TODAY

- The general refined MDL principle
  - model selection with > 2 models
  - parameter estimation
  - prediction
  - if NML is undefined
- Generic solution for undefined NML/Jeffreys prior
- Some Final Points

Comparing finitely many models

- Let $\mathcal{M}_1, \ldots \mathcal{M}_K$ be the list of candidate models. MDL selects

$$\arg\min_{i=1..K} - \log P_{\text{nml}}(x^n|\mathcal{M}_i)$$

- Reinterpretation: MDL selects model minimizing the total two-part code length for the data, where data are encoded by (1) uniform code for the model and (2) optimal universal code for the data given the model

$$\arg\min_{i=1..K} - \log P_{\text{nml}}(x^n|\mathcal{M}_i) + L(i)$$

- Here for $i = 1..k, L(i) = \log K$

Comparing infinitely many models

- Select $\arg \min_{i \in \{1,2,\ldots\}} - \log P_{\text{nml}}(x^n | \mathcal{M}_i) + L(i)$
- where now $L(i)$ is the length of some code for *all* the integers, e.g. $L(i) = \log i + \log(i+1)$
- If we simply picked $\mathcal{M}_i$ minimizing $- \log P_{\text{nml}}(x^n | \mathcal{M}_i)$ then indeed, things might go wrong:
  - If all the $\mathcal{M}_i$ are singleton sets, then we may overfit no matter how large n (for example, each $\mathcal{M}_i$ is a Markov chain of some order; the list is such that all Markov chains with rational-valued parameter of each order is included)

General MDL Principle, part I

- Relative to the given set of candidate models,
  - you first devise a single code to encode all possible sequences,
    - This code will be “partly two-part, partly one-part”
  - you then do all inferences based on that code
- This ‘works’ to avoid overfitting
- This will give a coherent grand picture!

Comparing infinitely many models

- Better not use two-part code for the parameters
  - NML, Bayes give smaller regret
- We are *forced* to use two-part code for encoding model index
  - Because we want to select a model, we explicity have to encode it
  - Note: complexity of models *not* due to model index!

MDL for Parameter Estimation

- Now we have to use two-part code for the parameters as well
  - even though NML, Bayes give smaller regret
- Because we want to select parameters, we explicity have to encode them
- If we want to select both model structure (meta-structure) and parameters, we use a many-stage code (as in the first few weeks of the lecture)

$$L_1(\gamma) + L_2(\ddot{\theta}_\gamma) - \log p_{\ddot{\theta}_\gamma}(x^n)$$

MDL for Prediction

- If our goal is merely to predict future data given the past, we create one big universal **one-part** code over the union of all the models and use that to derive a predictive distribution. For example, using a Bayesian universal code:

$$\bar{p}(x^n) := \sum_{\gamma \in \Gamma} \pi(\gamma) \cdot \int_{\theta \in \Theta_\gamma} p_\theta(x^n) w_\gamma(\theta) d\theta$$

- Predict X_{n+1} given x^n using $\bar{p}(X_{n+1} \mid x^n)$
- …nothing needs to be 'selected', hence nothing needs to be encoded explicitly
- Why Bayes rather than NML? NML has horizon problem, see end of lecture

General MDL Principle, part II

- Relative to the given set of candidate models,
  - you first devise a single code to encode all possible sequences,
    - This code will be “partly two-part, partly one-part”
  - you then do all inferences based on that code
- Explicitly encode the things you want to select between (parameters or meta-parameters – model indices)
- Use 1-part codes conditional on these explicitly encoded objects

General MDL Principle, part III

- Use 1-part codes conditional on the explicitly encoded objects γ (e.g. model indices)
  - Ideally, this code should achieve minimax regret

$$\bar{p}_\gamma := \arg\min_p \max_{x^n} -\log p(x^n) - [-\log p_{\hat{\theta}_\gamma(x^n)}(x^n)]$$

- Explicitly encode objects of interest γ

$$L_{\text{2-p};\pi}(x^n) = \min_{\gamma \in \ddot{\Gamma}} -\log \pi(\gamma) - \log \bar{p}_\gamma(x^n) = -\log \pi(\ddot{\gamma}) - \log \bar{p}_{\ddot{\gamma}}(x^n)$$

- Again, aim for minimax regret "at a meta-level"

$$\arg\min_\pi L_{\text{2-p};\pi}(x^n) - [-\log \bar{p}_{\hat{\gamma}(x^n)}(x^n)]$$

- This idea explains both the use of the uniform prior on the model indices and discretization based on $I(\theta)$ in two-part codes

General MDL Principle, Part IV

- What if NML undefined (cannot do minimax regret)?
- What if number of models to be compared is “too large”?
- …use luckiness!
- a bit like a Bayesian prior, but still with worst-case sequence guarantees…

Comparing finite but ‘exponentially large’ number of models

- Suppose the model expresses

$$Y_i = \sum_{j=0}^{p} \beta_j X_{i,j} + \epsilon_i, \;\; \epsilon_1, \epsilon_2, \ldots \text{ i.i.d. } \sim N(0, \sigma^2)$$

…i.e. linear regression with normally distributed noise, with $p \geq n$.

