Tutorial on Universal Prediction

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematisch Instituut – Universiteit Leiden

Universal Prediction

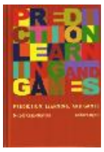
- Suppose data arrives sequentially in time.
- Let $\mathcal{M}$ be a set of predictors. There exist prediction strategies that, for each data sequence that can possibly be realized, predict essentially as well as the predictor in $\mathcal{M}$ that turns out to be best for that sequence 'with hindsight'

Provocation

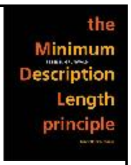
- Suppose data arrives sequentially in time.
- Let $\mathcal{M}$ be a set of predictors. There exist prediction strategies that, for each data sequence that can possibly be realized, predict essentially as well as the predictor in $\mathcal{M}$ that turns out to be best for that sequence 'with hindsight'
- **Hence, in some sense, it is possible to learn from data without making any assumptions at all about the data-generating process**

Universal Prediction

- Suppose data arrives sequentially in time.
- Let $\mathcal{M}$ be a set of predictors. There exist prediction strategies that, for each data sequence that can possibly be realized, predict essentially as well as the predictor in $\mathcal{M}$ that turns out to be best for that sequence 'with hindsight'
- Such prediction strategies are called universal (in individual-sequence rather than stochastic sense).
- Design of universal predictors is a central problem in information theory/machine learning theory



Where it Comes From



- Basic ideas reinvented independently many times...
  - Hannan (game theory, 1950s), Blackwell (statistics, 1950s), Rissanen, Shtarkov (information theory, 1980s), Vovk (probability theory, 1990), Warmuth and others (machine learning, 1990)
- ...but really took off in information theory only after Rissanen's MDL (minimum description length) papers (1980s) and in machine learning/game theory after the publication of Cesa-Bianchi and Lugosi (2006), Prediction, Learning and Games

Menu - Today

1. Universal Prediction
   with 'nice' scoring rules
2. Universal Prediction and Bayesian Inference
   - Complex Models

Menu – Tomorrow (2x)

1. Luckiness
   – “Objective Subjectivity”
2. Prediction with difficult loss functions
   – Vovk’s mixability, 0/1-loss, the Hedge Algorithm
3. Relations to Minimum Description Length
4. Relations to Kolmogorov Complexity KM(x)
   – Solomonoff prediction, superloss processes
5. Meta-Induction, Occam’s Razor

What You Will Learn

- MDL and Kolmogorov-complexity based approaches to inductive inference are **very different**
- Thinking in a completely different, nonstochastic way about Bayesian inference
  - Its remarkable (non)robustness properties
  - **Using Prior Distributions without Prior Assumptions**
- Occam’s Razor
  - A limited “simplicity bias” in inductive inference can be justified based on predictive considerations
- Kolmogorov complexity for ‘other loss functions’
- Time Permitting: Vovk-Shafer approach to probability founded on games rather than measures, avoiding ‘measure 0’ issues

Menu - Today

1. Universal Prediction
   with ‘nice’ scoring rules
2. Universal Prediction and Bayesian Inference
   - Complex Models

On-Line “Probabilistic” Prediction

- Consider sequence $(x_1,y_1),(x_2,y_2),\ldots$ where all $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
- Goal: sequentially predict y_i ,
  – given past $(x_1,y_1),\ldots,(x_{i-1},y_{i-1})$
  – using a ‘probabilistic prediction’ P_i (distribution on $\mathcal{Y}$)

On-Line Probabilistic Prediction

- Consider sequence $(x_1,y_1),(x_2,y_2),\ldots$ where all $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
- Goal: sequentially predict y_i ,
  – given past $(x_1,y_1),\ldots,(x_{i-1},y_{i-1})$
  – using a ‘probabilistic prediction’ P_i (distribution on $\mathcal{Y}$)
- Example: **weather forecaster**

$\mathcal{Y} = \{0,1\}$ (0 = no rain, 1 = rain)

$\mathcal{X} = \Big\{$ gigantic vector indicating humidity, air pressure temperature etc. at various locations

