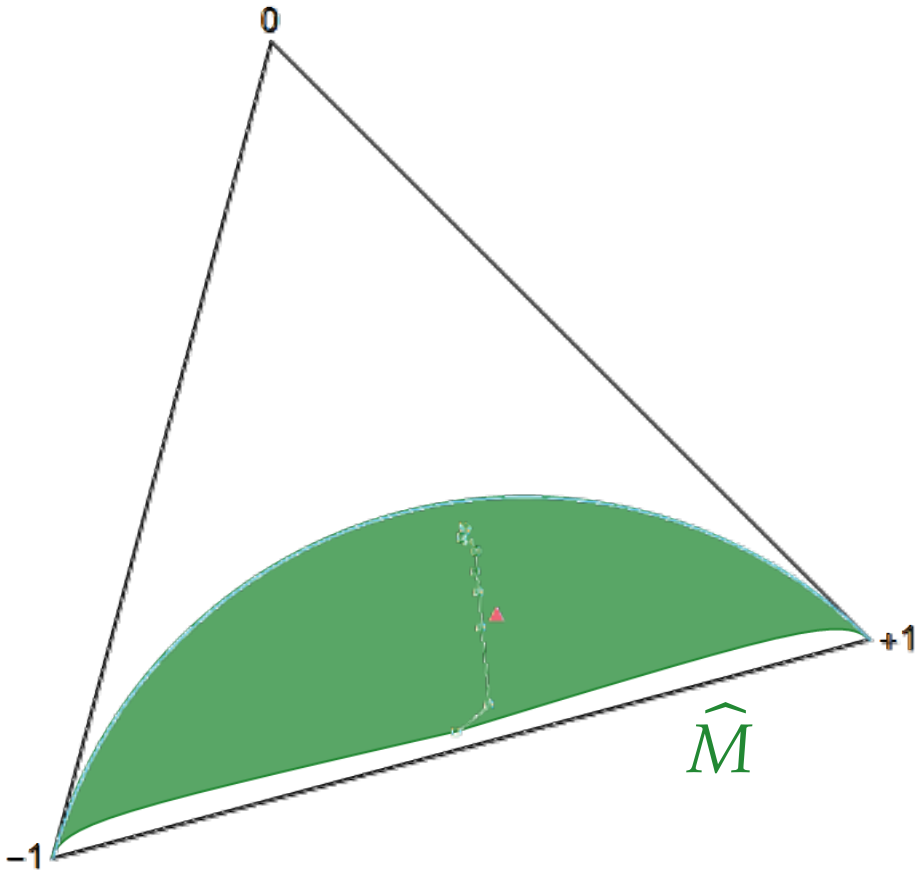


What does the cross-validation log-score actually measure?

$$?\prod_i \mathrm{p}(x_i'' \mid \overline{x'}^{(k)}\ M\ I)$$

$$\prod_i \mathrm{p}(x_i'' \mid \boldsymbol{F}^{(k)}\ \widehat{M}\ I)$$

$$= \mathrm{p}(\overline{x''} \mid \boldsymbol{F}^{(k)}\ \widehat{M}\ I)$$



Some conclusions about this kind of cross-validation

Some conclusions about this kind of cross-validation

The log-score is not assessing the model M

Some conclusions about this kind of cross-validation

The log-score is not assessing the model M

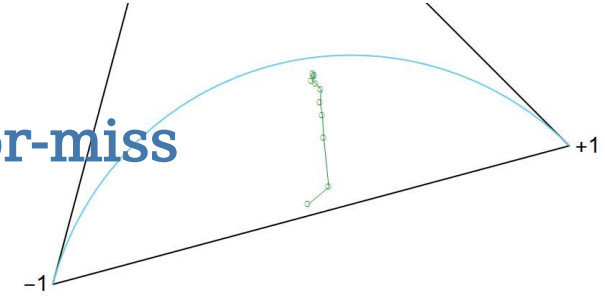
Log-score = *evidence* P(data | model) for several non-learning models $E^{(k)}, \widehat{M}$

