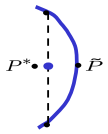
Learning from data when all models are wrong

Peter Grünwald
CWI / Leiden

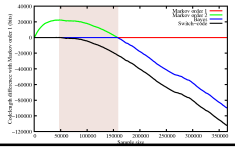
Joint work with John Langford, Tim van Erven, Steven de Rooij, Wouter Koolen, Wojciech Kotlowski

Menu – Two Pictures

1. Introduction
2. Learning when Models are **Seriously Wrong**

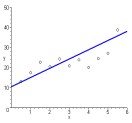


3. Learning when Models are **Almost True**



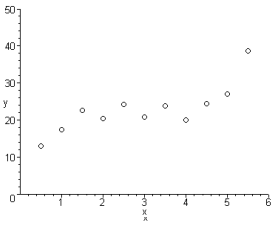
Wrong yet useful models

- Scientists routinely use models that are wrong…
  - nonlinear relations modeled as linear,
  - dependent variables modeled as independent, …
- …yet useful:
  - they lead to reasonable predictions

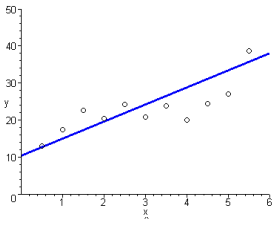


- Examples:
  - speech & digit recognition, DNA sequence analysis, …

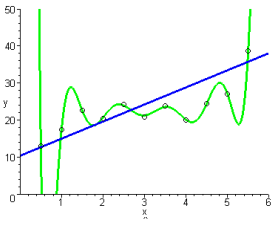
Example 1: Regression (Curve Fitting)



Regression

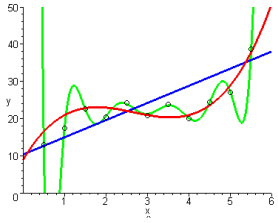


Regression

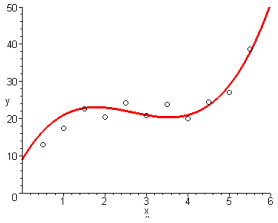


Overfitting!

Regression



Regression



Standard Setting

- "training data" $(x_1, y_1), \ldots, (x_n, y_n)$
- Usual assumption: independent identically distributed
  $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. $\sim P^*$
- Model or 'model class' $\mathcal{M}$: set of distributions $P(Y \mid X)$

Standard Setting

- "training data" $(x_1, y_1), \ldots, (x_n, y_n)$
- Usual assumption: independent identically distributed
  $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. $\sim P^*$ "true" distribution
- Model or 'model class' $\mathcal{M}$: set of distributions $P(Y \mid X)$

Polynomial Regression Example

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. $\sim P^*$

- Model class consists of parametric sub-models:
  $\mathcal{M} = \bigcup_{k \geq 1} \mathcal{M}_k \qquad \mathcal{M}_k = \{P_\mu \mid \mu \in \mathbb{R}^{k+1}\}$
- P_μ is conditional distribution of Y given X, expressing
  $Y = \sum_{j=0}^{k} \mu_j X^j + Z, Z \sim \text{Normal}(0, \sigma^2)$
  i.e. P_μ has density $f_\mu(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \sum_{j=0}^{k} \mu_j x^j)^2}{2\sigma^2}\right)$

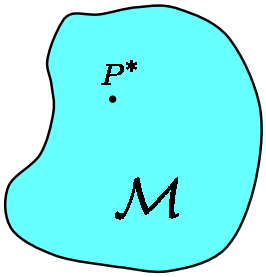
Standard Setting

- "training data" $(x_1, y_1), \ldots, (x_n, y_n)$
- Learning
  - based on training data, hypothesize an "estimate" $\hat{P} \in \mathcal{M}$ of P^*
- Learning Algorithm ("estimator") is function
  $\hat{P} : \bigcup_{n \geq 0} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}$
- Example Learning Algorithms:
  - (penalized) least squares, (penalized) maximum likelihood

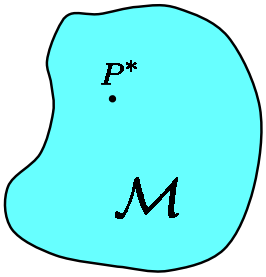
Standard Setting

- "training data" $(x_1, y_1), \ldots, (x_n, y_n)$
- Learning
  - based on training data, hypothesize an "estimate" $\hat{P} \in \mathcal{M}$ of P^*
- Prediction
  - Based on training data and X_{new} and $\mathcal{M}$, predict Y_{new}