Prediction Strategies

- prediction strategy S is function mapping, for all i, histories $(x_1,y_1),\ldots,(x_{i-1},y_{i-1})$ to distributions for i -th outcome

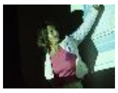
$$S : \cup_{n=1}^{\infty}(\mathcal{X}\times\mathcal{Y})^n \rightarrow \text{set of distributions on } \mathcal{Y}$$

Prediction Strategies

- prediction strategy S is function mapping, for all i, histories $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})$ to distributions for i -th outcome

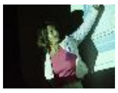
  $S : \cup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \rightarrow \text{set of distributions on } \mathcal{Y}$

- Weather forecasting example:
  - Prediction strategy is simply the prediction algorithm used by the weather forecaster, hopefully designed by meteorologists
  - Prediction for y_i will depend on data $(x_{i-1}, y_{i-1}), (x_{i-2}, y_{i-2}), \ldots$ observed on previous days



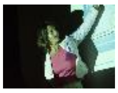
Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond (Dutch public TV)
  - Peter Timofeeff (Dutch commercial TV)
- On each i (day), Marjon and Peter announce the probability that $y_{i-1} = 1$,i.e. that it will rain on day $i+1$



Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond
  - Peter Timofeeff
- On each i (day), Marjon and Peter announce the probability that $y_{i-1} = 1$,i.e. that it will rain on day $i+1$
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence



Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond
  - Peter Timofeeff
- On each i (day), Marjon and Peter announce the probability that $y_{i-1} = 1$,i.e. that it will rain on day $i+1$
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence
  - If, with hindsight, Marjon was better, we predict as well as Marjon
  - If, with hindsight, Peter was better, we predict as well as Peter

Universal Prediction

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- Surprisingly, there exist prediction strategies that achieve this. These are called **universal**
  - "universal" is really a misnomer
- To formalize this idea, we need to define how we measure prediction quality
  - i.e., what do we mean by "the best forecaster"

Logarithmic Loss

- To compare performance of different prediction strategies, we need a measure of prediction quality
- THIS LECTURE, quality measured by log loss:

$$\text{loss}(y, P) := -\log_2 P(y)$$

$$\text{loss}(y_1 \ldots, y_n, S) := \sum_{i=1}^{n} \text{loss}(y_i, S(y_1, \ldots, y_{i-1}))$$

- corresponds to two important practical settings:
  - data compression: $\text{loss}(y_1 \ldots, y_n, S)$ is number of bits needed to encode $y_1, \ldots, y_n$ using code S
  - 'Kelly' gambling: loss related to log capital growth factor

Universal prediction with log loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{S}$ such that, for all $n, y_1, \ldots, y_n \in \{0,1\}^n$

$$\text{loss}(y_1 \ldots, y_n, \bar{S}) \leq$$
$$\min\{\text{loss}(y_1 \ldots, y_n, S_{\text{Marjon}}), \text{loss}(y_1 \ldots, y_n, S_{\text{Peter}})\} + 1.$$

Universal prediction with log loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{S}$ such that, for all $n, y_1, \ldots, y_n \in \{0,1\}^n$

$$\text{loss}(y_1 \ldots, y_n, \bar{S}) \leq$$
$$\min\{\text{loss}(y_1 \ldots, y_n, S_{\text{Marjon}}), \text{loss}(y_1 \ldots, y_n, S_{\text{Peter}})\} + 1.$$

- Losses increase linearly in n so this is very good!

$$\text{loss}(y_1 \ldots, y_n, S) := \sum_{i=1}^{n} \text{loss}(y_i, S(y_1, \ldots, y_{i-1}))$$

How to achieve universality

- How can we make sure that we always predict as well as the best forecaster?
  - Candidate Strategy 1 (“follow the leader”) :
    at each day n,
    - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
    - If Peter was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Peter