Some conclusions about this kind of cross-validation

The log-score is not assessing the model M

Log-score = *evidence* P(data | model) for several non-learning models $E^{(k)}, \widehat{M}$

Log-score is very hit-or-miss

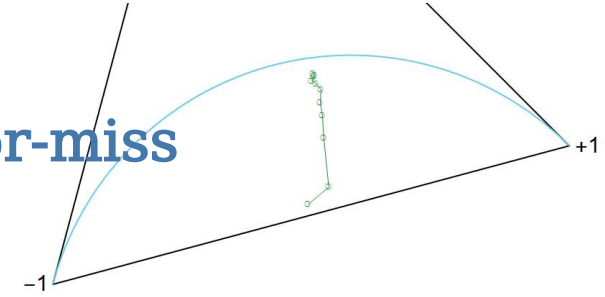


Some conclusions about this kind of cross-validation

The log-score is not assessing the model $\mathcal{M}$

Log-score = *evidence* P(data | model) for several non-learning mode $E^{(k)}\widehat{\mathcal{M}}$

Log-score is very hit-or-miss



→ More methodical approach: use model $\widehat{\mathcal{M}}$ $\in$ convex hull($\mathcal{M}$)

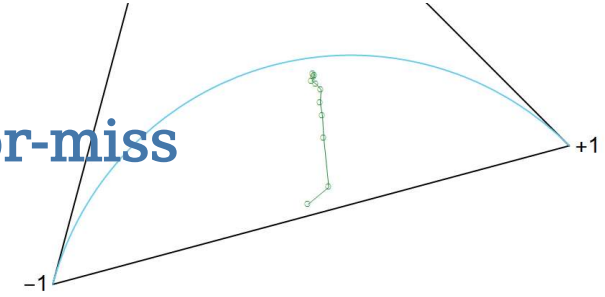
(or just go non-parametrics!)

Some conclusions about this kind of cross-validation

The log-score is not assessing the model $\mathcal{M}$

Log-score = *evidence* P(data | model) for several non-learning mode $E^{(k)}\widehat{\mathcal{M}}$

Log-score is very hit-or-miss



→ More methodical approach: use model $\widehat{\mathcal{M}}$ ∈ convex hull($\mathcal{M}$)
(or just go non-parametrics!)

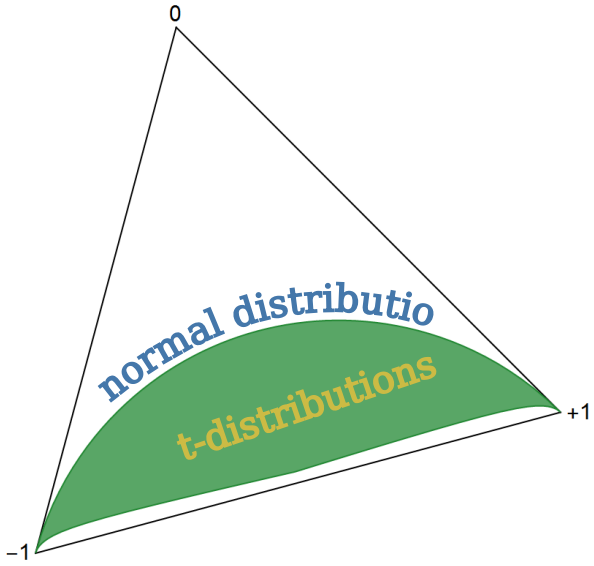
y stopping prevents ***underfitting***, not overfitting