1. "When the Model is True"



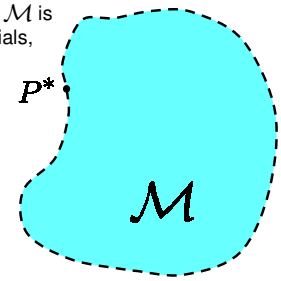
1. "When the Model is True"



e.g. 3rd degree polynomial, Gaussian noise

2. "When the Model is Almost True"

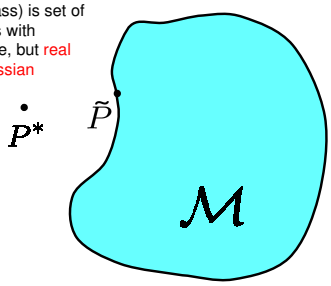
e.g. model (class) $\mathcal{M}$ is set of all polynomials, P^* is square root



$P^* \notin \mathcal{M}$, $\inf_{P \in \mathcal{M}} d(P^*, P) = 0$

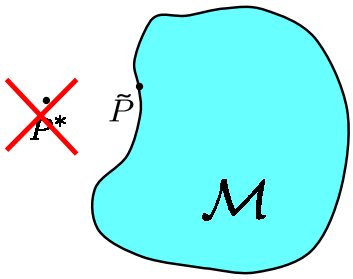
3. "When the Model is Wrong"

e.g. model (class) is set of all polynomials with Gaussian noise, but real noise not Gaussian



$P^* \notin \mathcal{M}$, $\inf_{P \in \mathcal{M}} d(P^*, P) > 0$

4. "When there is No Truth!"



4 Situations, 3 Goals

	Estimation, Model Selection	Prediction
model true	Bayes usually o.k.	Bayes usually o.k.
model almost true		
model wrong truth exists		
model wrong no truth		universal prediction Bayes ok for log-score, not for other scores

Theme 1: Bayes!

	Estimation, Model Selection	Prediction
model true	Bayes usually o.k.	Bayes usually o.k.
model almost true		
model wrong truth exists		Bayes ok for log-score, not for other scores
model wrong no truth		universal prediction Bayes ok for log-score, not for other scores

Theme 2: when the model is wrong, the distance measure becomes much more important

	Estimation, Model Selection	Prediction
model true		
model almost true		
model wrong truth exists	Bayes not ok in general unless model convex	Bayes ok for log-score, not for other scores
model wrong no truth		universal prediction Bayes ok for log-score, not for other scores

Theme 3: most existing results are from theoretical **machine learning** community (NIPS, COLT)

	Estimation, Model Selection	Prediction
model true		
model almost true	Bayes factor model selection suboptimal	Bayes factor model averaging suboptimal
model wrong truth exists	Bayes not ok in general unless model convex	
model wrong no truth		

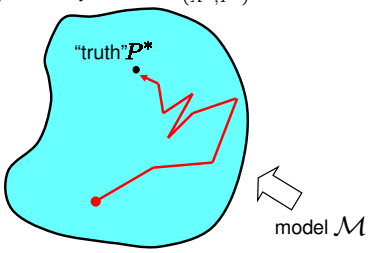
Bayes can be inconsistent!

Statistical Consistency

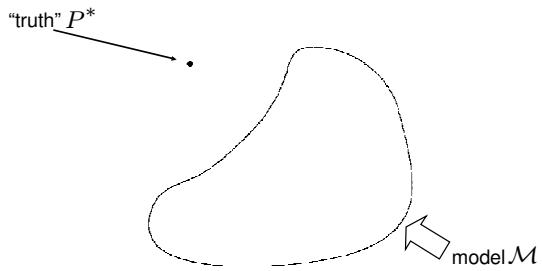
- Let $\hat{P} : \cup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}$ be an estimator/learner
- $\hat{P}$ is called consistent (relative to d) if for all $P^* \in \mathcal{M}$, with P^*-probability 1, $d(P^*, \hat{P}_{(X^n, Y^n)}) \to 0$

