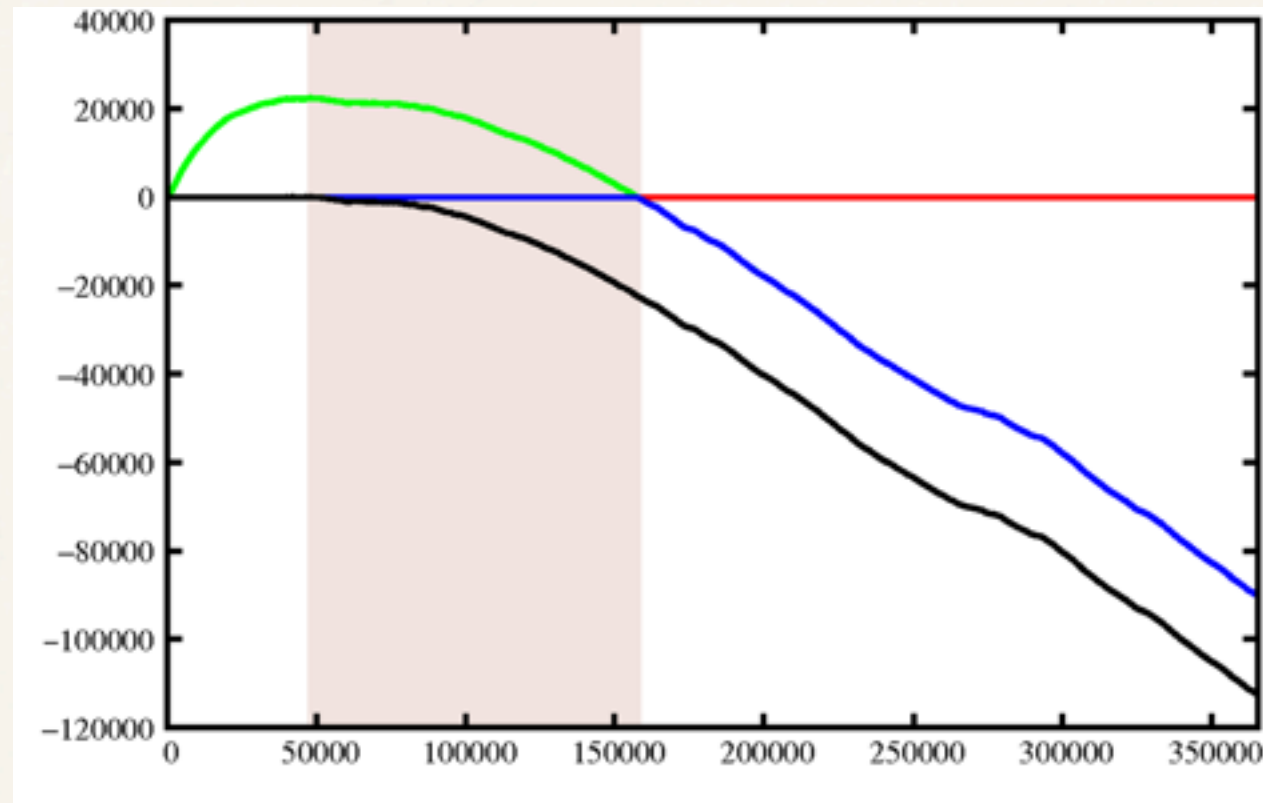


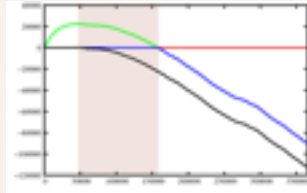
The Catch-up Phenomenon in Bayesian and MDL Model Selection



Tim van Erven
www.timvanerven.nl
23 May 2013

Joint work with **Peter Grünwald**, **Steven de Rooij** and **Wouter Koolen**

Outline

- ❖ *Bayes Factors* and *MDL* Model Selection
 - ❖ Consistent, but **suboptimal predictions**
- ❖ Explanation: the **Catch-up Phenomenon**
 - ❖ Predictive MDL interpretation of Bayes factors
 - ❖ Markov chain example 
- ❖ Solution: the **Switch Distribution**
 - ❖ Simulations & Theorems: consistent + **optimal predictions**
 - ❖ Cumulative risk

Two Desirable Properties in Model Selection

- ✧ Suppose $\mathcal{M}_1, \dots, \mathcal{M}_K$ are statistical models
(sets of probability distributions: $\mathcal{M}_k = \{p_\theta | \theta \in \Theta_k\}$)
- ✧ **Consistency**: If some p^* in model \mathcal{M}_{k^*} generates the data, then \mathcal{M}_{k^*} is selected with probability one as the amount of data goes to infinity.
- ✧ **Rate of convergence**: How fast does an estimator based on the available models converge to the true distribution?

AIC-BIC Dilemma

	Consistent	Optimal rate of convergence
BIC, Bayes, MDL	Yes	No
AIC, LOO Cross-validation	No	Yes

Two Desirable Properties in Model Selection

- ✧ Suppose $\mathcal{M}_1, \dots, \mathcal{M}_K$ are statistical models
(sets of probability distributions: $\mathcal{M}_k = \{p_\theta | \theta \in \Theta_k\}$)
- ✧ **Consistency**: If some p^* in model \mathcal{M}_{k^*} generates the data, then \mathcal{M}_{k^*} is selected with probability one as the amount of data goes to infinity.
- ✧ **Rate of convergence**: How fast does an estimator based on the available models converge to the true distribution?

AIC-BIC Dilemma

	Consistent	Optimal rate of convergence
BIC, Bayes, MDL	Yes	No ?
AIC, LOO Cross-validation	No	Yes

Bayesian Prediction

- ✧ Given model $\mathcal{M}_k = \{p_\theta | \theta \in \Theta_k\}$ with prior w_k and data $x^n = (x_1, \dots, x_n)$, the Bayesian **marginal likelihood** is

$$\bar{p}_k(x^n) \equiv p(x^n | \mathcal{M}_k) := \int_{\Theta_k} p_\theta(x^n) w_k(\theta) d\theta$$

- ✧ Given \mathcal{M}_k predict with **estimator**

$$\bar{p}_k(x_{n+1} | x^n) = \frac{\bar{p}_k(x^{n+1})}{\bar{p}_k(x^n)} = \int_{\Theta_k} p_\theta(x_{n+1} | x^n) w_k(\theta | x^n) d\theta$$

Bayes Factors and MDL Model Selection

- ✧ Suppose we have multiple models $\mathcal{M}_1, \mathcal{M}_2, \dots$
- ✧ **Bayes factors:** Put a prior π on model index k and choose $\hat{k}(x^n)$ to **maximize the posterior probability**

$$p(\mathcal{M}_k | x^n) := \frac{\bar{p}_k(x^n) \pi(k)}{\sum_{k'} \bar{p}_{k'}(x^n) \pi(k')}$$

- ✧ $\hat{k}(x^n)$ is **minimizing**

$$-\log \bar{p}_k(x^n) - \log \pi(k) \approx -\log \bar{p}_k(x^n)$$

Bayes Factors and MDL

Model Selection

- ✧ Suppose we have multiple models $\mathcal{M}_1, \mathcal{M}_2, \dots$
- ✧ **Bayes factors:** Put a prior π on model index k and choose $\hat{k}(x^n)$ to **maximize the posterior probability**

$$p(\mathcal{M}_k | x^n) := \frac{\bar{p}_k(x^n) \pi(k)}{\sum_{k'} \bar{p}_{k'}(x^n) \pi(k')}$$

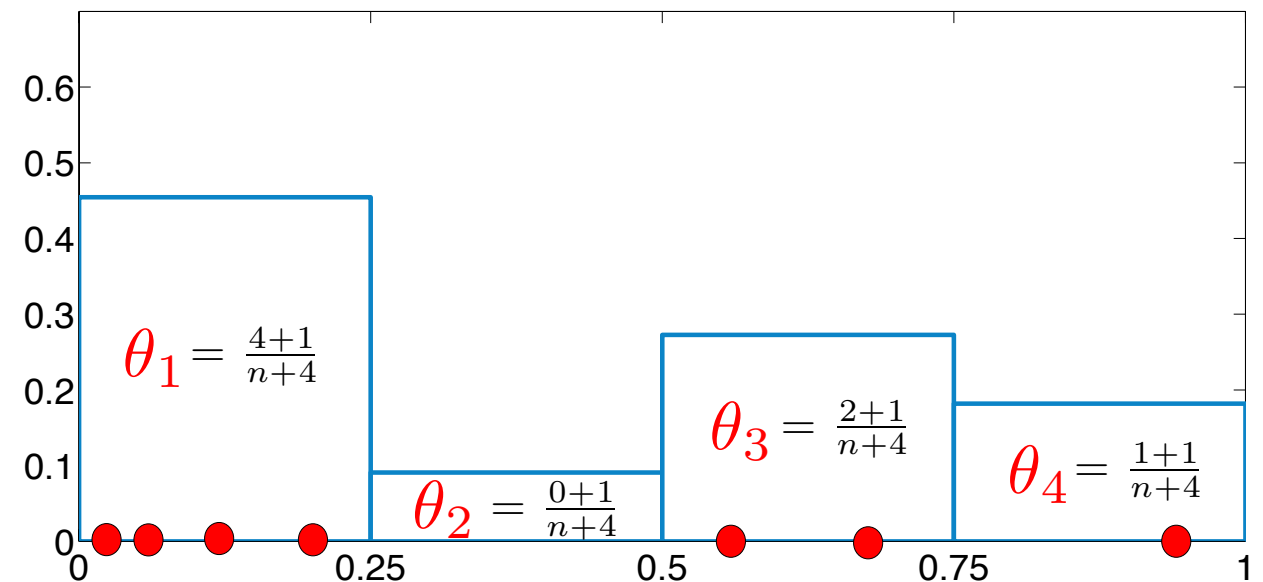
- ✧ $\hat{k}(x^n)$ is **minimizing**

$$\underbrace{-\log \bar{p}_k(x^n) - \log \pi(k)} \approx -\log \bar{p}_k(x^n)$$

Minimum Description Length (MDL)