More realistic example: normal model + normal-inverse-Wishart prior

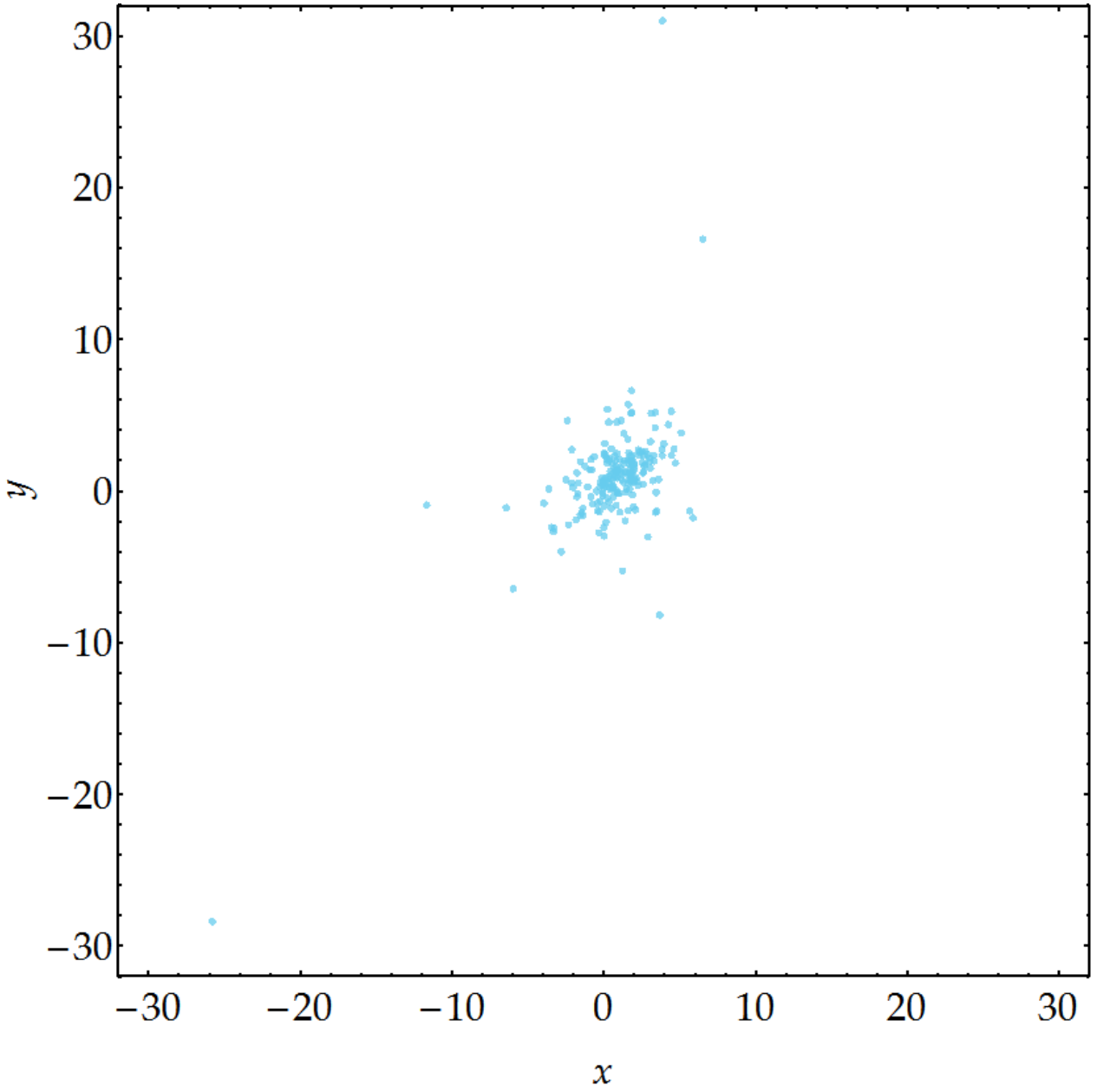
$$\mathrm{p}(x \mid N\ I) = \int \mathrm{d}\mu\, \mathrm{d}\sigma\, \mathrm{N}(x \mid \mu, \sigma)\ \mathrm{NiW}(\mu, \sigma \mid \zeta)$$