    …works reasonably well on most, but very bad on some sequences: there exist Peter and Marjon such that

$$\max_{y_1, \ldots, y_n} \{ \text{loss}(y_1, \ldots, y_n, \bar{S}) - \min_{S \in \text{Peter,Marjon}} \text{loss}(y_1, \ldots, y_n, S) \} = 0.25n$$

How to achieve universality

- FTL : at each day n,
  - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
  - If Peter was best on $y_1, \ldots, y_n$ then predict y_{n+1} like Peter
- Proposition: there exist Peter and Marjon such that

$$\max_{y_1, \ldots, y_n} \{ \text{loss}(y_1, \ldots, y_n, \bar{S}) - \min_{S \in \text{Peter,Marjon}} \text{loss}(y_1, \ldots, y_n, S) \} = 0.25n$$

How to achieve universality

- FTL : at each day n,
  - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
  - If Peter was best on $y_1, \ldots, y_n$ then predict y_{n+1} like Peter
- Proposition: there exist Peter and Marjon such that

$$\max_{y_1, \ldots, y_n} \{ \text{loss}(y_1, \ldots, y_n, \bar{S}) - \min_{S \in \text{Peter,Marjon}} \text{loss}(y_1, \ldots, y_n, S) \} = 0.25n$$

Each day Peter says ‘it rains with probability 1/4’
Marjon says ‘it rains with probability 3/4’

How to achieve universality

- FTL : at each day n,
  - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
  - If Peter was best on $y_1, \ldots, y_n$ then predict y_{n+1} like Peter
- Proposition: there exist Peter and Marjon such that

$$\max_{y_1, \ldots, y_n} \{ \text{loss}(y_1, \ldots, y_n, \bar{S}) - \min_{S \in \text{Peter,Marjon}} \text{loss}(y_1, \ldots, y_n, S) \} = 0.25n$$

Each day Peter says ‘it rains with probability 1/4’
Marjon says ‘it rains with probability 3/4’

0101010101010101(

How to achieve universality

- FTL : at each day n,
  - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
  - If Peter was best on $y_1, \ldots, y_n$ then predict y_{n+1} like Peter
- Proposition: there exist Peter and Marjon such that

$$\max_{y_1,\ldots,y_n} \{ \text{loss}(y_1,\ldots,y_n,\bar{\text{S}}) - \min_{S\in\text{Peter,Marjon}} \text{loss}(y_1,\ldots,y_n,S) \} = 0.25n$$

Cumulative Losses after day i , i odd:

$$\text{Peter:} \quad \frac{i-1}{2}\cdot\text{large}+\frac{i+1}{2}\cdot\text{small}$$

$$\text{Marjon:} \quad \frac{i+1}{2}\cdot\text{large}+\frac{i-1}{2}\cdot\text{small}$$

$$\text{large} = -\log\frac{1}{4} = 2 \quad ; \quad \text{small} = -\log\frac{3}{4} \approx 0.4$$

How to achieve universality

- FTL : at each day n,
  - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
  - If Peter was best on $y_1, \ldots, y_n$ then predict y_{n+1} like Peter
- Proposition: there exist Peter and Marjon such that

$$\max_{y_1,\ldots,y_n} \{ \text{loss}(y_1,\ldots,y_n,\bar{\text{S}}) - \min_{S\in\text{Peter,Marjon}} \text{loss}(y_1,\ldots,y_n,S) \} = 0.25n$$

Cumulative Losses after day i , i odd:

$$\text{Peter:} \quad \frac{i-1}{2}\cdot\text{large}+\frac{i+1}{2}\cdot\text{small}$$

$$\text{Marjon:} \quad \frac{i+1}{2}\cdot\text{large}+\frac{i-1}{2}\cdot\text{small}$$

$$\text{FTL:} \quad \approx\frac{3}{4}\cdot\text{large}+\frac{i}{4}\cdot\text{small}$$

How to achieve universality

- How can we make sure that we always predict as well as the best forecaster?
  - Candidate Strategy 1 ("follow the leader") :
    at each day n,
    - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
    - If Peter was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Peter

