Inconsistency of Bayesian inference when the model is wrong, and how to repair it

Peter Grünwald Thijs van Ommen

CWI

Centrum Wiskunde & Informatica, Amsterdam

Universiteit Leiden

Universiteit Leiden

June 3, 2015

Outline

Outline

1. Introduction
2. Bayes when the Model is Wrong
3. Learning rates and SafeBayes
4. Discussion and Conclusion

Setup

We have one or more models:

- Each model is a set of hypotheses;
- Each hypothesis is a probability distribution

Setup

We have one or more models:

- Each model is a set of hypotheses;
- Each hypothesis is a probability distribution

We want to learn from the training data which of these distributions we can use to predict new data

Setup: Regression

We will consider regression models:

$$\mathcal{M}_k = \{p_{(k,\beta,\sigma^2)} \mid \beta \in \mathbf{R}^{k+1}, \sigma > 0\};$$

Setup: Regression

We will consider regression models:

$$\mathcal{M}_k = \{p_{(k,\beta,\sigma^2)} \mid \beta \in \mathbf{R}^{k+1}, \sigma > 0\};$$

Hypothesis $p_{(k,\beta,\sigma^2)}$ expresses that

$$Y \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{k} \beta_i g_i(X), \sigma^2\right)$$

In this presentation: g_i is a polynomial of degree i
So model $\mathcal{M}_k$ represents *all* polynomials of degree up to k

Bayesian statistics

Big idea: Use probability distributions over the *models and hypotheses* to represent our uncertainty

Bayesian statistics

Big idea: Use probability distributions over the *models and hypotheses* to represent our uncertainty

- Introduce a prior distribution

$$\pi(k, \beta, \sigma^2)$$

Bayesian statistics

Big idea: Use probability distributions over the *models and hypotheses* to represent our uncertainty

- Introduce a prior distribution

$$\pi(k, \beta, \sigma^2)$$

- so that we can define the joint distribution:

$$p_{\text{Bayes}}(Y^n, k, \beta, \sigma^2 \mid X^n) = p_{(k,\beta,\sigma^2)}(Y^n \mid X^n)\pi(k, \beta, \sigma^2)$$

Bayesian statistics: Posterior and predictive

We can use the joint distribution p_{Bayes} to compute interesting things:

- The Bayesian posterior distribution

$$\pi(k, \beta, \sigma^2 \mid X^n, Y^n)$$

  tells us how to update our prior beliefs after have seen the data

Bayesian statistics: Posterior and predictive

We can use the joint distribution p_{Bayes} to compute interesting things:

- The Bayesian posterior distribution

$$\pi(k, \beta, \sigma^2 \mid X^n, Y^n)$$

  tells us how to update our prior beliefs after have seen the data
- The Bayesian predictive distribution

$$p_{\text{Bayes}}(Y_i \mid Y^{i-1}, X^i)$$

  tells what new data should look like, based on that posterior

Bayesian statistics: Advantages

Bayesian methods are very successful, in both theory and practice:

- Keep track of uncertainty in a very elegant way;

Bayesian statistics: Advantages

Bayesian methods are very successful, in both theory and practice:

- Keep track of uncertainty in a very elegant way;
- Incorporate this uncertainty into (mixed) predictions;

Bayesian statistics: Advantages

Bayesian methods are very successful, in both theory and practice:

- Keep track of uncertainty in a very elegant way;
- Incorporate this uncertainty into (mixed) predictions;
- Bayes avoids overfitting

Bayesian statistics: Advantages

Bayesian methods are very successful, in both theory and practice:

- Keep track of uncertainty in a very elegant way;
- Incorporate this uncertainty into (mixed) predictions;
- Bayes avoids overfitting

. . . usually (see next section)

Experiment: Model correct

Experiment: We let Bayes choose from 51 different models (polynomials of degrees 0 up to 50);
The data are actually drawn according to a distribution P^* (the true distribution), which is in the simplest model:

$$X \sim U(-1, 1);$$
$$Y \sim \mathcal{N}(f^*(X), 0.05)$$

with $f^*(x) = 0$ for all x.