More realistic example: normal model + normal-inverse-Wishart prior

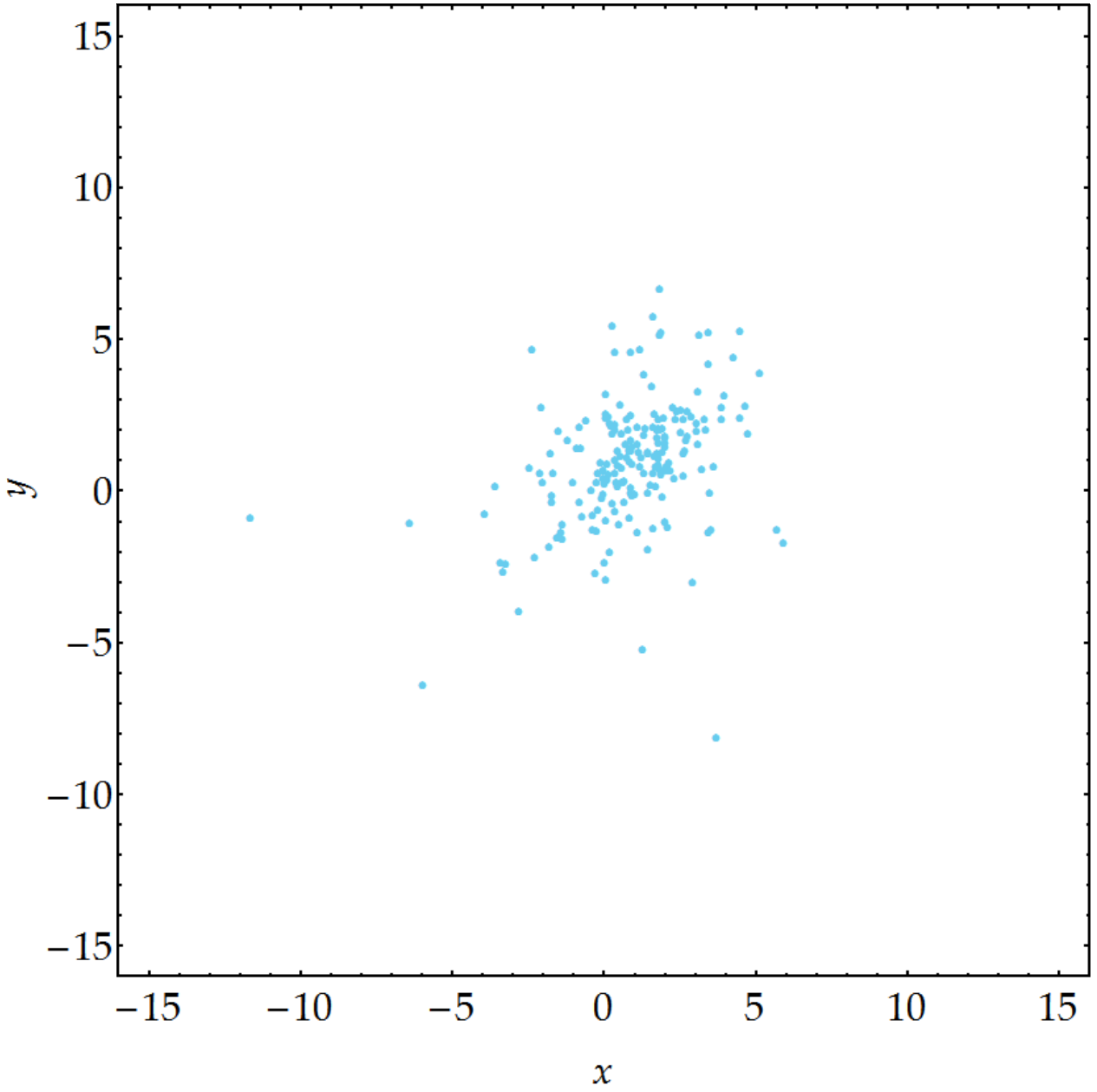
$$\mathrm{p}(x \mid N\ I) = \int \mathrm{d}\mu\, \mathrm{d}\sigma\ \mathrm{N}(x \mid \mu, \sigma)\ \mathrm{NiW}(\mu, \sigma \mid \zeta) \equiv \mathrm{t}(x \mid \ldots)$$