Statistical Consistency

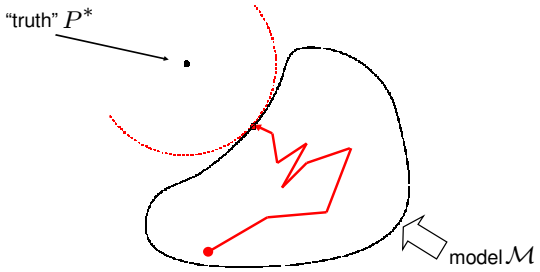
- Let $\hat{P} : \cup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}$ be an estimator/learner
- $\hat{P}$ is called consistent (relative to d) if for all $P^* \in \mathcal{M}$, with P^*-probability 1, $d(P^*, \hat{P}_{(X^n, Y^n)}) \to 0$



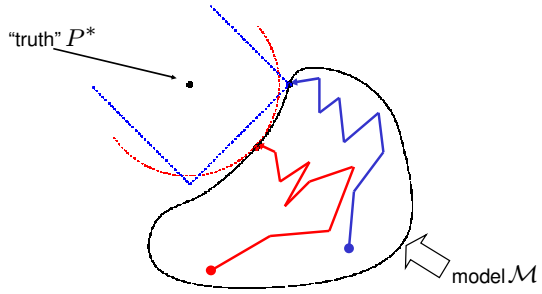
Consistency when Model is Wrong



Consistency when Model is Wrong



Distance More Important when Model is Wrong



Bayesian Inconsistency when Model is Wrong

Grünwald and Langford, *Machine Learning 66(2-3)*, 2007

- There exist
  - a countable set of distributions $\mathcal{M} = \{P_0, P_1, P_2, \ldots\}$
  - a "true distribution" P^*, and a "best approximation" $\tilde{P} = P_0$

$$\tilde{P} = \arg\min_{P \in \mathcal{M}} d(P^*, \tilde{P}) = \epsilon > 0$$

  - a prior distribution W on $\mathcal{M}$ with $W(\tilde{P}) = 1/2$

...such that, P^*-almost surely, the standard Bayesian estimators never "converge" to $\tilde{P}$

Bayesian Inconsistency when Model is Wrong

Grünwald and Langford, *Machine Learning 66(2-3)*, 2007

- There exist
  - a countable set of distributions $\mathcal{M} = \{P_0, P_1, P_2, \ldots\}$
  - a "true distribution" P^*, and a "best approximation" $\tilde{P} = P_0$

$$\tilde{P} = \arg\min_{P \in \mathcal{M}} d(P^*, \tilde{P}) = \epsilon > 0$$

  - a prior distribution W on $\mathcal{M}$ with $W(\tilde{P}) = 1/2$

...such that, P^*-almost surely, the standard Bayesian estimators never "converge" to $\tilde{P}$

logistic regression setting

d=KL divergence! (so this is real bad)

- There exist
  - a countable set of distributions $\mathcal{M} = \{P_0, P_1, P_2, \ldots\}$
  - a "true distribution" P^*, and a "best approximation" $\tilde{P} = P_0$,

$$\tilde{P} = \arg\min_{P \in \mathcal{M}} d(P^*, \tilde{P}) = \epsilon > 0$$

  - a prior distribution W on $\mathcal{M}$ with $W(\tilde{P}) = 1/2$

...such that, P^*-almost surely, the standard Bayesian estimators never "converge" to $\tilde{P}$

- There exist
  - a countable set of distributions $\mathcal{M} = \{P_0, P_1, P_2, \ldots\}$
  - a "true distribution" P^*, and a "best approximation" $\tilde{P} = P_0$,
    $\tilde{P} = \arg\min_{P \in \mathcal{M}} d(P^*, \tilde{P}) = \epsilon > 0$
  - a prior distribution W on $\mathcal{M}$ with $W(\tilde{P}) = 1/2$

  ...such that, P^*-almost surely, the standard Bayesian estimators never "converge" to $\tilde{P}$

- Bayesian inference is based on posterior $W(P \mid X^n, Y^n)$

  prediction ⟶ average over posterior

  estimation ⟶ output subset of $\mathcal{M}$ with large posterior probability

- There exist
  - a countable set of distributions $\mathcal{M} = \{P_0, P_1, P_2, \ldots\}$
  - a "true distribution" P^*, and a "best approximation" $\tilde{P} = P_0$,
    $\tilde{P} = \arg\min_{P \in \mathcal{M}} d(P^*, \tilde{P}) = \epsilon > 0$
  - a prior distribution W on $\mathcal{M}$ with $W(\tilde{P}) = 1/2$