Example: Histogram Density Estimation

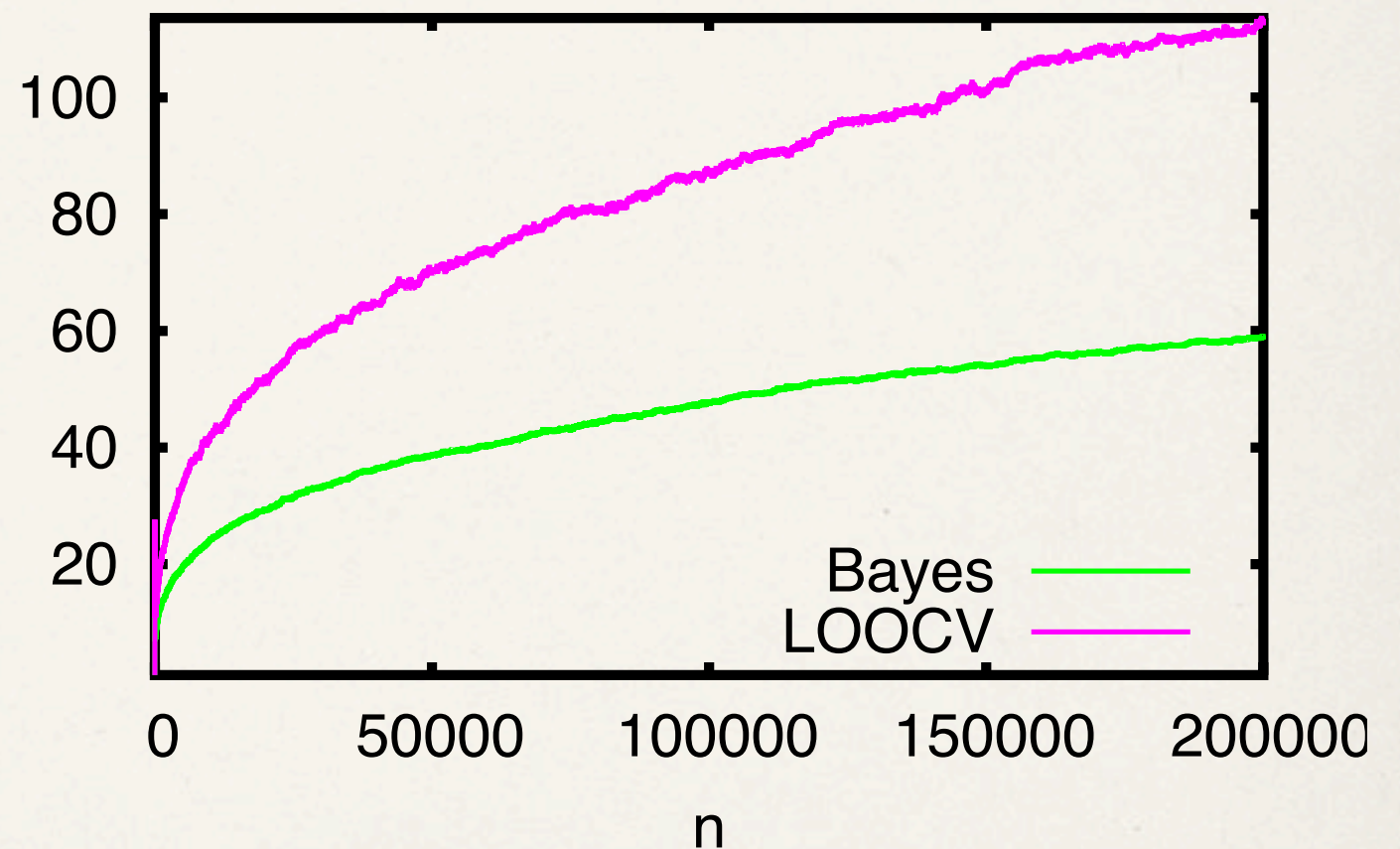
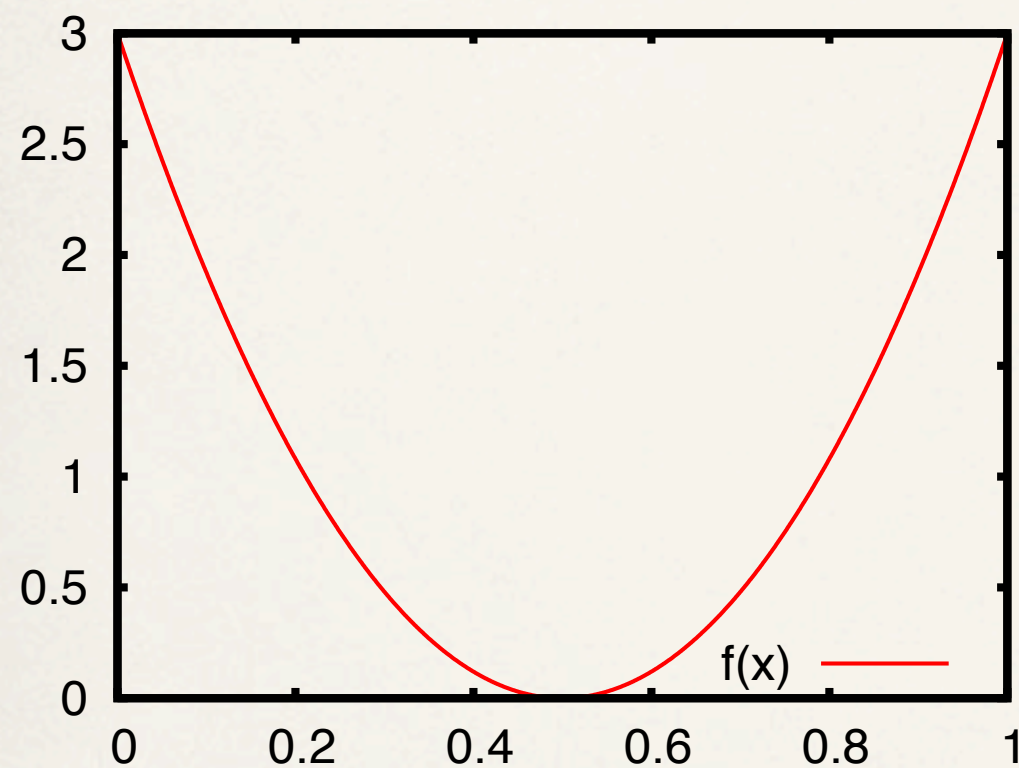
$$\mathcal{M}_k = \{p_\theta | \theta \in \Theta_k \subset \mathbb{R}^k\}$$



- * I.I.D. data in interval [0,1]
- * Given k, estimate density by the estimator in the figure
- * This is equivalent to \bar{p}_k for conjugate Dirichlet(1,...,1) prior
- * How should we **choose the number of bins k**?
 - * Too few: does not capture enough structure
 - * Too many: overfitting (many bins will be empty)
- * [Yu, Speed, '92]: Bayes does not achieve the optimal rate of convergence!