More realistic example: normal model + normal-inverse-Wishart prior



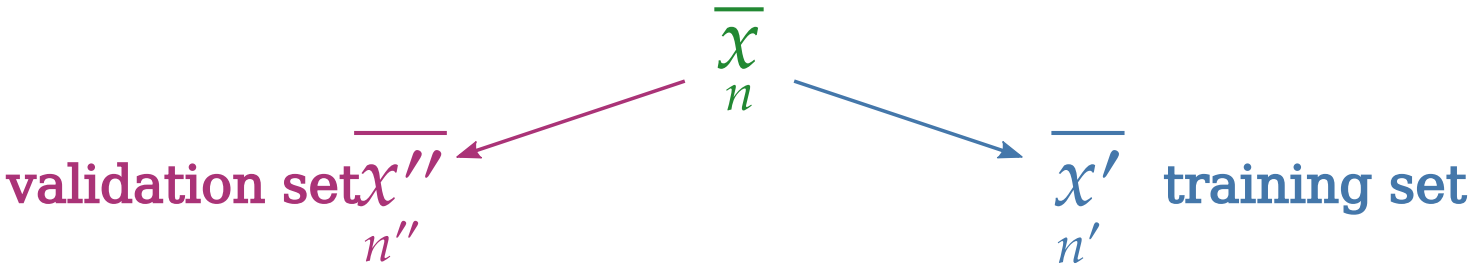
More realistic example: normal model + normal-inverse-Wishart prior



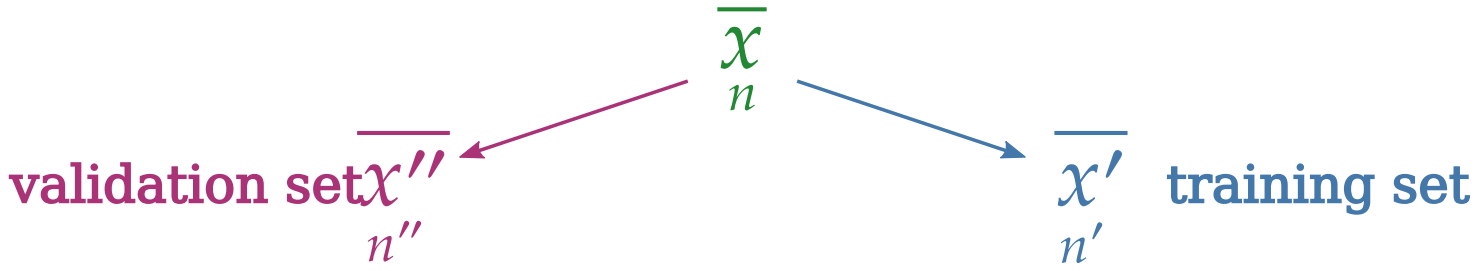
Let's cross-validate!

$$\overline{x}_n$$

Let's cross-validate!



Let's cross-validate!