Experiment: Model correct

Experiment: We let Bayes choose from 51 different models (polynomials of degrees 0 up to 50);
The data are actually drawn according to a distribution P^* (the true distribution), which is in the simplest model:

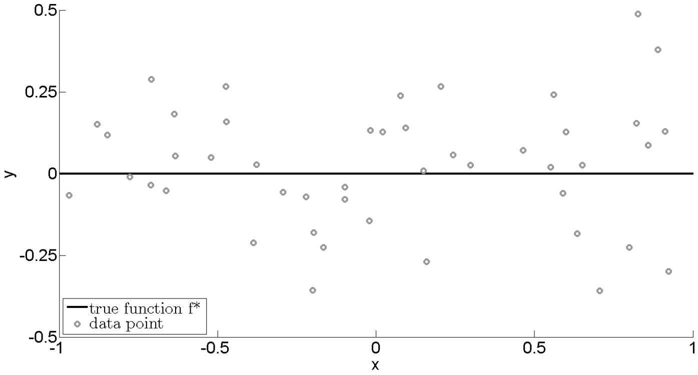
$$X \sim U(-1, 1);$$
$$Y \sim \mathcal{N}(f^*(X), 0.05)$$

with $f^*(x) = 0$ for all x.

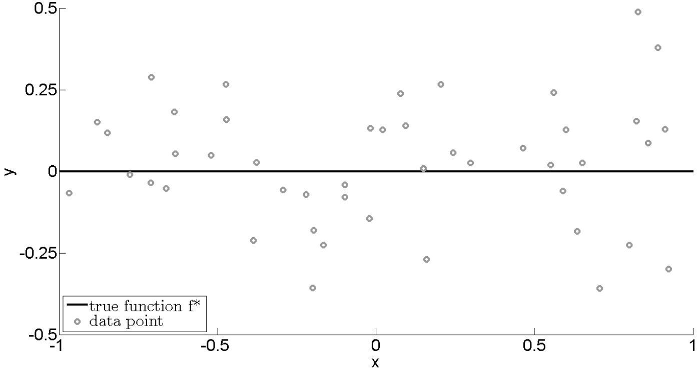
We use standard priors:

- More-or-less uniform on k;
- Gaussian with large variance on β;
- Inverse-gamma on σ^2.

Experiment: Model correct

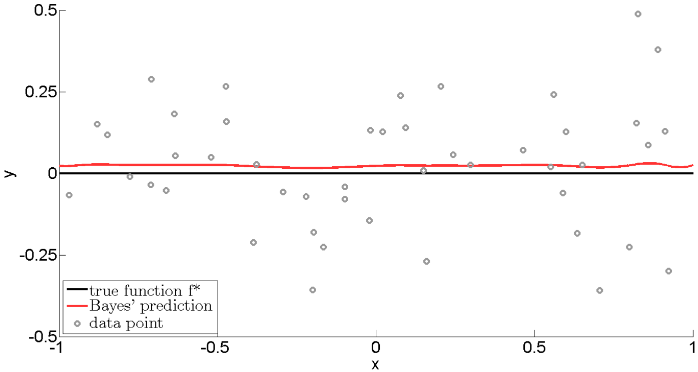


Experiment: Model correct



- For these data, Bayes puts most weight on *smallest model*

Experiment: Model correct



- For these data, Bayes puts most weight on *smallest model*

Outline

Experiment: Model wrong

New experiment:

- Same models;
- Different true distribution:
  For each data point, flip a fair coin
  - if heads, data point is drawn randomly as before;
  - if tails, data point is exactly at $(0, 0)$