    …works reasonably well on most, but very bad on some sequences: there exist Peter and Marjon such that

$$\max_{y_1,\ldots,y_n} \{ \text{loss}(y_1,\ldots,y_n,\text{S}) - \min_{S\in\text{Peter,Marjon}} \text{loss}(y_1,\ldots,y_n,S) \} = 0.25n$$

How to achieve universality

- How can we make sure that we always predict as well as the best forecaster?
  - Candidate Strategy 1 ("follow the leader") :
    at each day n,
    - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
    - If Peter was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Peter

    …works reasonably well on most, but very bad on some sequences: there exist Peter and Marjon such that **regret** of using FTL increases linearly

How to achieve universality

- How can we make sure that we always predict as well as the best forecaster?
  - Candidate Strategy 1 ("follow the leader") :
    at each day n,
    - if Marjon was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Marjon
    - If Peter was best on $y_1, \ldots, y_n$, then predict y_{n+1} like Peter

    …works reasonably well on most, but very bad on some sequences: there exist Peter and Marjon such that **regret** of using FTL increases linearly

    …moreover, FTL fails dramatically if set of candidate predictors is infinite: **overfitting!**

Overfitting: The Main Problem of Machine Learning and Statistics

- World Cup Soccer 2010: Paul the Octopus predicts each game

Overfitting: The Main Problem of Machine Learning and Statistics

- World Cup Soccer 2010: Paul the Octopus predicts each game

- World Cup 2014: 100s of animals are predicting!

Menu

1. Universal Prediction
2. Universal Prediction and Bayes
   - Complex Models

On-Line Probabilistic Prediction

- Consider sequence $y_1, y_2, \ldots$, all $y_i \in \mathcal{Y}$
- Goal: sequentially predict y_i given past $y_1, \ldots, y_{i-1}$ using a 'probabilistic prediction' P_i (distribution on $\mathcal{Y}$)
- prediction strategy S is function mapping, for all i, 'histories' $y_1, \ldots, y_{i-1}$ to distributions for i -th outcome

$$S : \cup_{n=1}^{\infty} \mathcal{Y}^n \rightarrow \text{set of distributions on } \mathcal{Y}$$

prediction strategy = distribution

- If we think that $Y_1, \ldots, Y_n \sim P$ (not necessarily i.i.d !) then we should predict Y_i using the conditional distribution

$$P(\cdot \mid y^{i-1}) := P(Y_i = \cdot \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$$

- Conversely, every prediction strategy S may be thought of as a distribution on $(Y_1, \ldots, Y_n)$, by defining:

$$P(\cdot \mid y^{i-1}) := S(y^{i-1})$$
$$P(y_1, \ldots, y_n) := \prod_{i=1}^{n} P(y_i \mid y^{i-1})$$

prediction strategy = distribution

- If we think that $Y_1, \ldots, Y_n \sim P$ (not necessarily i.i.d !) then we should predict Y_i using the conditional distribution

$$P(\cdot \mid y^{i-1}) := P(Y_i = \cdot \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$$

- Conversely, every prediction strategy S may be thought of as a distribution on $(Y_1, \ldots, Y_n)$, by defining:

$$P(\cdot \mid y^{i-1}) := S(y^{i-1})$$
$$P(y_1, \ldots, y_n) := \prod_{i=1}^{n} P(y_i \mid y^{i-1}) = \prod_{i=1}^{n} \frac{P(y_1, \ldots, y_i)}{P(y_1, \ldots, y_{i-1})}$$
$$= \frac{P(y^n)}{P(y^{n-1})} \cdot \frac{P(y^{n-1})}{P(y^{n-2})} \cdot \frac{P(y^{n-2})}{P(y^{n-3})} \cdot \ldots P(y_1)$$

Log loss & likelihood

- For every "prediction strategy" P, all n,

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} - \log P(y_i \mid y^{i-1}) = - \log P(y_1, \ldots, y_n)$$

$$\uparrow$$

$$\sum_{i=1}^{n} - \log P(y_i \mid y^{i-1}) = - \log \prod_{i=1}^{n} P(y_i \mid y^{i-1}) = - \log \prod \frac{P(y_i)}{P(y^{i-1})} = - \log P(y_1, \ldots, y_n)$$