CV Selects More Bins than Bayes

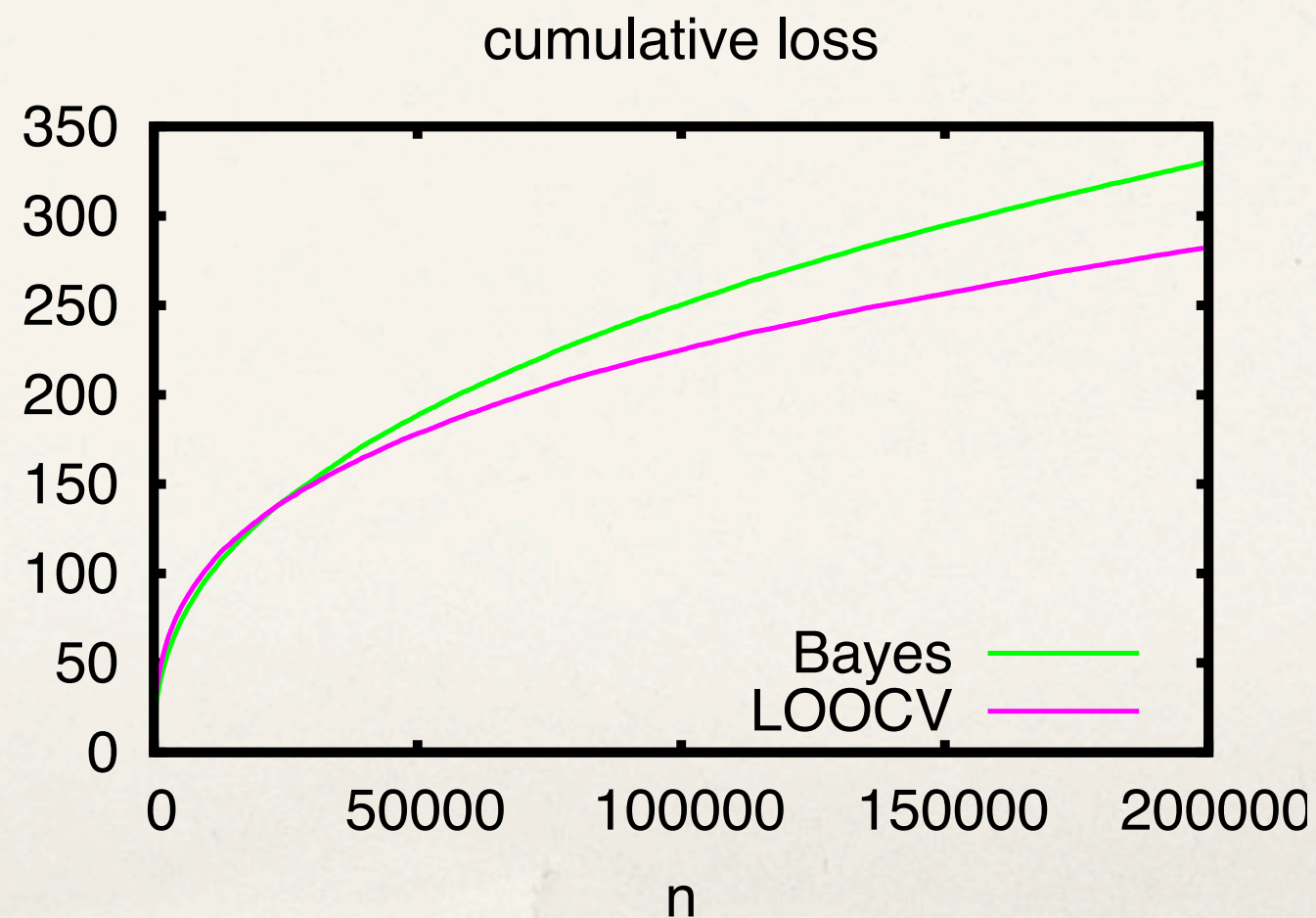
average # bins selected



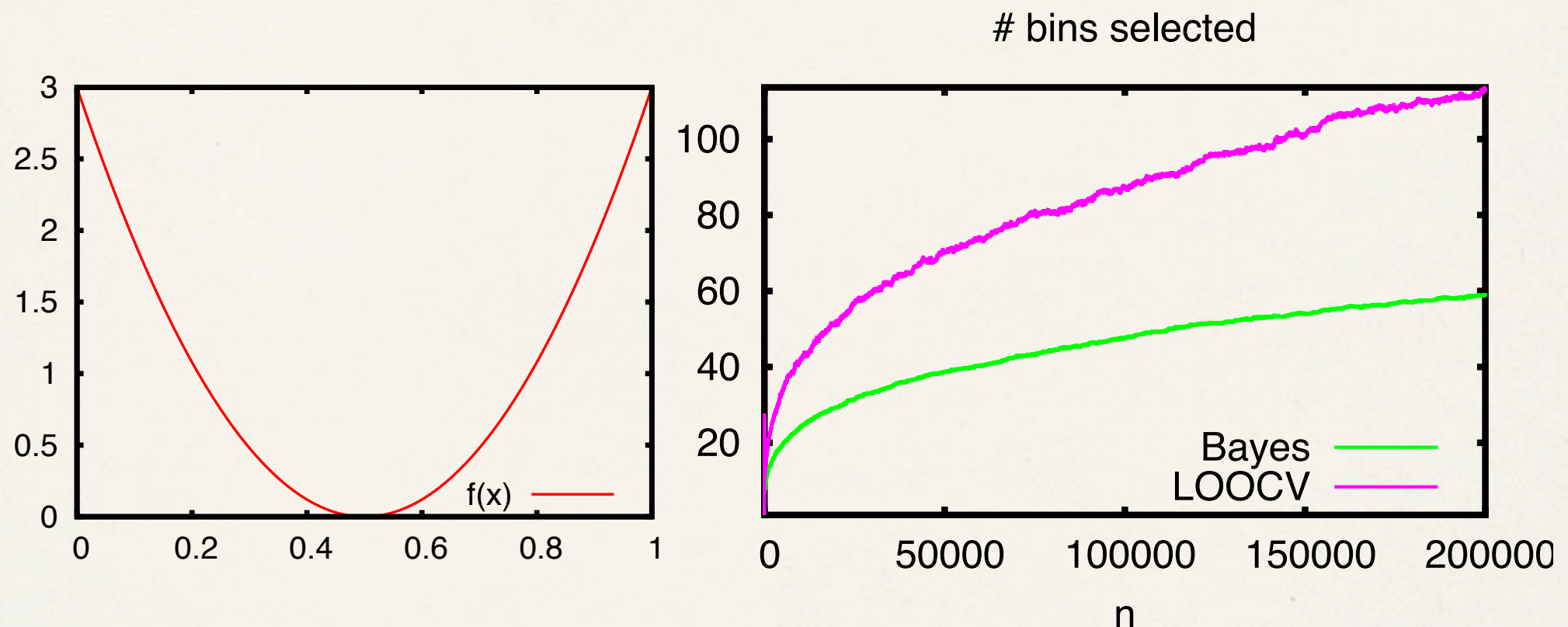
CV Predicts Better than Bayes

Prediction error in log loss at sample size n : $-\log \bar{p}_{\hat{k}(x^n)}(x_{n+1}|x^n)$

cumulative prediction error: $\sum_{i=1}^n -\log \bar{p}_{\hat{k}(x^{i-1})}(x_i|x^{i-1})$



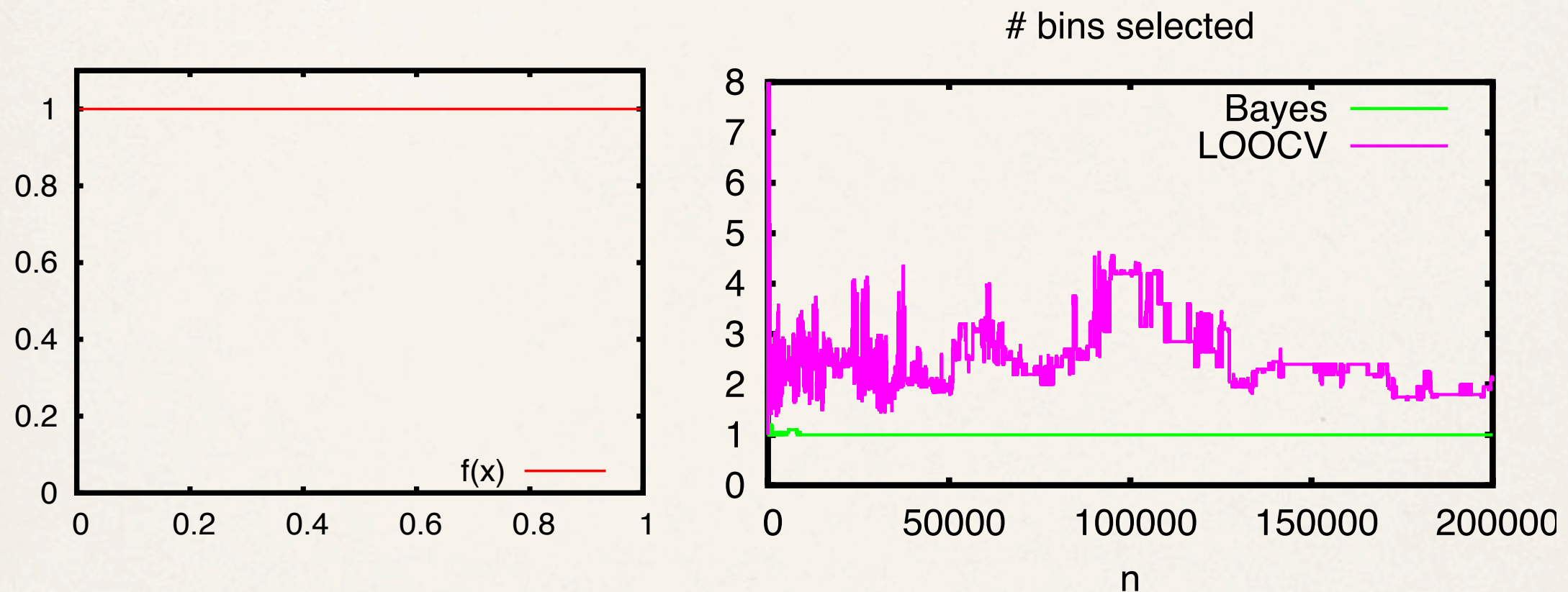
CV Predicts Better than Bayes...



- ❖ Density not a histogram, but can be approximated arbitrarily well
- ❖ LOO-CV, AIC converge at optimal rate
- ❖ Bayesian model selection selects too few bins (**underfits**)!

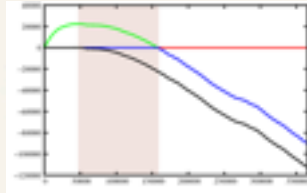
... but CV is Inconsistent!

- Now suppose data are sampled from the uniform distribution



- LOO cross-validation selects 2.5 bins on average: it is **inconsistent!**

Outline

- ❖ **Bayes Factors** and **MDL** Model Selection
 - ❖ Consistent, but **suboptimal predictions**
- ❖ *Explanation: the **Catch-up Phenomenon***
 - ❖ Predictive MDL interpretation of Bayes factors
 - ❖ Markov chain example 
- ❖ Solution: the **Switch Distribution**
 - ❖ Simulations & Theorems: consistent + **optimal predictions**
 - ❖ Cumulative risk

Logarithmic Loss

If we measure prediction quality by **log loss**

$$\text{loss}(p, x) := -\log p(x)$$

then **minus log likelihood = cumulative log loss**:

$$-\log p(x_1, \dots, x_n) = \sum_{i=1}^n -\log p(x_i | x^{i-1})$$

where $x^{i-1} = (x_1, \dots, x_{i-1})$

Proof. Take the negative logarithm of the chain rule: $p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i | x^{i-1})$

The Most Important Slide

Bayes factors and MDL pick the k minimizing

$$-\log \bar{p}_k(x_1, \dots, x_n) = \sum_{i=1}^n \underbrace{-\log \bar{p}_k(x_i | x^{i-1})}_{\text{Prediction error for model } \mathcal{M}_k \text{ at sample size } i!}$$

Prediction error for model \mathcal{M}_k at sample size i !

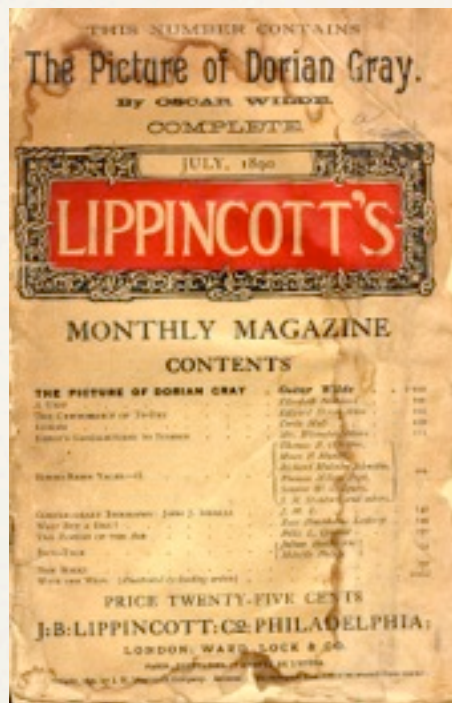
Prequential/predictive MDL interpretation:

select the model \mathcal{M}_k such that, when used as a sequential prediction strategy, \bar{p}_k minimizes **cumulative sequential prediction error**

[Dawid '84, Rissanen '84]

Example: Markov Chains

Natural language text: “The Picture of Dorian Gray” by Oscar Wilde

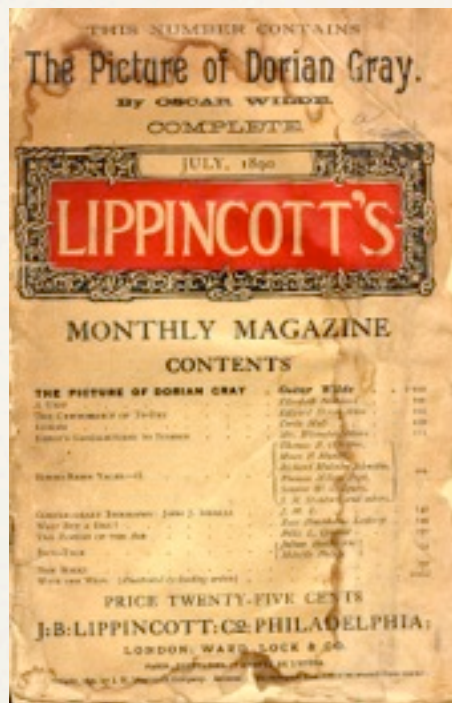


"... But beauty, real beauty, ends where an intellectual expression begins. Intellect is in itself a mode of exaggeration, and destroys the harmony of any face. The moment one sits down to think, one becomes all nose, or all forehead, or something horrid. Look at the successful men in any of the learned professions. How perfectly hideous they are! ..."

Compare the first-order and second-order **Markov chain models** on the first n characters in the book, with uniform priors on the transition probabilities

Example: Markov Chains

Natural language text: “The Picture of Dorian Gray” by Oscar Wilde



"... But beauty, real beauty, ends where an intellectual expression begins. Intellect is in itself a mode of exaggeration, and destroys the harmony of any face. The moment one sits down to think, one becomes all nose, or all forehead, or something horrid. Look at the successful men in any of the learned professions. How perfectly hideous they are! ..."