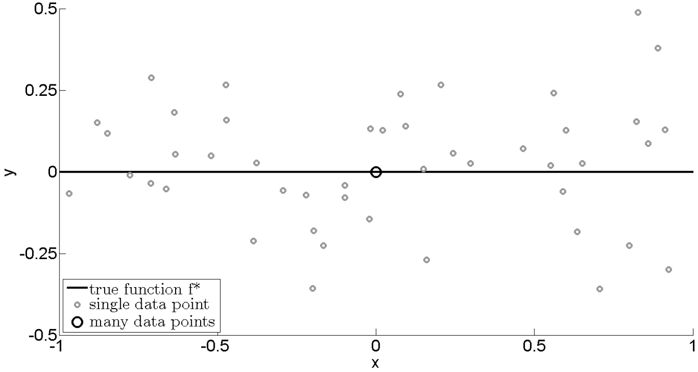
Experiment: Model wrong

New experiment:

- Same models;
- Different true distribution:
  For each data point, flip a fair coin
  - if heads, data point is drawn randomly as before;
  - if tails, data point is exactly at $(0, 0)$

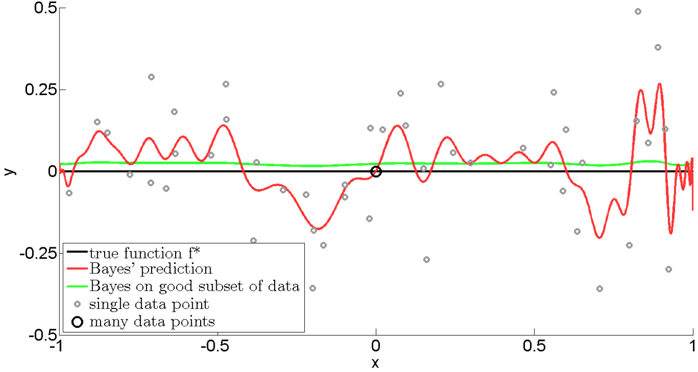
Simplest model is still best! (in a sense we will see later)

Experiment: Model wrong



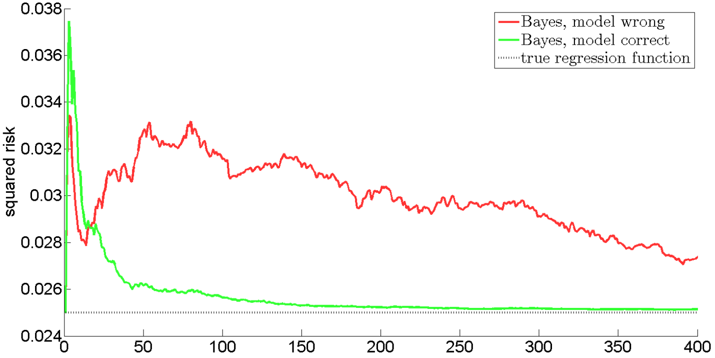
- This should just make it easier, right?

Experiment: Model wrong



- This should just make it easier, right?
- Now Bayes puts most weight on *largest models*!

Experiment: Model wrong



- This should just make it easier, right?
- Now Bayes puts most weight on *largest models*!

Results from other experiments

We ran many more experiments, eg.

- different models;
- different priors;
- different true distributions.

The problems with Bayes occur in all of them.

Results from other experiments

We ran many more experiments, eg.

- different models;
- different priors;
- different true distributions.