Log loss & likelihood

- For every “prediction strategy” P, all n,

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

Accumulated log loss = minus log likelihood

Dawid ‘84, Rissanen ‘84

Universal Prediction

- Let $\mathcal{M} = \{P_1, P_2, \ldots\}$ be a finite or countable set of predictors (identified with probability distributions on $\mathcal{Y}^\infty$)
  - Example: $\mathcal{M}$ is set of all Markov chains of each order with rational-valued parameters
- GOAL: given $\mathcal{M}$, construct a new predictor predicting data ‘essentially as well’ as any of the $P_\theta \in \mathcal{M}$

A Bayesian Strategy

- One possibility is to act Bayesian:
  1. Put some prior W on (parameter space of) $\mathcal{M}$
  2. Define Bayesian marginal distribution

$$P_{\text{Bayes}}(y_1, \ldots, y_n) := \sum_{\theta=1}^{\infty} P_\theta(y_1, \ldots, y_n) W(\theta)$$

  3. Predict with Bayesian (posterior) predictive distribution

$$P_{\text{Bayes}}(y_{i+1} \mid y_1, \ldots, y_i) = \frac{P_{\text{Bayes}}(y_1, \ldots, y_{i+1})}{P_{\text{Bayes}}(y_1, \ldots, y_i)}$$

why is this called ‘Bayesian’?

A Bayesian Strategy

- One possibility is to act Bayesian:
  1. Put some prior W on (parameter space of) $\mathcal{M}$
  2. Define Bayesian marginal distribution

$$P_{\text{Bayes}}(y_1, \ldots, y_n) := \sum_{\theta=1}^{\infty} P_\theta(y_1, \ldots, y_n) W(\theta)$$

  3. Predict with Bayesian (posterior) predictive distribution

$$P_{\text{Bayes}}(y_{i+1} \mid y_1, \ldots, y_i) = \frac{P_{\text{Bayes}}(y_1, \ldots, y_{i+1})}{P_{\text{Bayes}}(y_1, \ldots, y_i)}$$

why is this called ‘Bayesian’? …because we can write:

$$P_{\text{Bayes}}(y_{i+1} \mid y_1, \ldots, y_i) = \sum_{\theta=1}^{\infty} P_\theta(y_{i+1} \mid y^i) W(\theta \mid y^i)$$

$$W(\theta \mid y^i) = \frac{P_\theta(y^i) \cdot W(\theta)}{\sum_{\theta=1}^{\infty} P_\theta(y^i) W(\theta)}$$ is Bayes posterior!

Evaluating Bayes

- For arbitrary strategies P :

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

Evaluating Bayes

- For arbitrary strategies P :

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

- Moreover, for Bayes strategy P_{Bayes} ,for all n, y^n ,all θ_0:

$$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) = -\log P_{\text{Bayes}}(y_1, \ldots, y_n)$$

$$= -\log \sum_{\theta=1}^{\infty} P_\theta(y_1, \ldots, y_n) W(\theta) \le \underbrace{-\log P_{\theta_0}(y_1, \ldots, y_n)}_{\text{linear increase in } n} \underbrace{-\log W(\theta_0)}_{\text{constant in } n}$$

Bayesian strategy is universal

- For all n, y^n, all θ :
  $$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) \leq$$
  $$-\log P_\theta(y_1,\ldots,y_n)+C_\theta = \sum_{i=1}^{n} \text{loss}(y_i, P_\theta(\cdot \mid y^{i-1}))+C_\theta$$
- For all sequences of each length n, total loss of Bayes strategy bounded by constant depending on θ, not on n (Marjon vs. Peter: $w(\theta)=\frac{1}{2}, C_\theta = -\log w(\theta) = 1$)