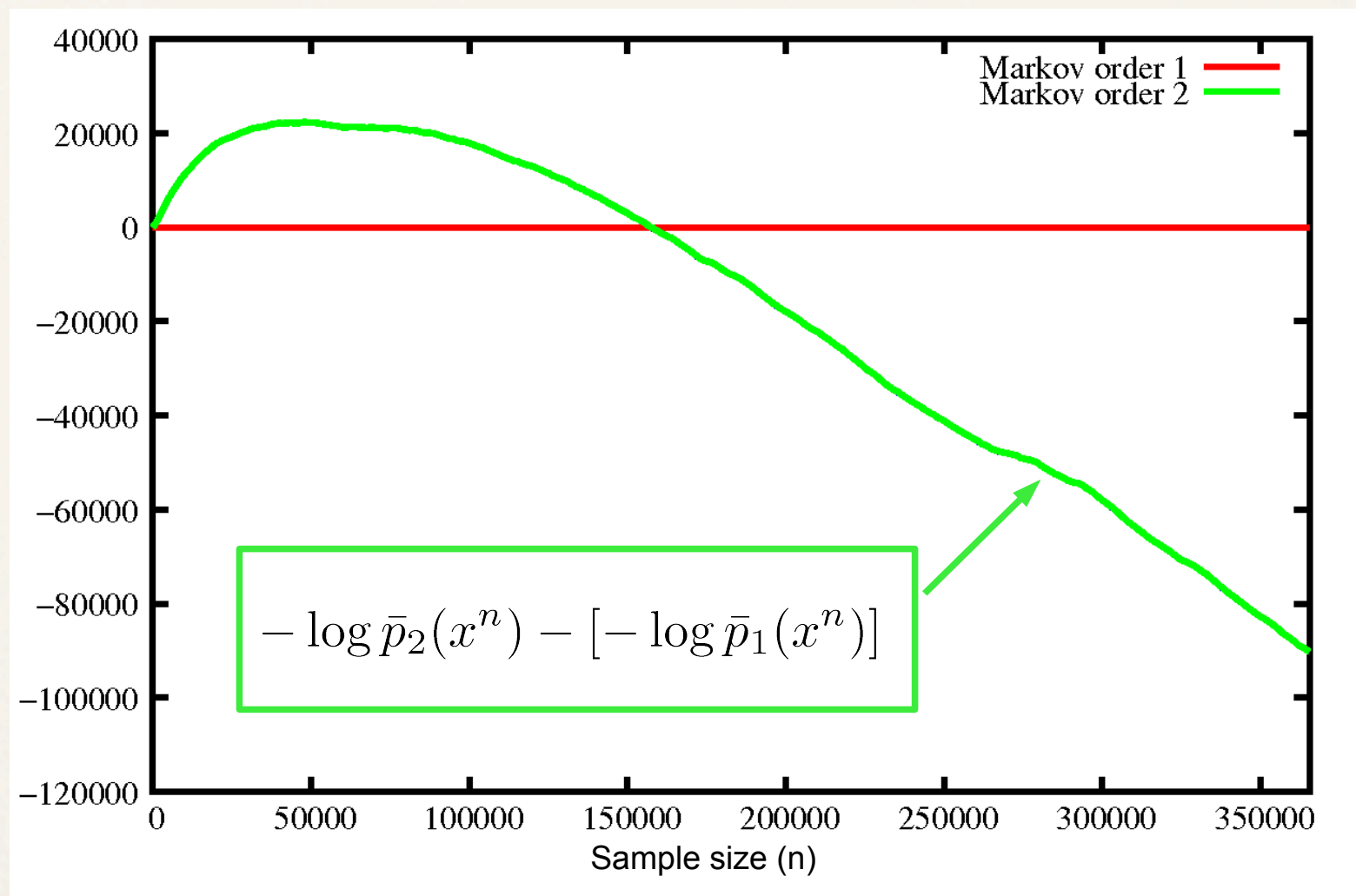
128x127 parameters

128x128x127 parameters

Compare the first-order and second-order **Markov chain models** on the first n characters in the book, with uniform priors on the transition probabilities

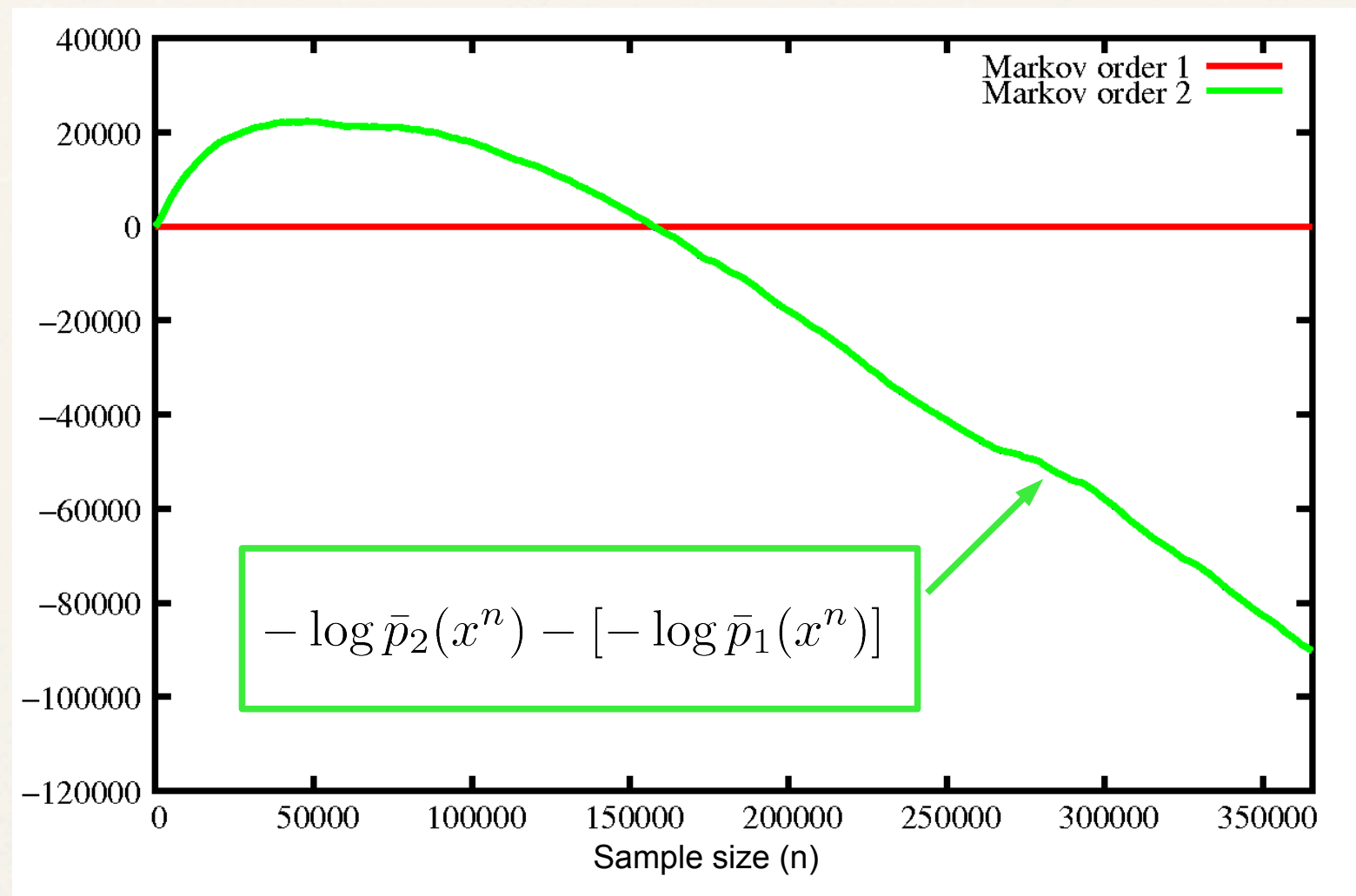
Example: Markov chains

Compare the marginal likelihoods



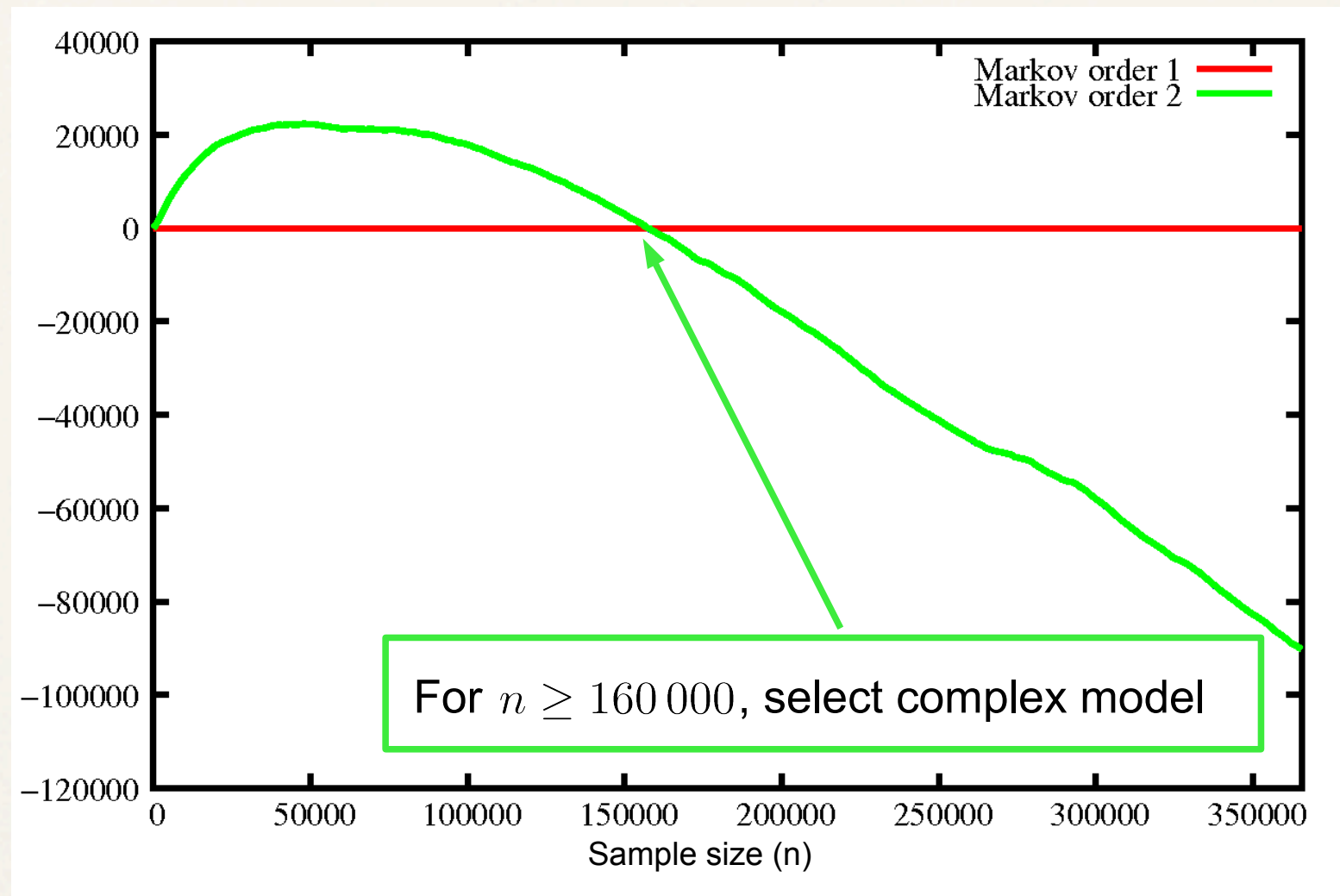
(green line equals the log of the Bayes factor)