The problems with Bayes occur in all of them.
Problems get worse if there are more models;

- By comparison: In model-correct experiment, Bayes is hardly affected by extra models

KL divergence

If the Bayesian posterior concentrates, it is around the hypothesis $\tilde{P}$ that is closest to P^* in terms of KL divergence among all elements in the model:

$$D(P^* \| \tilde{P}) = \mathbf{E}_{X,Y \sim P^*}[-\log \tilde{P}(Y \mid X)] - C_{P^*}$$

[Kleijn and Van der Vaart 2006]

KL divergence

If the Bayesian posterior concentrates, it is around the hypothesis $\tilde{P}$ that is closest to P^* in terms of KL divergence among all elements in the model:

$$D(P^* \| \tilde{P}) = \mathbf{E}_{X,Y \sim P^*}[-\log \tilde{P}(Y \mid X)] - C_{P^*}$$

[Kleijn and Van der Vaart 2006]

In our experiment, $\tilde{P}$ is the hypothesis that

- predicts $Y = 0$ for all X (coincides with f^*);
- sets $\sigma^2 = 0.025$ ($=$ variance of Y).

This $\tilde{P}$ also minimizes the squared risk!
Conclusion: Bayesian posterior did not concentrate!

Outline

Introducing Generalized Bayes

Bayes:

$$\pi(\theta \mid \text{data}) \propto p(\text{data} \mid \theta) \cdot \pi(\theta)$$

On-line prediction; PAC-Bayes; Lasso/Ridge; . . . :

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \cdot \text{loss}_{\theta}(\text{data})} \cdot \pi(\theta)$$

(η: 'learning rate')

Introducing Generalized Bayes

Bayes:

$$\pi(\theta \mid \text{data}) \propto p(\text{data} \mid \theta) \cdot \pi(\theta)$$
$$= e^{-\text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

for $\text{loss}_\theta(\text{data}) = -\log p(\text{data} \mid \theta)$

On-line prediction; PAC-Bayes; Lasso/Ridge; . . . :

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \cdot \text{loss}_\theta(\text{data})} \cdot \pi(\theta)$$

(η: 'learning rate')

Introducing Generalized Bayes

Generalized Bayes: [Vovk 1990; Barron & Cover 1991; Walker & Hjort 2002; McAllister 2003; ...]

$$\pi(\theta \mid \text{data}) \propto p(\text{data} \mid \theta)^{\eta} \cdot \pi(\theta)$$

$$= e^{-\eta \cdot \text{loss}_{\theta}(\text{data})} \cdot \pi(\theta)$$

for $\text{loss}_{\theta}(\text{data}) = -\log p(\text{data} \mid \theta)$

On-line prediction; PAC-Bayes; Lasso/Ridge; ...:

$$\pi(\theta \mid \text{data}) \propto e^{-\eta \cdot \text{loss}_{\theta}(\text{data})} \cdot \pi(\theta)$$

(η: 'learning rate')

Choosing the learning rate

- $\eta = 1$: standard Bayes
- $\eta = 0$: no learning occurs (posterior remains equal to prior)

Choosing the learning rate

- $\eta = 1$: standard Bayes
- $\eta = 0$: no learning occurs (posterior remains equal to prior)
- $\eta \in (0, 1]$ small enough: posterior concentrates again, even when model is wrong! eg. [Zhang 2006]

But if η too small, we are learning more slowly than we could

Choosing the learning rate

- $\eta = 1$: standard Bayes
- $\eta = 0$: no learning occurs (posterior remains equal to prior)
- $\eta \in (0, 1]$ small enough: posterior concentrates again, even when model is wrong! eg. [Zhang 2006]

But if η too small, we are learning more slowly than we could

Theoretical prescriptions for η are often suboptimal in practice

Choosing the learning rate

- $\eta = 1$: standard Bayes
- $\eta = 0$: no learning occurs (posterior remains equal to prior)
- $\eta \in (0, 1]$ small enough: posterior concentrates again, even when model is wrong! eg. [Zhang 2006]

But if η too small, we are learning more slowly than we could

Theoretical prescriptions for η are often suboptimal in practice

Grand aim:
Find generic method (a theory, if you will) for determining the learning rate is all such problems