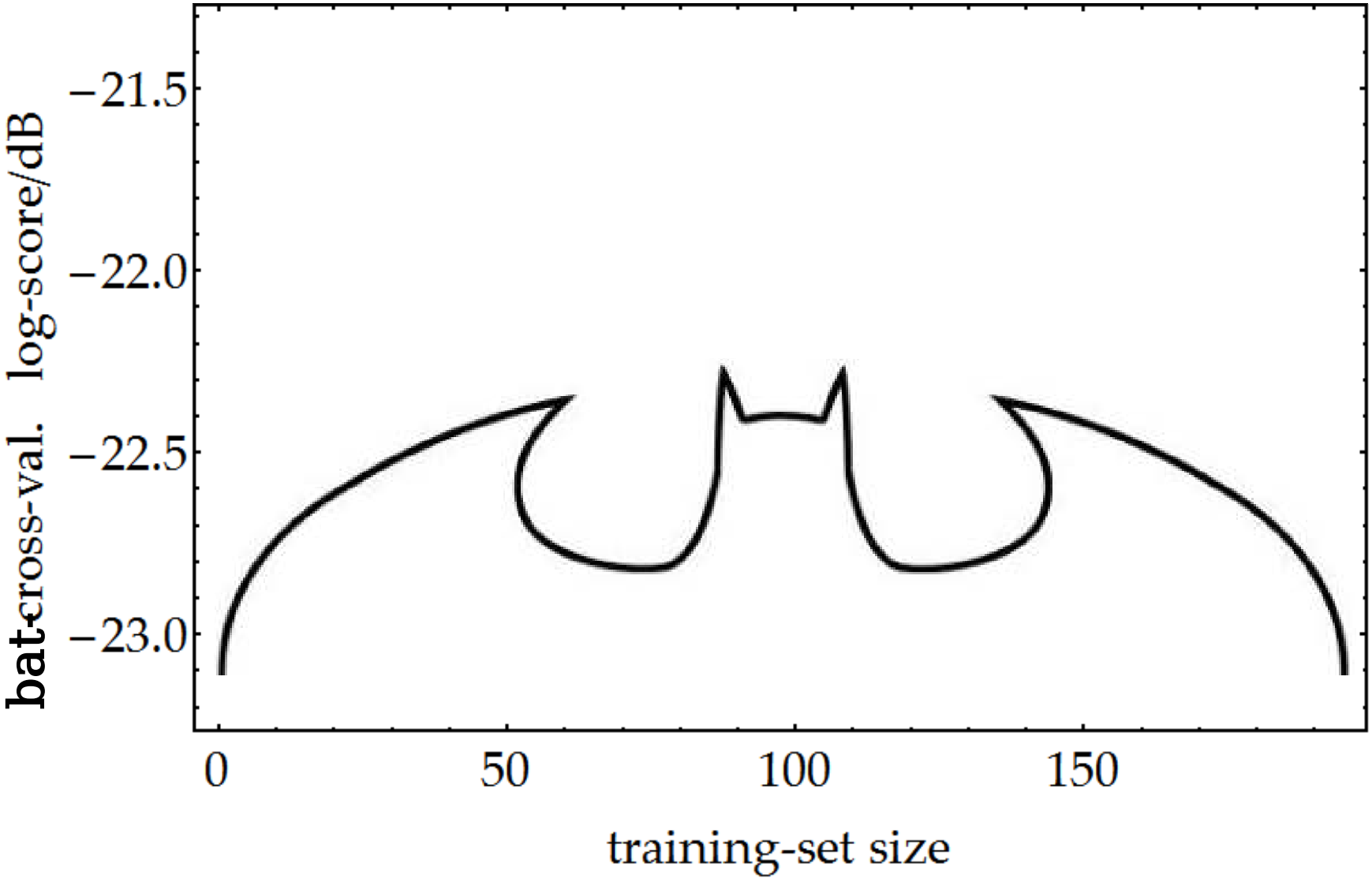


n'-fold cross-validation log-score

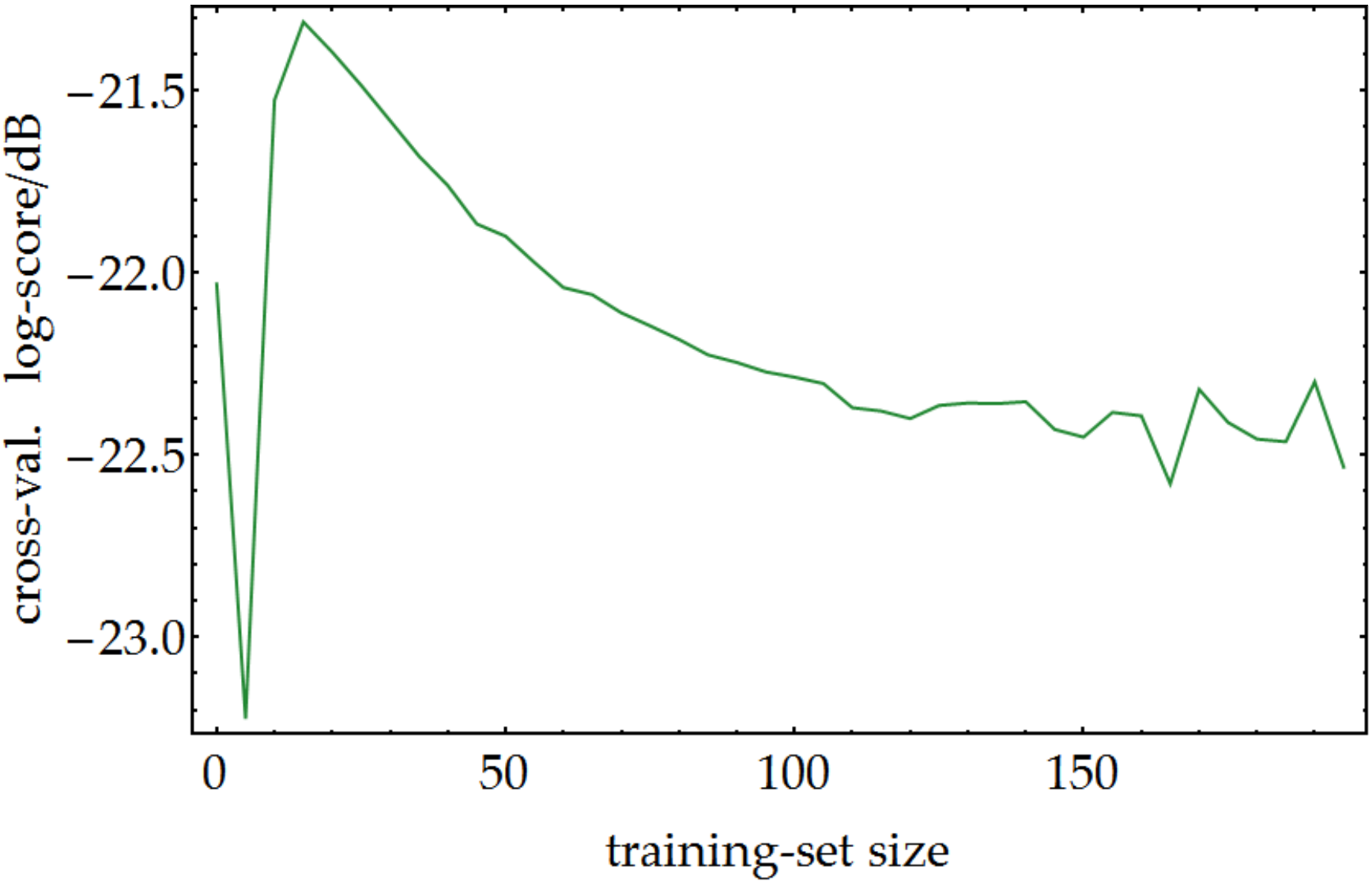
$$\frac{1}{\binom{n}{n'}} \sum_{\overline{x''}, \overline{x'}} \frac{1}{n''} \sum_{i} \ln \mathrm{p}(x''_i \mid \overline{x'}\ N\ I)$$