Example: Markov chains



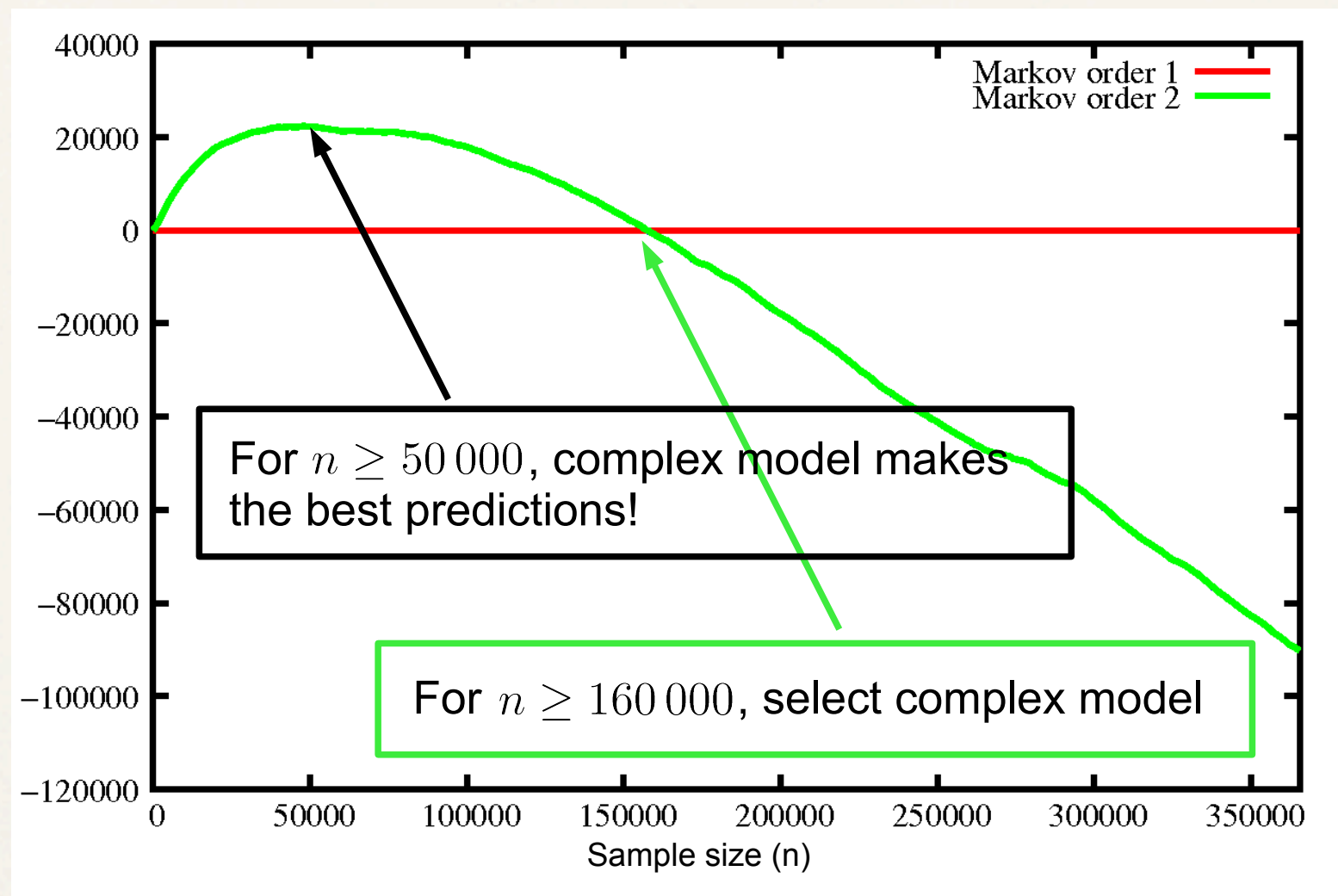
$$-\log \bar{p}_2(x^n) - [-\log \bar{p}_1(x^n)] = \sum_{i=1}^n \text{loss}(\bar{p}_2, x_i) - \sum_{i=1}^n \text{loss}(\bar{p}_1, x_i)$$

Example: Markov chains



$$-\log \bar{p}_2(x^n) - [-\log \bar{p}_1(x^n)] = \sum_{i=1}^n \text{loss}(\bar{p}_2, x_i) - \sum_{i=1}^n \text{loss}(\bar{p}_1, x_i)$$

Example: Markov chains



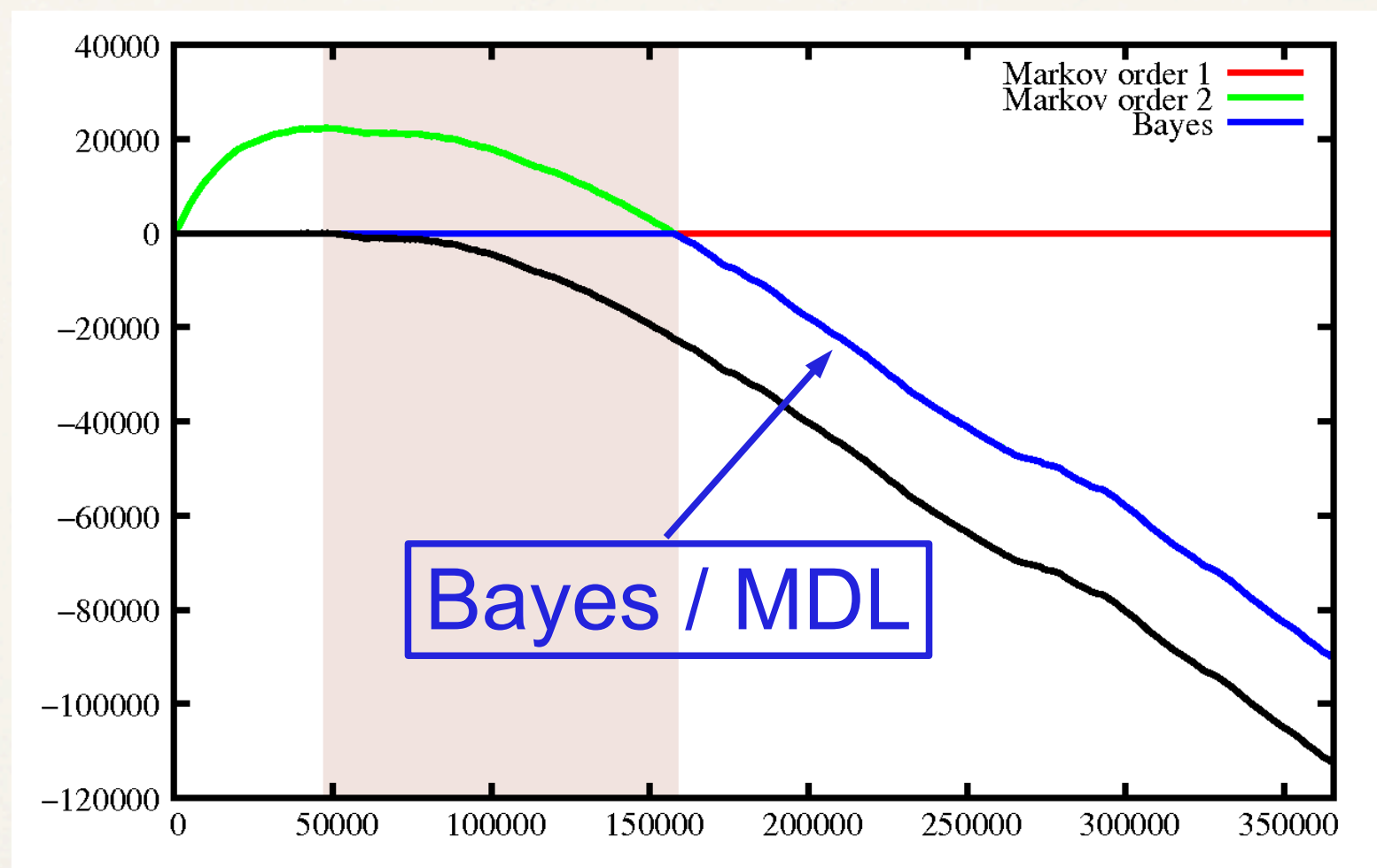
$$-\log \bar{p}_2(x^n) - [-\log \bar{p}_1(x^n)] = \sum_{i=1}^n \text{loss}(\bar{p}_2, x_i) - \sum_{i=1}^n \text{loss}(\bar{p}_1, x_i)$$

The Catch-up Phenomenon

- ✧ Given “simple” model \mathcal{M}_1 and a “complex” model \mathcal{M}_2
- ✧ Common phenomenon: for some sample size s
 - ✧ simple model predicts better if $n \leq s$
 - ✧ complex model predicts better if $n > s$
- ✧ **Catch-up Phenomenon:** Bayes / MDL exhibit **inertia**
 - ✧ complex model has to “**catch up**”,
so we prefer simpler model for a while even after $n > s$!
- ✧ **Remark:** Methods similar to Bayes factors (e.g. BIC) will also exhibit the catch-up phenomenon. Bayesian model averaging does not help either!