Bayesian model selection

$$p_{\text{Bayes}}(Y^n \mid X^n, k)$$
$$= \int_{(\beta,\sigma^2)} p_{(k,\beta,\sigma^2)}(Y^n \mid X^n)\pi(\beta,\sigma^2 \mid k)\,\mathrm{d}(\beta,\sigma^2)$$

is the Bayesian marginal probability of the data given model $\mathcal{M}_k$

Bayes factor model selection: from a collection of models $\mathcal{M}_0, \mathcal{M}_1, \ldots, cM_K$, select the model $\mathcal{M}_k$ that maximizes this quantity

Bayesian model selection as forward validation

$$-\log p_{\text{Bayes}}(Y^n \mid X^n, k) = \sum_{i=1}^{n} -\log p_{\text{Bayes}}(Y_i \mid Y^{i-1}, X^i, k)$$

Minus log likelihood = sum of logarithmic prediction errors

[Dawid 1984; Rissanen 1984]

Bayesian model selection as forward validation

$$-\log p_{\text{Bayes}}(Y^n \mid X^n, k) = \sum_{i=1}^{n} -\log p_{\text{Bayes}}(Y_i \mid Y^{i-1}, X^i, k)$$

Minus log likelihood = sum of logarithmic prediction errors

[Dawid 1984; Rissanen 1984]

Viewed this way (‘prequential’), Bayes is similar to leave-one-out cross-validation — but goes through the data in only one direction

The SafeBayesian algorithm

Can we use the same approach to learn η instead of k?

$$
\begin{aligned}
-\log p_{\text{Bayes}}(Y^n \mid X^n, \eta) &= \sum_{i=1}^{n} -\log p_{\text{Bayes}}(Y_i \mid Y^{i-1}, X^i, \eta) \\
&= \sum_{i=1}^{n} -\log \mathbf{E}_{(k,\beta,\sigma^2)\sim\pi \mid Y^{i-1}, X^i, \eta}\left[p_{(k,\beta,\sigma^2)}(Y_i)\right]
\end{aligned}
$$

The SafeBayesian algorithm

Can we use the same approach to learn η instead of k?

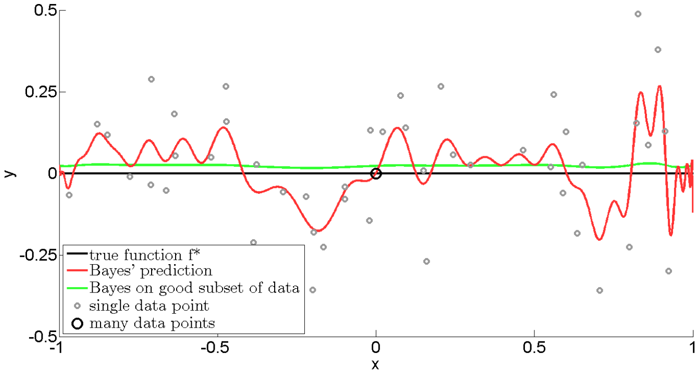
$$-\log p_{\text{Bayes}}(Y^n \mid X^n, \eta) = \sum_{i=1}^{n} -\log p_{\text{Bayes}}(Y_i \mid Y^{i-1}, X^i, \eta)$$

$$= \sum_{i=1}^{n} -\log \mathbf{E}_{(k,\beta,\sigma^2)\sim\pi \mid Y^{i-1}, X^i, \eta}[p_{(k,\beta,\sigma^2)}(Y_i)]$$