all possible (n'', n') partitions

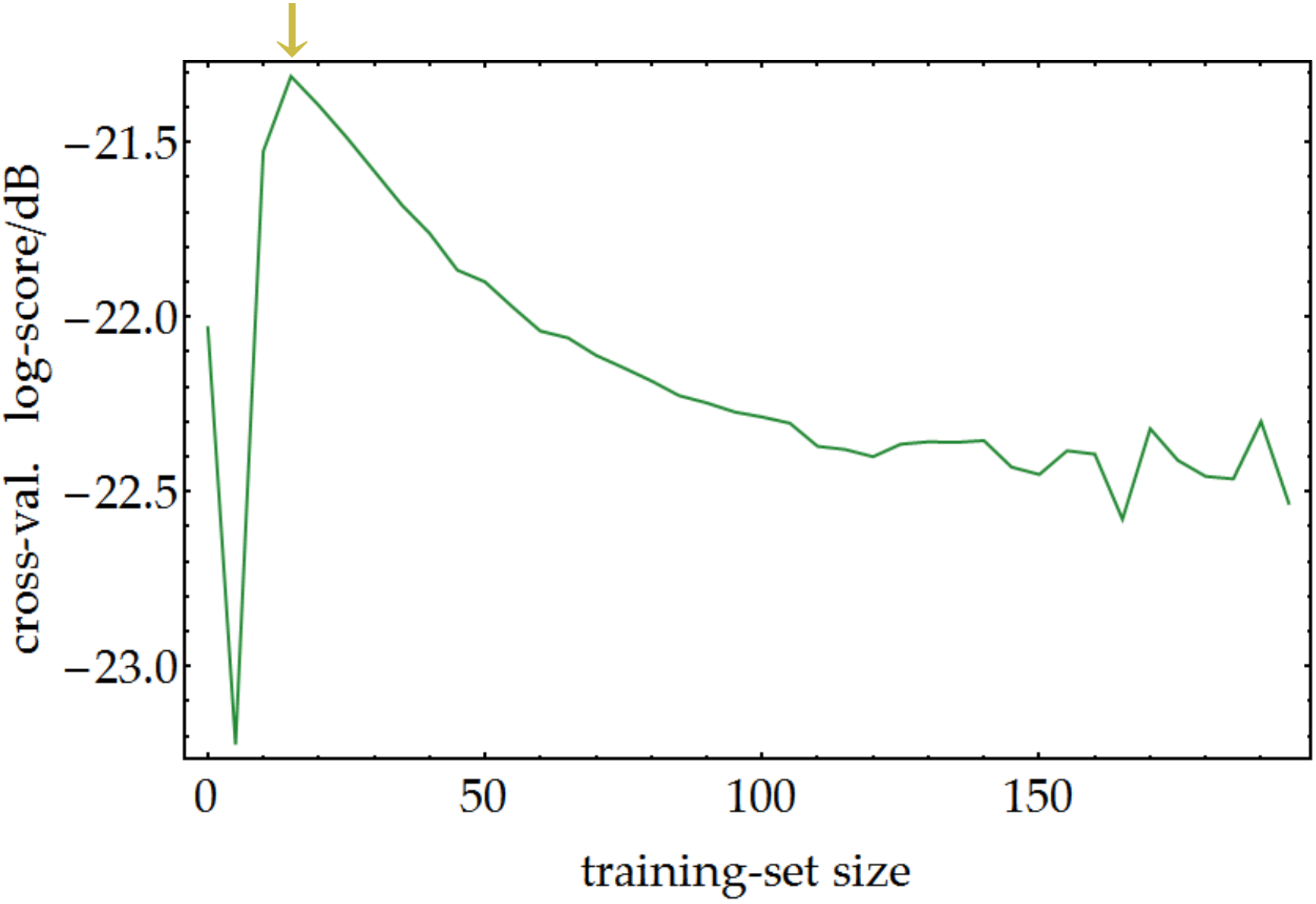
Let's cross-validate!



Let's cross-validate!



Let's cross-validate!



Cross-validation: a Bayesian perspective?

parametric model M with sufficient statistics $s(x)$

$$\mathrm{p}(x \mid \theta, M\ I) = g(x)\ \exp[\theta\ s(x) - \ln Z(\theta)]$$

Cross-validation: a Bayesian perspective?

parametric model M with sufficient statistics $s(x)$

$$\mathrm{p}(x \mid \theta, M\ I) = g(x)\ \exp[\theta\ s(x) - \ln Z(\theta)]$$

cross-validation = hit-or-miss Bayesian analysis of different parametric model $\widehat{M}$

$$\mathrm{p}(x \mid n, \sigma, \widehat{M}\ I) = \frac{\int \mathrm{d}\theta\ g(x)\ \exp[\theta\ s(x) + n\,\sigma\theta - (n+1)\ln Z(\theta)]\ f(\theta)}{\int \mathrm{d}\theta\ g(x)\ \exp[n\,\sigma\theta - n\ln Z(\theta)]\ f(\theta)}$$

Questions & thoughts

Questions & thoughts

Could we study the larger model more systematically instead?

Questions & thoughts

Could we study the larger model more systematically instead?

Would that be worthwhile?

Questions & thoughts

Could we study the larger model more systematically instead?

Would that be worthwhile?

Is cross-validation worthwhile? (it's only exploring a slightly larger model)

Questions & thoughts

Could we study the larger model more systematically instead?

Would that be worthwhile?

Is cross-validation worthwhile? (it's only exploring a slightly larger model)

How about the machine-learning case? – very different!

Thank you!