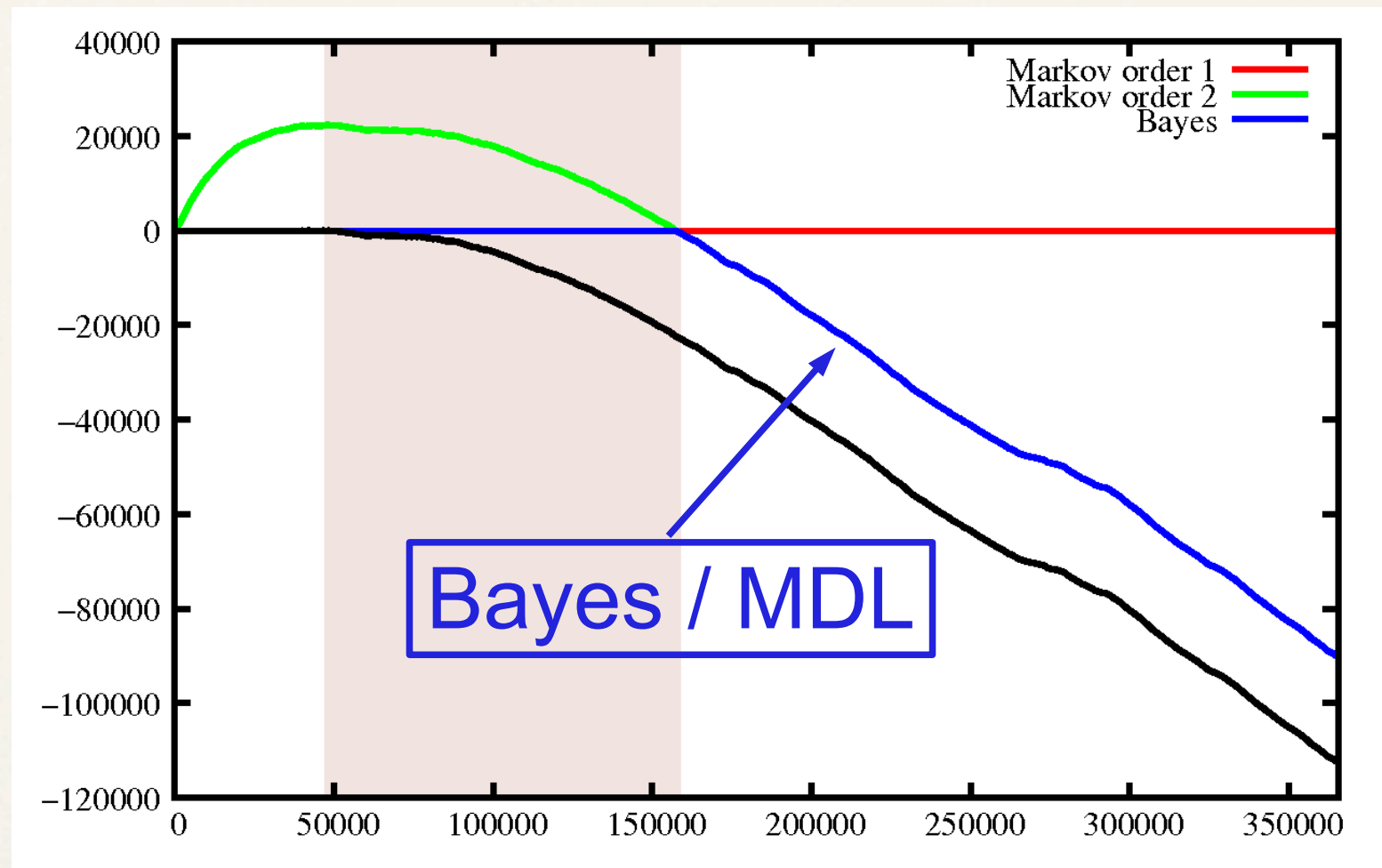


Example: Markov chains



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

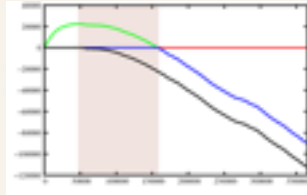
Example: Markov chains



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

Almost!

Outline

- ❖ **Bayes Factors** and **MDL** Model Selection
 - ❖ Consistent, but **suboptimal predictions**
- ❖ Explanation: the **Catch-up Phenomenon**
 - ❖ Predictive MDL interpretation of Bayes factors
 - ❖ Markov chain example 
- ❖ *Solution: the **Switch Distribution***
 - ❖ Simulations & Theorems: consistent + **optimal predictions**
 - ❖ Cumulative risk

The Best of Both Worlds

- ❖ **Catch-up phenomenon**: new explanation for poor predictions of Bayes (and other BIC-like methods)
- ❖ We want a model selection / averaging method that, in a wide variety of circumstances,
 - ❖ is provably **consistent**,
 - ❖ provably achieves **optimal convergence rates**
- ❖ But it has previously been suggested that this is impossible!
[Yang '05]

The Best of Both Worlds

- ❖ **Catch-up phenomenon**: new explanation for poor predictions of Bayes (and other BIC-like methods)
- ❖ We want a model selection / averaging method that, in a wide variety of circumstances,
 - ❖ is provably **consistent**,
 - ❖ provably achieves **optimal convergence rates**
- ❖ **But it has previously been suggested that this is impossible!**
[Yang '05]
- ❖ So we have to be careful to avoid impossibility results...

The Switch Distribution

- ✧ To avoid the catch-up phenomenon we would like to switch between models at switch-point s :

$$p_{\text{sw}}(x^n | s) := \prod_{i=1}^s \bar{p}_1(x_i | x^{i-1}) \times \prod_{i=s+1}^n \bar{p}_2(x_i | x^{i-1})$$

- ✧ Q. But how do we know when to switch?!

The Switch Distribution

- ✧ To avoid the catch-up phenomenon we would like to switch between models at switch-point s :

$$p_{\text{sw}}(x^n | s) := \prod_{i=1}^s \bar{p}_1(x_i | x^{i-1}) \times \prod_{i=s+1}^n \bar{p}_2(x_i | x^{i-1})$$

- ✧ Q. But how do we know when to switch?!
- ✧ A. **Switch distribution**: do not put a prior π on models, but on when to switch between models:

$$p_{\text{sw}}(x^n) := \sum_{s \geq 0} p_{\text{sw}}(x^n | s) \pi(s)$$

The Switch Distribution

- ✧ To avoid the catch-up phenomenon we would like to switch between models at switch-point s :

$$p_{\text{sw}}(x^n | s) := \prod_{i=1}^s \bar{p}_1(x_i | x^{i-1}) \times \prod_{i=s+1}^n \bar{p}_2(x_i | x^{i-1})$$

- ✧ Q. But how do we know when to switch?!
- ✧ A. **Switch distribution**: do not put a prior π on models, but on when to switch between models:

$$p_{\text{sw}}(x^n) := \sum_{s \geq 0} p_{\text{sw}}(x^n | s) \pi(s)$$

- ✧ Generalizes to an arbitrary (unknown) number of switches between any countable number of models.

The Switch Distribution

- ❖ To avoid the catch-up phenomenon we would like to switch between models at switch-point s :

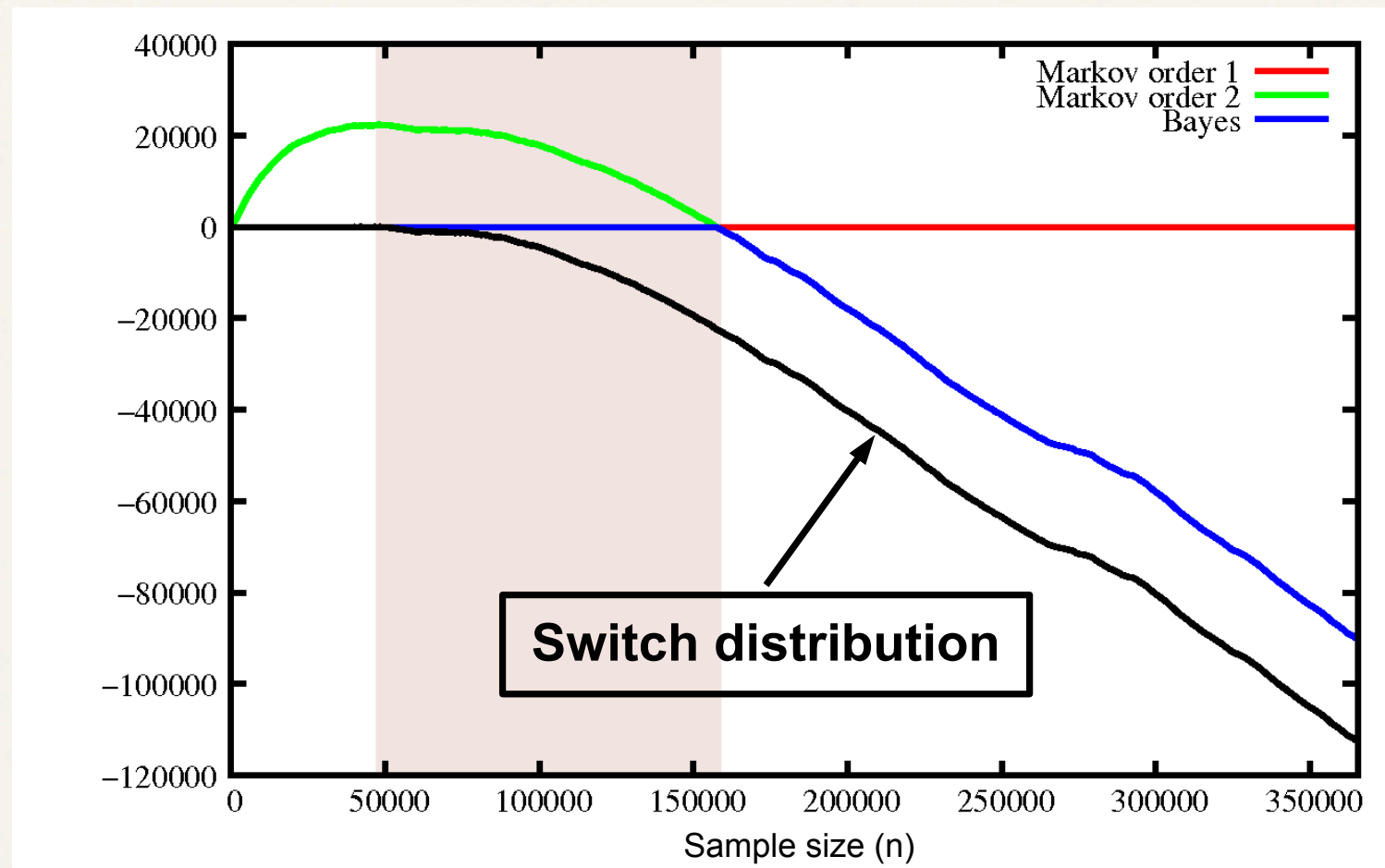
$$p_{\text{sw}}(x^n | s) := \prod_{i=1}^s \bar{p}_1(x_i | x^{i-1}) \times \prod_{i=s+1}^n \bar{p}_2(x_i | x^{i-1})$$

- ❖ Q. But how do we know when to switch?!
- ❖ A. **Switch distribution**: do not put a prior π on models, but on when to switch between models:

$$p_{\text{sw}}(x^n) := \sum_{s \geq 0} p_{\text{sw}}(x^n | s) \pi(s)$$

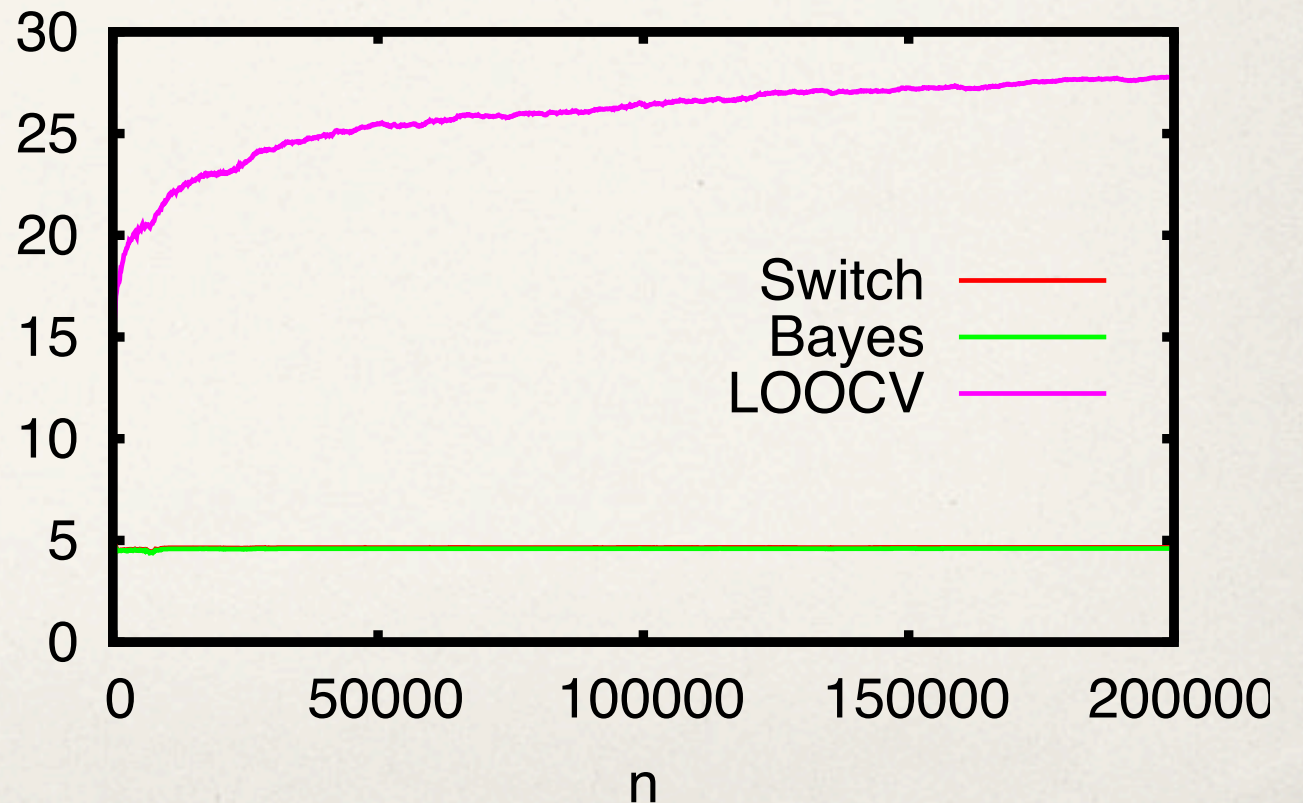
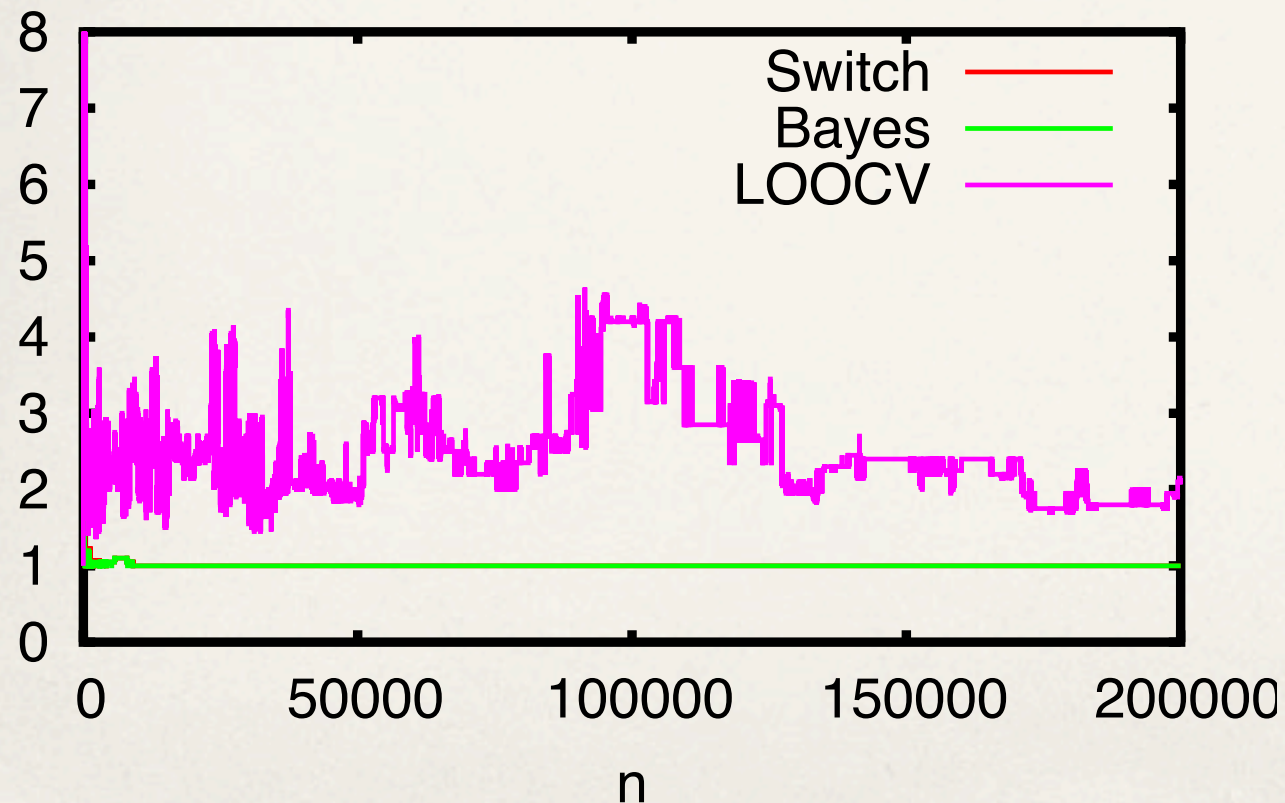
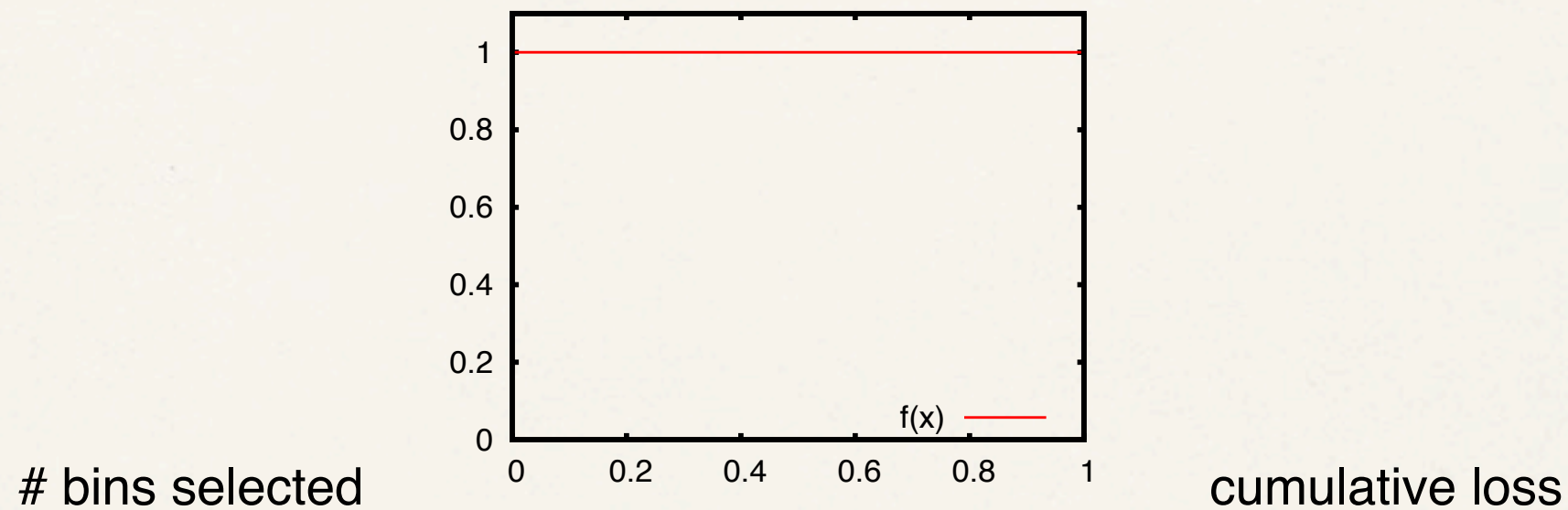
- ❖ Generalizes to an arbitrary (unknown) number of switches between any countable number of models.
- ❖ For many model classes, method is **computationally feasible**.

Switching Resolves the Catch-up Phenomenon

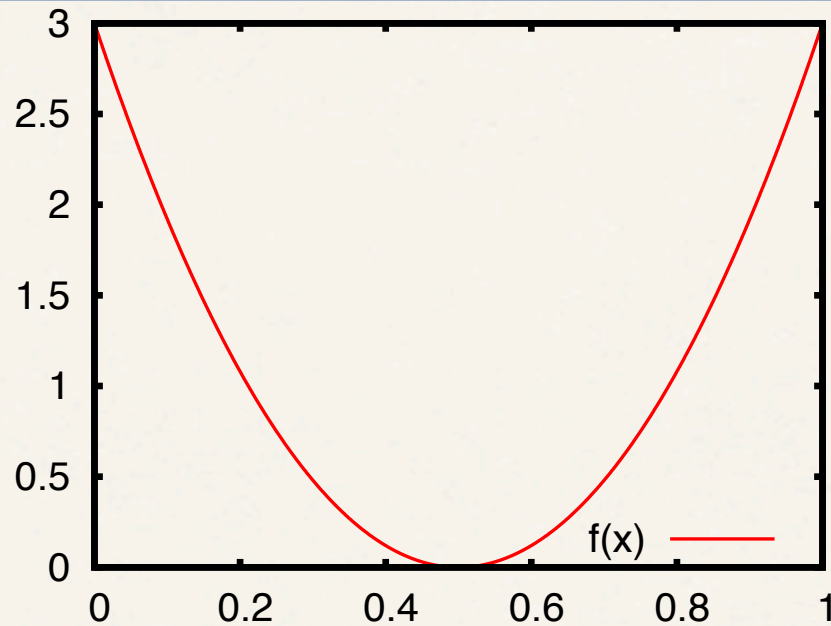


- ❖ Pay less than $2 \log(s + 1) = 2 \log(50\,001) \approx 32$ bits for not knowing s
- ❖ Gain more than 20 000 bits by switching
- ❖ Almost as good as knowing in advance when to switch!