Bayesian strategy is universal

- For all n, y^n, all θ :
  $$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) \leq$$
  $$-\log P_\theta(y_1,\ldots,y_n)+C_\theta = \sum_{i=1}^{n} \text{loss}(y_i, P_\theta(\cdot \mid y^{i-1}))+C_\theta$$
- For all sequences of each length n, total loss of Bayes strategy bounded by constant depending on θ, not on n
- So that average loss *per outcome*
  - either converges to loss of θ at rate at least $O(1/n)$
  - or becomes smaller than loss of θ for all large n

Bayesian strategy is universal

- We say: “a prediction strategy $\bar{P}$ is ‘universal’ “
  - relative to $\mathcal{M}$,
  - with respect to the log loss
  - in an individual sequence sence

  if for all $P \in \mathcal{M}$:
  $$\sup_{y^n \in \mathcal{Y}^n} \left\{ -\log \bar{P}(y^n) \underset{\boxed{\text{regret}}}{\underset{\uparrow}{-}} (-\log P(y^n)) \right\} = o(n)$$
- Clearly, Bayesian strategies are universal

Uncountable $\mathcal{M}$

- What if $\mathcal{M}$ uncountable, and ‘really big’?
  - e.g., $\mathcal{M}$ is set of all Markov chains of each order, not just with rational valued parameters!
  - as long as our priors are not too crazy, the Bayes strategy for the set of all MC’s with rational-valued parameters is still universal relative to $\mathcal{M}$
  - For example, we can construct priors such that for all k, all k-parameter Markov chains P, all n :

$$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) \leq \text{loss}(y_i, P(\cdot \mid y^{i-1}))+\frac{k}{2}\log n+O(1)$$

Uncountable $\mathcal{M}$

- What if $\mathcal{M}$ uncountable, and ‘really big’?
  - e.g., $\mathcal{M}$ is set of all Markov chains of each order, not just with rational valued parameters!
  - as long as our priors are not too crazy, the Bayes strategy for the set of all MC’s with rational-valued parameters is still universal relative to $\mathcal{M}$
  - For example, we can construct priors such that for all k, all k-parameter Markov chains P, all n :

$$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) \leq \text{loss}(y_i, P(\cdot \mid y^{i-1}))+\frac{k}{2}\log n+O(1)$$

We deal with overfitting!

Beyond Bayes…

- Bayesian strategy works remarkably well for universal prediction with respect to logarithmic loss…
  - that is, in coding and ‘freely mixable (Kelly) gambling games’
  - We also say ‘Bayesian strategy is a universal code’
- …but it is by no means the only good ‘universal’ strategy for log loss!
  - two-part codes/strategies
  - normalized maximum likelihood (Shtarkov) codes

‘nonstochastic statistics’

$$\sum_{i=1}^{n} \text{loss}(y_i, S_{\text{Bayes}}(i)) \le \sum_{i=1}^{n} \text{loss}(y_i, S_{\theta_0}(i)) - \log W(\theta_0)$$

- We made no assumptions whatsoever about the data generating mechanism
  - Not even that data are probabilistically generated
  - The ‘distributions’ in Θ were **only used as predictors**
- Still, no matter what sequence obtains, if some $\theta_0 \in \Theta$ suffers small log loss, then Bayes’ prediction suffers at most almost as small loss!

‘nonstochastic statistics’

$$\sum_{i=1}^{n} \text{loss}(y_i, S_{\text{Bayes}}(i)) \le \sum_{i=1}^{n} \text{loss}(y_i, S_{\theta_0}(i)) - \log W(\theta_0)$$

- We made no assumptions whatsoever about the data generating mechanism
  - Not even that data are probabilistically generated
  - The ‘distributions’ in Θ were **only used as predictors**
- Still, no matter what sequence obtains, if some $\theta_0 \in \Theta$ suffers small log loss, then Bayes’ prediction suffers at most almost as small loss!
  - idea invented independently by Hannan, Blackwell (1950s), Shtarkov (1988), Foster (1990s), others…

‘objective subjectivity’

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - “ $W(\theta)$ is your prior belief that the world is in state θ ”
  - Savage/De Finetti/Cox justification of Bayes: much more subtle, but assumptions still strong: decision maker has a complete preference order on acts (Why should she?)