  ...such that, P^*-almost surely, the standard Bayesian estimators never "converge" to $\tilde{P}$

- Bayesian inference is based on posterior $W(P \mid X^n, Y^n)$

  prediction ⟶ average over posterior

  estimation ⟶ output subset of $\mathcal{M}$ with large posterior probability

- There exist
  - a countable set of distributions $\mathcal{M} = \{P_0, P_1, P_2, \ldots\}$
  - a "true distribution" P^*, and a "best approximation" $\tilde{P} = P_0$,
    $\tilde{P} = \arg\min_{P \in \mathcal{M}} d(P^*, \tilde{P}) = \epsilon > 0$
  - a prior distribution W on $\mathcal{M}$ with $W(\tilde{P}) = 1/2$

  ...such that, P^*-almost surely, the standard Bayesian estimators never "converge" to $\tilde{P}$

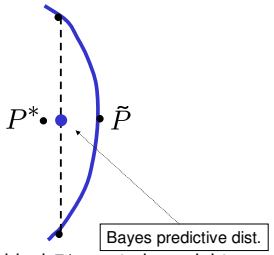
For all $K > 0, W(P : d(P^*, P) \leq K \mid X^n, Y^n) \to 0$, P^*–a.s.

- Bayesian inference is based on posterior $W(P \mid X^n, Y^n)$

  prediction ⟶ average over posterior

  estimation ⟶ output subset of $\mathcal{M}$ with large posterior probability

The "Geometry" of Inconsistency

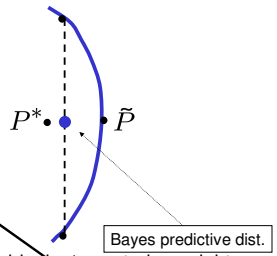


- Note that each individual P's posterior weight $W(P \mid X^n, Y^n)$ does tend to 0, eventually, a.s.

The "Geometry" of Inconsistency

essential that we take 'nonparametric' model (or model selection with infinite nr of models)

if model is 'parametric' then we get consistency but we need 'too much data'

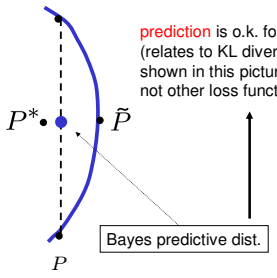


- Note that each individual P's posterior weight $W(P \mid X^n, Y^n)$ does tend to 0, eventually, a.s.

The "Geometry" of Inconsistency

essential that we take 'nonparametric' model (or model selection with infinite nr of models)

prediction is o.k. for log loss (relates to KL divergence, shown in this picture) but not other loss functions

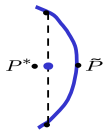


The Role of Convexity

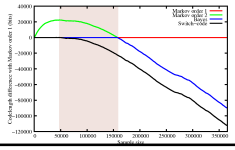
- Bayes predictive distribution is a mixture of $P \in \mathcal{M}$
- Indeed, Bayes *is* consistent with KL divergence, countable set $\mathcal{M}$, $P^* \notin \mathcal{M}$ if model is convex (closed under taking mixtures)
- Bayes implicitly "punishes" complex sub-models. To get consistency when model is wrong, need to punish them significantly more
  - if the prior $W(P)$ is changed to $W(P)^{\sqrt{n}}$ then we do get "consistency" even for nonconvex models

Menu – Two Pictures

1. Introduction
2. Learning when Models are **Seriously Wrong**



3. Model Selection when Models are **Almost True**



Model Selection Methods

- Suppose we observe data $y^n = y_1, \ldots, y_n \in \mathcal{Y}^n$
- We want to know which model in our list of candidate models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ best explains the data
  - $\hat{\theta}_k$ is maximum likelihood estimator (best-fit) within model $\mathcal{M}_k$
- A model selection method

$$\hat{k} : \bigcup_{n \geq 1} \mathcal{Y}^n \to \mathbb{N}$$

is a function mapping data sequences of arbitrary length to model indices
  - $\hat{k}(y^n)$ is model chosen for data y^n

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion (AIC, 1973)

$$\hat{k}(y^n) \text{ is } k \text{ minimizing } -\log p_{\hat{\theta}_k}(y^n) + k$$