- Now we want to learn which variables are ‘relevant, i.e. which $\beta_j \neq 0$.
- $2^p \geq 2^n$ models under consideration at sample size n.

What if NML distribution undefined?

- In **most** interesting applications, NML distribution undefined
  - Examples: linear regression, normal distribution: $\bar{p}_{\text{nml}}$ should have density

$$\bar{p}_{\mathsf{nml}}(x^n) = \frac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n}$$

  - Undefined since exponentiated complexity

$$\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n$$

diverges!

What if NML distribution/ Jeffreys Prior undefined?

- Typically (though there are exceptions):

$$\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n$$

diverges iff

$$\int_{\theta \in \Theta} \sqrt{\det I(\theta)} d\theta$$

diverges. Examples:

- Bernoulli/multinomial/Markov (both converge);
- normal/exponential/Gamma/Poisson/geometric (both diverge)

…so using Bayes with Jeffreys' prior does not save the day!

Solution that's better than just truncating parameter space

- Replace $\bar{p}_{\text{nml}}(x^n) = \frac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n}$

- by $\bar{p}_{\text{l-nml}}(x^n) = \frac{\sup_{\theta \in \Theta} p_\theta(x^n) v(\theta)}{\int_{y^n \in \mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(y^n) v(\theta) dy^n}$

…for some **luckiness function** v such that the denominator becomes finite: **Luckiness NML (in book: L-NML II).** $\bar{p}_{\mathrm{l-nml}}$ achieves minimax luckiness regret:

$$\min_p \max_{x^n} - \log p(x^n) - [\min_{\theta \in \Theta} \{- \log p_\theta(x^n) - \log v(\theta)\}]$$

General MDL Principle,

- Use mixed 1-part/2-part code based on achieving minimax regret at various levels
- If not possible…use luckiness

- But what luckiness function to use? For undefined NML/Jeffreys prior, there is sometimes an alternative, which gets closer to ‘real’ minimax regret

Conditional NML (-II)

Suppose $\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n < \infty$

Then codelength for $x_2, \ldots x_n$ given x_n is

$-\log \bar{p}_{\mathsf{nml}}(x_2, \ldots, x_n \mid x_1)$ where

$$\bar{p}_{\mathsf{nml}}(x_2, \ldots, x_n \mid x_1) =$$

$$\frac{\dfrac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n}}{\int_{z_2,\ldots,z_n \in \mathcal{Y}^{n-1}} \dfrac{p_{\hat{\theta}(x_1,z_2,\ldots,z_n)}(x_1,z_2\ldots,z_n) dz_2 \ldots dz_n}{\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n}}$$

$$= \frac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{z_2,\ldots,z_n \in \mathcal{Y}^{n-1}} p_{\hat{\theta}(x_1,z_2,\ldots,z_n)}(x_1, z_2 \ldots, z_n) dz_2 \ldots dz_n}$$

Conditional NML (-II)

Suppose $\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n < \infty$

Then codelength for $x_2, \ldots x_n$ given x_n is

$-\log \bar{p}_{\mathsf{nml}}(x_2, \ldots, x_n \mid x_1)$ where

$$\bar{p}_{\mathsf{nml}}(x_2, \ldots, x_n \mid x_1) =$$

$$= \frac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{z_2,\ldots,z_n \in \mathcal{Y}^{n-1}} p_{\hat{\theta}(x_1, z_2, \ldots, z_n)}(x_1, z_2 \ldots, z_n) dz_2 \ldots dz_n}$$

Conditional NML (-II)

Suppose $\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n < \infty$

Then codelength for $x_2, \ldots x_n$ given x_n is

$-\log \bar{p}_{\text{nml}}(x_2, \ldots, x_n \mid x_1)$ where

$$\bar{p}_{\text{nml}}(x_2, \ldots, x_n \mid x_1) =$$

$$= \frac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{z_2,\ldots,z_n \in \mathcal{Y}^{n-1}} p_{\hat{\theta}(x_1,z_2,\ldots,z_n)}(x_1, z_2 \ldots, z_n) dz_2 \ldots dz_n}$$

…but now suppose $\int_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n) dy^n = \infty$

Then often still

$$\int_{z_2,\ldots,z_n \in \mathcal{Y}^{n-1}} p_{\hat{\theta}(x_1,z_2,\ldots,z_n)}(x_1, z_2 \ldots, z_n) dz_2 \ldots dz_n < \infty$$