‘objective subjectivity’

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - “ $W(\theta)$ is your prior belief that the world is in state θ ”
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term $$-\log W(\theta)$$ *no matter what data you will observe*

‘objective subjectivity’

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - “ $W(\theta)$ is your prior belief that the world is in state θ ”
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term $$-\log W(\theta)$$ *no matter what data you will observe*

‘objective subjectivity’

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - “ $W(\theta)$ is your prior belief that the world is in state θ ”
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term $$-\log W(\theta)$$ *no matter what data you will observe*

Allows you to use Bayesian methods without adopting the (still controversial) Bayesian philosophy

‘objective subjectivity’

- Bayesian statistics in its usual interpretation relies on heavy assumptions about the world:
  - “ $W(\theta)$ is your prior belief that the world is in state θ ”
- Our reinterpretation justifies Bayesian approaches in *some* situations with hardly *any* assumptions
  - With prior W, your predictions will never be worse than the predictions made by any θ up to luckiness term $$-\log W(\theta)$$ *no matter what data you will observe*

In other situations (loss functions) need to change Bayes a little

Relevance

- We analyzed one aspect of inductive inference (prediction) without making any distributional assumptions.
- Why is this relevant?
  1. Philosophical Reasons
  2. Practical Reasons

Philosophical Relevance

- People not immersed into statistics before a certain critical age often feel ill at ease when reading
  ‘assume $Y_1, Y_2, \ldots$ distributed according to some P, P being a member of some family $\mathcal{M}$ ’
- Does randomness exist in the real world?

Practical Relevance

- People not immersed into statistics before a certain critical age often feel ill at ease when reading
  ‘assume $Y_1, Y_2, \ldots$ distributed according to some P, P being a member of some family $\mathcal{M}$ ’
- Does randomness exist in our application?

  Is it possible to include Tolstoy’s ***War and Peace*** in a reasonable way into the set of ‘all possible novels’ and further to postulate the existence of a certain probability distribution in this set? Must we assume that the individual scenes in this book form a random sequence with stochastic relations that damp out quite rapidly over a distance of several pages?

  - A.N. Kolmogorov (1965)

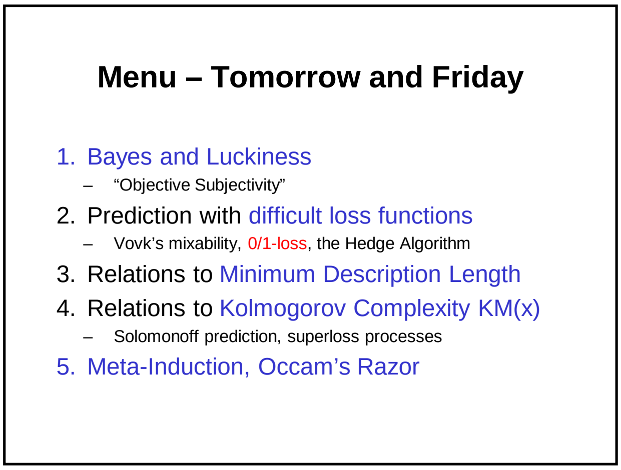
Practical Relevance

- People not immersed into statistics before a certain critical age often feel ill at ease when reading
  ‘assume $Y_1, Y_2, \ldots$ distributed according to some P, P being a member of some family $\mathcal{M}$ ’
- Does randomness exist in our application?
  - to be fair, I have applications in mind far beyond what was envisaged by the founding fathers of 20th century statistics:

Menu - Today

1. Universal Prediction

   with ‘nice’ scoring rules
2. Universal Prediction and Bayesian Inference
   - Complex Models



Menu – Tomorrow and Friday

1. Bayes and Luckiness
   - “Objective Subjectivity”
2. Prediction with difficult loss functions
   - Vovk’s mixability, 0/1-loss, the Hedge Algorithm
3. Relations to Minimum Description Length
4. Relations to Kolmogorov Complexity KM(x)
   - Solomonoff prediction, superloss processes
5. Meta-Induction, Occam’s Razor