2. **BIC-type**
   - Bayesian Information Criterion (BIC, 1978)

$$\hat{k}(y^n) \text{ is } k \text{ minimizing } -\log p_{\hat{\theta}_k}(y^n) + \frac{k}{2} \log n$$

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion (AIC, 1973)
   - **leave-one-out cross-validation**
   - DIC, C_p
2. **BIC-type**
   - Bayesian Information Criterion (BIC, 1978)
   - prequential validation
   - **Bayes factor model selection**
   - standard Minimum Description Length (MDL)

The AIC-BIC Dilemma

asymptotic overfitting

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion
   - **leave-one-out cross-validation**
   - DIC, C_p

inconsistent

2. **BIC-type**
   - Bayesian Information Criterion
   - prequential validation
   - **Bayes factor model selection**
   - standard MDL

consistent

The AIC-BIC Dilemma

- Two main types of **model selection** methods:

1. **AIC-type**
   - Akaike Information Criterion
   - **leave-one-out cross-validation**
   - DIC, C_p

   inconsistent
   optimal rate

2. **BIC-type**
   - Bayesian Information Criterion
   - prequential validation
   - **Bayes factor model selection**
   - standard MDL

   consistent
   slower rate

asymptotic underfitting

The Best of Both Worlds

- We give a novel analysis of the slower convergence rate of BIC-type methods: *the catch-up phenomenon*

The Best of Both Worlds

- We give a novel analysis of the slower convergence rate of BIC-type methods: *the catch-up phenomenon*
- This allows us to define a model selection/averaging method that, in a wide variety of circumstances,
  1. is provably **consistent**
  2. provably achieves **optimal convergence rates**

The Best of Both Worlds

- We give a novel analysis of the slower convergence rate of BIC-type methods: *the catch-up phenomenon*
- This allows us to define a model selection/averaging method that, in a wide variety of circumstances,
  1. is provably **consistent**
  2. provably achieves **optimal convergence rates (often…)**
- …even though it had been suggested that this is impossible! Yang 2005, Forster 2001, Sober 2004
- For many model classes, method is computationally feasible

Sub-Menu

1. Bayes Factor Model Selection
   - Predictive interpretation
2. The Catch-Up Phenomenon
   …. as exhibited by the Bayes factor method
3. Conclusion

Bayes Factor Model Selection

$\mathcal{M}_k = \{p_\theta \mid \theta \in \Theta_k\} \qquad \Theta_k \subseteq \mathbb{R}^k \qquad k \in \mathcal{K} \subset \mathbb{N}$

$\hat{k}(y^n)$ is k maximizing a posteriori probability

$$p(\mathcal{M}_k \mid y^n) = \frac{p(y^n \mid \mathcal{M}_k)\pi(k)}{\sum_{k \in \mathcal{K}} p(y_n \mid \mathcal{M}_k)\pi(k)}$$

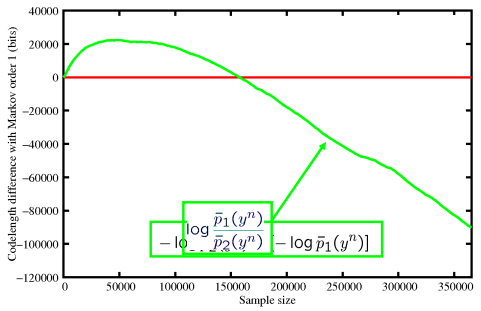
$$\bar{p}_k := p(y^n \mid \mathcal{M}_k) = \int_{\theta \in \Theta_k} p_\theta(y^n) w_k(\theta) d\theta$$

$\pi(k)$ is prior

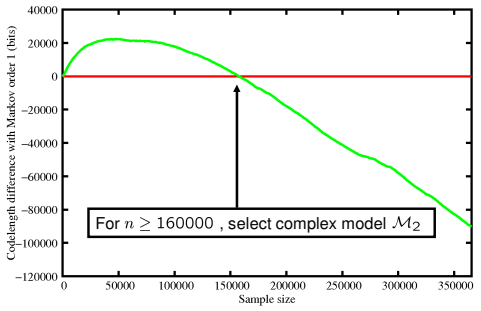
$w_1, w_2, \ldots$ are priors

$\hat{k}(y^n)$ is k minimizing $-\log \bar{p}_k(y^n) - \log \pi(k) \approx -\log \bar{p}_k(y^n)$

$$\approx\approx \frac{k}{2} \log n - \log p_{\hat{\theta}_k}(y^n)$$



Bayes factor model selection between 1st-order and 2nd-order Markov model for "The Picture of Dorian Gray"