Switch Distribution is Consistent for Histograms

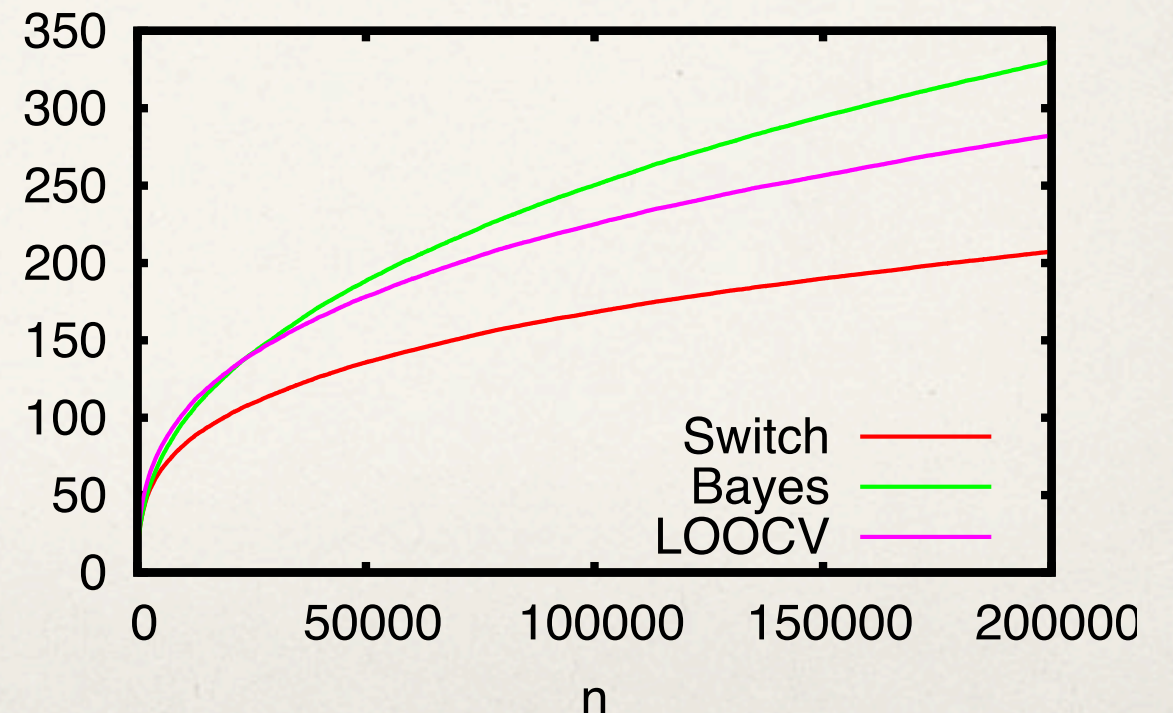
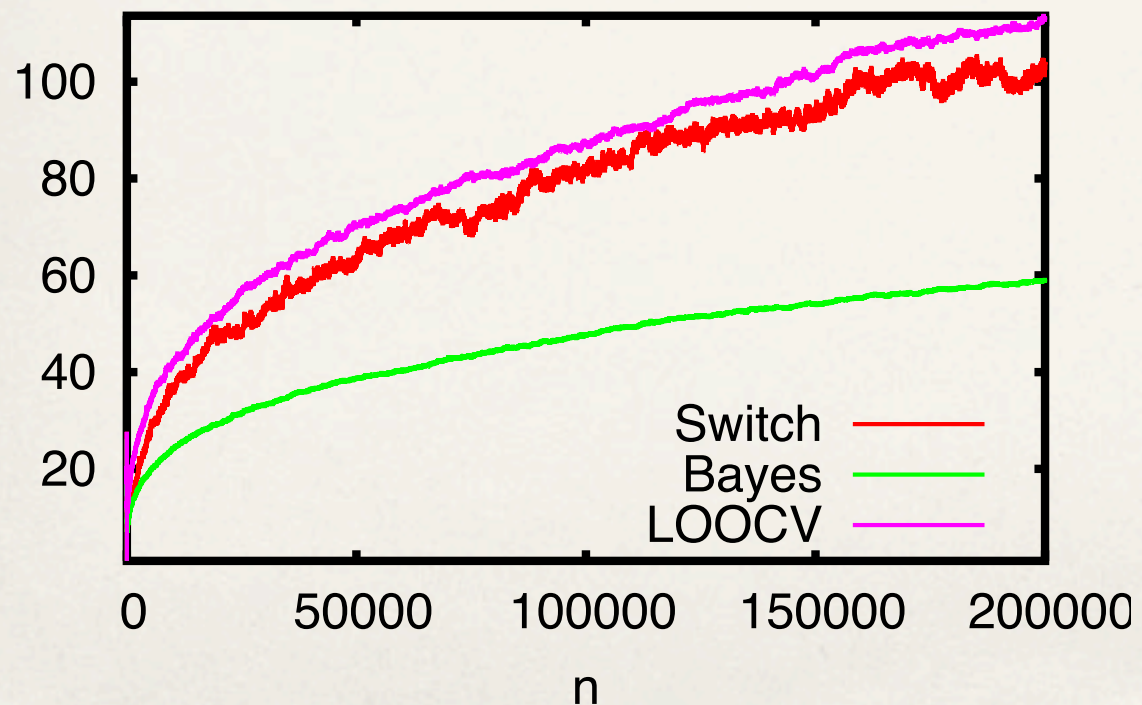


Switch Distribution Predicts Well with Histograms



bins selected

cumulative loss



Theorem: Switching is Consistent

- ♦ Let $\mathcal{M}_1, \mathcal{M}_2, \dots$ be models with priors w_1, w_2, \dots on parameter sets $\Theta_1, \Theta_2, \dots$ and marginal likelihoods $\bar{p}_1, \bar{p}_2, \dots$
- ♦ Suppose $\bar{p}_1, \bar{p}_2, \dots$ are **asymptotically sufficiently distinguishable** in a suitable sense.
 - ♦ For example, it is sufficient if the models consist of i.i.d. or Markov distributions, the parameter sets $\Theta_1, \Theta_2, \dots$ are of different dimensionality and the priors have a density w.r.t. Lebesgue measure.
- ♦ Then, for all k^* and all $p^* \in \mathcal{M}_{k^*}$, except for a subset of \mathcal{M}_{k^*} with prior w_{k^*} -probability 0, the switch distribution is **consistent** in that

$$\lim_{n \rightarrow \infty} p_{\text{sw}}(\mathcal{M}_{k^*} \mid X^n) = 1$$

with p^* -probability 1.

Setting for Prediction

- ✧ Let $\mathcal{M}_1, \mathcal{M}_2, \dots$ be i.i.d. models that can approximate a large set of i.i.d. distributions \mathcal{M}^* arbitrarily well (in Kullback-Leibler divergence)
- ✧ For example, \mathcal{M}^* may be the set of all densities on $[0,1]$ with bounded derivatives and $\mathcal{M}_1, \mathcal{M}_2, \dots$ may be histograms
- ✧ Suppose data $X^n = (X_1, \dots, X_n)$ are i.i.d. with distribution $p^* \in \mathcal{M}^*$

Risk

- ✧ Let $p_{X^{n-1}}$ be the prediction of outcome X_n for some **estimator** p
 - ✧ For example, p may be based on the Bayesian marginal likelihood
- ✧ The **risk** is the expected divergence of the predictions of p from p^* :

$$r_n(p^*, p) := \mathbf{E}_{X^{n-1} \sim p^*} D(p^* \| p_{X^{n-1}})$$

- ✧ We take D to be the Kullback-Leibler divergence:

$$D(p^* \| p_{X^{n-1}}) = \mathbf{E}_{X_n \sim p^*} [\text{loss}(p_{X^{n-1}}, X_n) - \text{loss}(p^*, X_n)]$$

Cumulative Risk

The **cumulative risk** is

$$R_n(p^*, p) = \sum_{i=1}^n r_i(p^*, p) = \mathbf{E}_{X^n} \left[\sum_{i=1}^n \text{loss}(p_{X^{i-1}}, X_i) - \sum_{i=1}^n \text{loss}(p^*, X_i) \right]$$

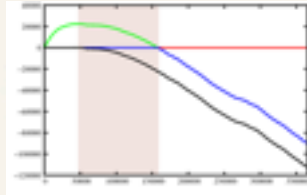
Motivation:

- ✧ Appropriate when the goal is sequential prediction
- ✧ Can convert to ordinary risk via online-to-batch conversion [Yang, Barron, '99]
- ✧ Equals redundancy in universal coding
- ✧ Avoids Yang's impossibility results

Theorem: Switching Achieves Minimax Cumulative Rate

- ✧ Let $\bar{p}_1, \bar{p}_2, \dots$ be estimators for the models $\mathcal{M}_1, \mathcal{M}_2, \dots$.
An **oracle** chooses model $k^\circ \equiv k^\circ(p^*, X^n)$, knowing the true distribution and the data.
- ✧ Suppose the cumulative risk of the oracle grows fast enough that
$$\frac{(\log n)^{2+\alpha}}{\sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}_{k^\circ})} \rightarrow 0$$
for some $\alpha > 0$ and the effective number of models is polynomial in n , i.e. $k^\circ(p^*, X^n) \leq n^\beta$ for some $\beta > 0$.
- ✧ Then the switch distribution, with suitable prior π , **predicts at least as well as the oracle**:
$$\limsup_{n \rightarrow \infty} \frac{\sup_{p^* \in \mathcal{M}^*} R_n(p^*, p_{\text{sw}})}{\sup_{p^* \in \mathcal{M}^*} R_n(p^*, \bar{p}_{k^\circ})} \leq 1.$$

Outline

- ❖ **Bayes Factors** and **MDL** Model Selection
 - ❖ Consistent, but **suboptimal predictions**
- ❖ Explanation: the **Catch-up Phenomenon**
 - ❖ Predictive MDL interpretation of Bayes factors
 - ❖ Markov chain example 
- ❖ Solution: the **Switch Distribution**
 - ❖ Simulations & Theorems: consistent + **optimal predictions**
 - ❖ Cumulative risk

Conclusion

- ❖ Bayes and other BIC-like methods select the model that minimizes cumulative prediction error.
- ❖ If the best-predicting model depends on the sample size, then they suffer from the **catch-up phenomenon**.
- ❖ This explains the **AIC-BIC dilemma**.
- ❖ The switch-distribution provably resolves the catch-up phenomenon:

	Consistent	Optimal rate of convergence
BIC, Bayes, MDL	Yes	No
AIC, LOO Cross-validation	No	Yes
Switch distribution	Yes	Yes (for cumulative risk)

References

T. van Erven, P. Grünwald and S. de Rooij,
Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC-BIC dilemma,
Journal of the Royal Statistical Society, Series B, vol. 74, no. 3, pp. 361-417, 2012

MATLAB software available from my website: www.timvanerven.nl/publications/

Other:

- * A.P. Dawid, *Statistical theory: the prequential approach*, Journal of the Royal Statistical Society, Series A 147, Part 2 (1984), 278-292
- * D. Haussler and M. Oppen, *Mutual information, metric entropy and cumulative relative entropy risk*, The Annals of Statistics, Vol. 25, no. 6 (1997), 2451-2492
- * J. Rissanen, *Universal coding, information, prediction, and estimation*, IEEE Transactions on Information Theory IT-30, no. 4 (1984), 629-636
- * B. Yu and T. P. Speed, *Data compression and histograms*, Probability Theory and Related Fields 92 (1992), 195-229
- * Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Annals of Statistics, Vol. 27, no. 5 (1999), 1564-1599
- * Y. Yang, *Can the strengths of AIC and BIC be shared?*, Biometrika 92(4), 2005, 937-950

Bayesian Prediction

- ✧ Given model $\mathcal{M}_k = \{p_\theta | \theta \in \Theta_k\}$ with prior w_k and data $x^n = (x_1, \dots, x_n)$, the Bayesian **marginal likelihood** is

$$\bar{p}_k(x^n) \equiv p(x^n | \mathcal{M}_k) := \int_{\Theta_k} p_\theta(x^n) w_k(\theta) d\theta$$

- ✧ Given \mathcal{M}_k predict with **estimator**

$$\bar{p}_k(x_{n+1} | x^n) = \frac{\bar{p}_k(x^{n+1})}{\bar{p}_k(x^n)} = \int_{\Theta_k} p_\theta(x_{n+1} | x^n) w_k(\theta | x^n) d\theta$$

- ✧ If k is unknown, **Bayesian model averaging** also puts a prior π on k :

$$p(x_{n+1} | x^n) = \sum_k \bar{p}_k(x_{n+1} | x^n) \pi(k | x^n)$$