Conditional NML (-II)

We can often simply **define**

$$\bar{p}_{\text{nml}}(x_2, \ldots, x_n \mid x_1) := \frac{p_{\hat{\theta}(x^n)}(x^n)}{\int_{z_2,\ldots,z_n \in \mathcal{Y}^{n-1}} p_{\hat{\theta}(x_1,z_2,\ldots,z_n)}(x_1, z_2 \ldots, z_n) dz_2 \ldots dz_n}$$

- The first data point is a ‘start-up’ point (similarly to the situation for the prequential plug-in model)
- For special luckiness function v , L-NML can often be seen to be equivalent to conditional NML with a specific ‘start-up point’

Improper Priors

- Let $w(\theta)$ be the density of a measure on Θ .
- If $\int_{\theta \in \Theta} w(\theta) d\theta$ diverges, it is often called an '**improper** prior density
- Even with improper priors, it is often the case that "formal" Bayesian posterior

$$w(\theta \mid x^n) := \frac{p_\theta(x^n) w(\theta)}{\int_{\theta \in \Theta} p_\theta(x^n) w(\theta) d\theta}$$

is **proper** after all (i.e. it is a probability density).

Improper Priors

- Let $w(\theta)$ be the density of a measure on Θ .
- If $\int_{\theta \in \Theta} w(\theta) d\theta$ diverges, it is often called an '**improper** prior density
- Even with improper priors, it is often the case that "formal" Bayesian posterior

$$w(\theta \mid x^n) := \frac{p_\theta(x^n) w(\theta)}{\int_{\theta \in \Theta} p_\theta(x^n) w(\theta) d\theta}$$

is **proper** after all (i.e. it is a probability density).

Now Bayes' rule is an algorithm rather than a theorem

Improper Jeffreys Prior

- Let $w_{\mathrm{j}}(\theta) = \sqrt{\det I(\theta)}$
- Even if Jeffreys prior undefined (**improper**) it is often the case that $w_{\mathrm{j}}(\theta \mid x_1)$ is **proper**. In that case:
- Define $L_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1) =$

$$- \log \bar{p}_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1)$$

with

$$\bar{p}_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1) := \int p_\theta(x_2, \ldots, x_1) w_{\mathrm{j}}(\theta \mid x_1) d\theta$$

Improper Jeffreys Prior

- Let $w_{\text{j}}(\theta) = \sqrt{\det I(\theta)}$
- Even if Jeffreys prior undefined (**improper**) it is often the case that $w_{\text{j}}(\theta \mid x_1)$ is **proper**. In that case:
- Define $L_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1) =$

$$- \log \bar{p}_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1)$$

with

$$\bar{p}_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1) := \int p_\theta(x_2, \ldots, x_1) w_{\text{j}}(\theta \mid x_1) d\theta$$

Example: normal location family $N(\mu, \sigma^2)$ with fixed σ^2. Jeffreys' prior on μ is Lebesgue (uniform) measure (improper). Posterior after x_1 is normal with mean x_1 , variance σ^2.

Insight: the unconditional story generalizes to the conditional one

- The codelength obtained by conditional NML on $x_2, \ldots, x_n$ is asymptotically the same as the codelength obtained by Bayes-Jeffreys $x_2, \ldots, x_n$ conditioned on x_1 (i.e. Jeffreys' posterior based on x_1 is used as prior)

- Example: normal location family $N(\mu, \sigma^2)$ with fixed σ^2. Jeffreys' posterior after x_1 is normal with mean x_1 , variance σ^2. We have precisely (in this case not just asymptotically):

$$- \log \bar{p}_{\text{Bayes-Jeffreys}}(x_2, \ldots, x_n \mid x_1) = - \log \bar{p}_{\text{nml}}(x_2, \ldots, x_n \mid x_1)$$

Two Final Issues

Random Processes: NML vs Bayes/Prequential Plug-In

- For most models $\mathcal{M}$, the NML distribution based on $\mathcal{M}$ does not define a random process / probabilistic source:

$$\bar{p}^{(n)}{}_{\text{nml}}(x^n) = \frac{p_{\hat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{Y}^n} p_{\hat{\theta}(y^n)}(y^n)} \neq \sum_{y \in \mathcal{Y}} \bar{p}^{(n+1)}{}_{\text{nml}}(x^n, y)$$

- Prequential interpretation of NML only works if you know the horizon n_{final} in advance!
- Bayesian marginal distributions and prequential plug-in distributions *always* define a probabilistic source
  - predictive interpretation much cleaner than for NML

Computational Issues

- We **NEVER NEVER** have to do any real coding!
- Only the code**LENGTHS** matter!
- Calculating codelengths is far easier than actually encoding a sequence