Bayes factor model selection between 1st-order and 2nd-order Markov model for "The Picture of Dorian Gray"

The Catch-Up Phenomenon

- Suppose we select between "simple" model $\mathcal{M}_1$ and "complex" model $\mathcal{M}_2$
- Common Phenomenon: for some n_{switch}
  simple model predicts better if $n < n_{\text{switch}}$
  complex model predicts better if $n \geq n_{\text{switch}}$
  - this seems to be the very reason why it makes sense to prefer a simple model even if the complex one is true
- We would expect Bayes factor method to switch at about $n \approx n_{\text{switch}}$....
  but is this really where Bayes switches!?

Menu

1. Bayes Factor Model Selection
   - Predictive interpretation
2. The Catch-Up Phenomenon
   .... as exhibited by the Bayes factor method
3. Conclusion

Bayesian prediction

- Given model $\mathcal{M}_k$, Bayesian marginal likelihood is
  $\bar{p}_k(y^n) = p(y^n \mid \mathcal{M}_k) := \int_{\Theta_k} p_\theta(y^n) w(\theta) d\theta$
- Given model $\mathcal{M}_k$, predict by predictive distribution
  $\bar{p}_k(y_{n+1} \mid y^n) = \frac{\bar{p}_k(y^{n+1})}{\bar{p}_k(y^n)} = \int_{\Theta_k} p_\theta(y_{n+1} \mid y^n) w(\theta \mid y^n) d\theta$

Logarithmic Loss

- If we measure prediction quality by 'log loss',
  $\text{loss}(y, p) := -\log p(y)$
  then accumulated loss satisfies

$$\sum_{i=1}^{n} \text{loss}(y_i, p(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} \left[-\log p(y_i \mid y^{i-1})\right]$$
$$= -\log \prod_{i=1}^{n} p(y_i \mid y_1, \ldots, y_{i-1}) = -\log \prod_{i=1}^{n} \frac{p(y^i)}{p(y^{i-1})}$$
$$= -\log p(y_1, \ldots, y_n)$$

so that **accumulated log loss = minus log likelihood**

The Most Important Slide

- Bayes picks the k minimizing

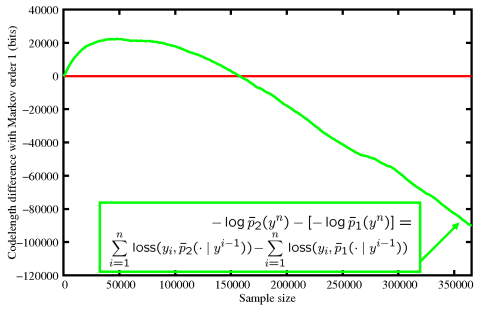
$$-\log \bar{p}_k(y_1,\ldots,y_n) = \sum_{i=1}^{n} \text{loss}(y_i, \bar{p}_k(\cdot \mid y^{i-1}))$$

- **Prequential interpretation** of Bayes model selection: select the model $\mathcal{M}_k$ such that, when used as a sequential prediction strategy, $\bar{p}_k = p(\cdot \mid \mathcal{M}_k)$ minimizes accumulated sequential prediction error

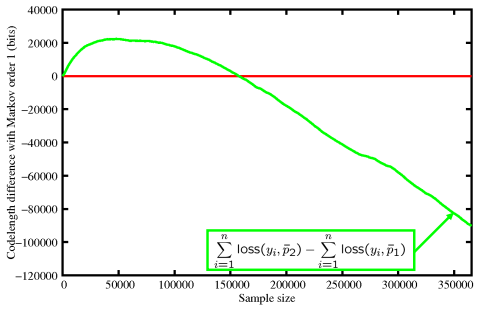
Dawid '84, Rissanen '84

Menu

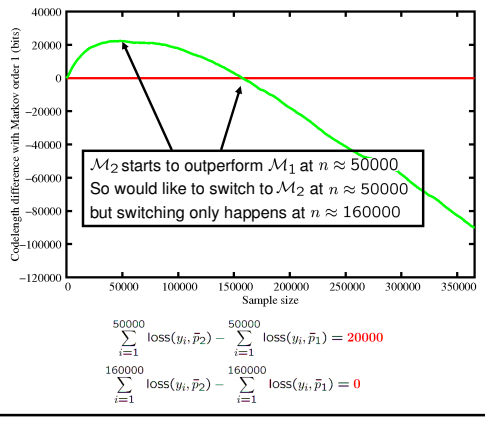
1. Bayes Factor Model Selection
   - Predictive interpretation
2. The Catch-Up Phenomenon
   .... as exhibited by the Bayes factor method
3. Conclusion



Green curve depicts difference in accumulated prediction error between predicting with $\mathcal{M}_2$ and predicting with $\mathcal{M}_1$

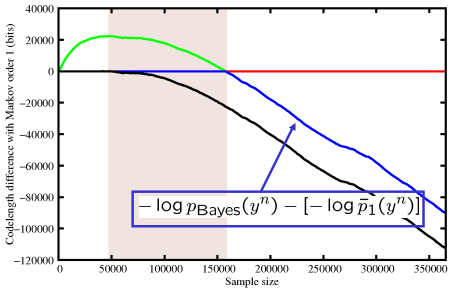


Green curve depicts difference in accumulated prediction error between predicting with $\mathcal{M}_2$ and predicting with $\mathcal{M}_1$



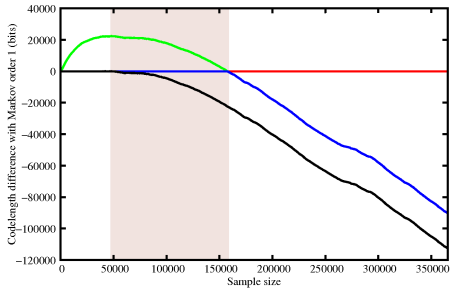
The Catch-Up Phenomenon

- Suppose we select between "simple" model $\mathcal{M}_1$ and "complex" model $\mathcal{M}_2$
- Common Phenomenon: for some n_{switch}
  simple model predicts better if $n < n_{\text{switch}}$
  complex model predicts better if $n \geq n_{\text{switch}}$
- Bayes exhibits **inertia**: complex model has to "**catch up**", so we prefer simpler model for a while even after $n \geq n_{\text{switch}}$



Model averaging does not help!

$$p_{\text{Bayes}}(y^n) = \frac{1}{2}\bar{p}_1(y^n) + \frac{1}{2}\bar{p}_2(y^n)$$



Can we modify Bayes so as to do as well as the black curve? Almost!

The Switch Distribution

- Suppose we switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at sample size ***s***
- Our total prediction error is then

$$\sum_{i=1}^{s} \text{loss}(y_i, \bar{p}_1) + \sum_{s+1}^{n} \text{loss}(y_i, \bar{p}_2)) =$$
$$-\log \bar{p}_1(y^s) - \log \bar{p}_2(y_{s+1}, \ldots, y_n \mid y^s)$$

The Switch Distribution

- Suppose we switch from $\mathcal{M}_1$ to $\mathcal{M}_2$ at sample size ***s***
- Our total prediction error is then

$$\sum_{i=1}^{s} \text{loss}(y_i, \bar{p}_1) + \sum_{s+1}^{n} \text{loss}(y_i, \bar{p}_2)) =$$
$$-\log \bar{p}_1(y^s) - \log \bar{p}_2(y_{s+1}, \ldots, y_n \mid y^s)$$

- If we define

$$\bar{p}_{\text{switch}}(y^n \mid s) = \bar{p}_1(y^s) \cdot \bar{p}_2(y_{s+1}, \ldots, y_n \mid y^s)$$

then total prediction error is $-\log \bar{p}_{\text{switch}}(y^n \mid s)$

- $\bar{p}_{\text{switch}}$ may be viewed both as a prediction strategy and as a distribution over infinite sequences

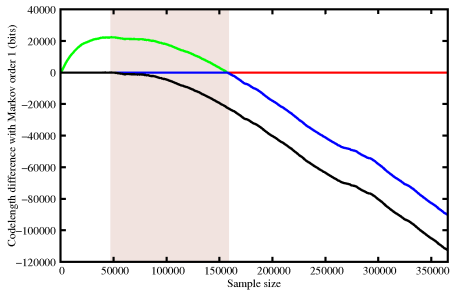
The Switch Distribution

- We want to predict $y_1, y_2, \ldots$ using some distribution $\bar{p}$ such that *no matter what data are observed*, i.e. for *all* $y^n \in \mathcal{Y}^n$,

$$-\log \bar{p}(y^n) \approx -\log \bar{p}_{\text{switch}}(y^n \mid \hat{s}(y^n))$$

where $\hat{s}(y^n)$ maximizes $\bar{p}_{\text{switch}}(y^n \mid s)$

- We achieve this by treating s as a parameter, putting a prior on it, and then integrating s out
(adopt a Bayesian solution to a Bayesian problem…)



Switch Distribution (very briefly)

- The switch distribution is obtained by putting a (cleverly constructed) prior on sequences of models rather than individual models
- $\pi(1,1,2,2,2,1,1,2,2,2,2,\ldots)$ is the prior mass you put on the meta-hypothesis that
- " $\mathcal{M}_1$ is "best" at sample size 1,2,6,7
  $\mathcal{M}_2$ is "best" at sample size 3,4,5,8,9,..."
  - readily extended to > 2 models
- Now we apply "Bayes" using this new prior for prediction and model selection
  - can be done efficiently by dynamic programming

Results

- Theorem 1: "In all situations in which Bayes factor model selection is consistent, model selection based on Switching Prior is consistent (like BIC)
- Theorem 2: "In many 'models almost true' situations prediction based on the Switching Prior achieves the optimal rate of convergence (like AIC)
- Practical Experiments confirm this
- Method is "formally" Bayes, but Bayesian interpretation very tenuous!

It's prequential!

Discussion

- Model(s) 'almost true': standard Bayes is 'too slow'. We can interpret 'switch distribution' as a 'reparation' of Bayes
  ...but also as a frequentist hypothesis test to check whether the Bayesian model and prior assumptions are justified, and to 'step out of the model' if necessary

Discussion

- Model(s) 'almost true': standard Bayes is 'too slow'. We can interpret 'switch distribution' as a 'reparation' of Bayes
  ...but also as a frequentist hypothesis test to check whether the Bayesian model and prior assumptions are justified, and to 'step out of the model' if necessary
- Model(s) 'seriously wrong': standard Bayes can be inconsistent unless model **convex**.

Discussion

- Model(s) 'almost true': standard Bayes is 'too slow'. We can interpret 'switch distribution' as a 'reparation' of Bayes
  ...but also as a frequentist hypothesis test to check whether the Bayesian model and prior assumptions are justified, and to 'step out of the model' if necessary
- Model(s) 'seriously wrong': standard Bayes can be inconsistent unless model convex.
  In recent work we provide a frequentist 'empirical convexity test' (does the convex closure of the model fit the data better?) to check whether Bayesian model and prior assumptions are justified, and to 'step out of the model' if necessary

Discussion

- Model(s) 'almost true': standard Bayes is 'too slow'. We can interpret 'switch distribution' as a 'reparation' of Bayes
  ...but also as a frequentist hypothesis test to check whether the Bayesian model and prior assumptions are justified, and to 'step out of the model' if necessary
- Model(s) 'seriously wrong': standard Bayes can be inconsistent unless model convex.
  In recent work we provide a frequentist 'empirical convexity test' (does the convex closure of the model fit the data better?) to check whether Bayesian model and prior assumptions are justified, and to 'step out of the model' if necessary

We are effectively combining Bayes+Popper+(..new principle..)

Conclusion

We are effectively combining Bayes+Popper+(..new principle..)

- Bayes+Popper has been proposed before, on an informal level, by eminent Bayesians & others:
  - A.P. Dawid (1982): 'the well-calibrated Bayesian'
  - A. Gelman (various papers, blog)
- New principle: **our ongoing work!**

Conclusion

We are effectively combining Bayes+Popper+(..new principle..)

- Bayes+Popper has been proposed before, on an informal level, by eminent Bayesians & others:
  - A.P. Dawid (1982): 'the well-calibrated Bayesian'
  - A. Gelman (various papers, blog)
- New principle: **our ongoing work!**
- **Literature:** Grünwald & Langford, Machine Learning, 2007. Van Erven, Grünwald and De Rooij, NIPS 2007. Van Erven's PhD. Thesis (2010). (notice: *machine learning* is important!)