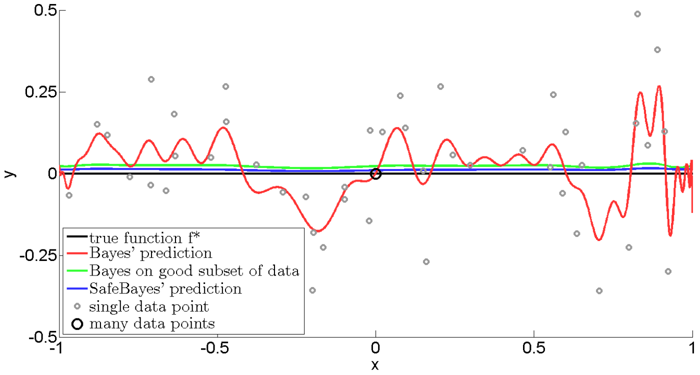
Doesn't work! Instead SafeBayes finds η minimizing

$$= \sum_{i=1}^{n} \mathbf{E}_{(k,\beta,\sigma^2)\sim\pi \mid Y^{i-1}, X^i, \eta} -\log[p_{(k,\beta,\sigma^2)}(Y_i)]$$

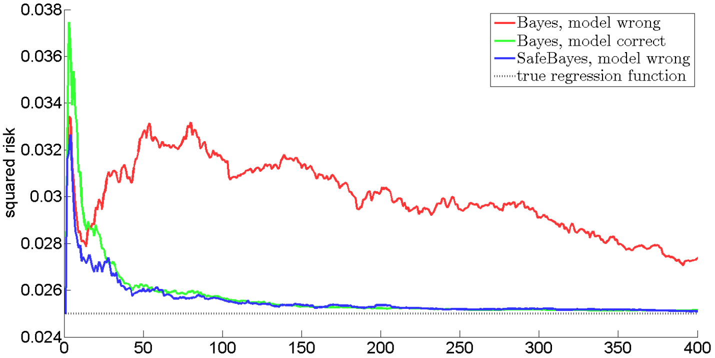
Experiment: Wrong model (continued)



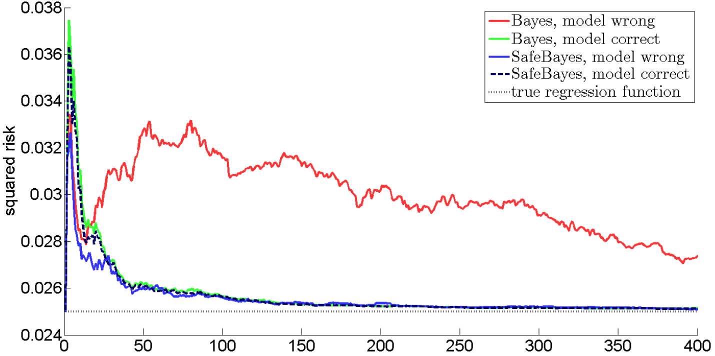
Experiment: Wrong model (continued)



Experiment: Wrong model (continued)



Experiment: Wrong model (continued)



Outline

Bayes and logarithmic loss

When measured in terms of logarithmic loss
($\text{loss}_\theta(\text{data}) = -\log p(\text{data} \mid \theta)$),

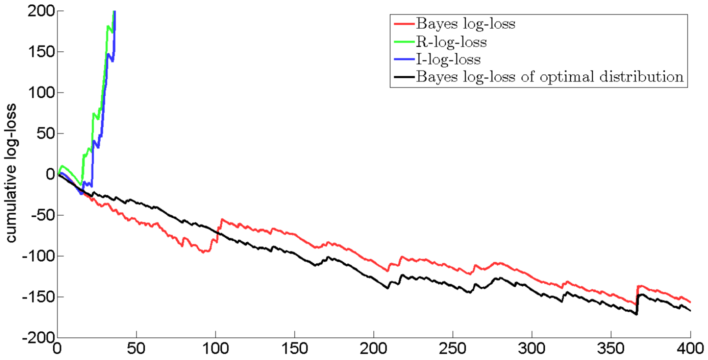
- normally, Bayes learns to predict almost as well as the best element in the model
(slightly worse because we don't know which element is best)

Bayes and logarithmic loss

When measured in terms of logarithmic loss ($\text{loss}_\theta(\text{data}) = -\log p(\text{data} \mid \theta)$),

- normally, Bayes learns to predict almost as well as the best element in the model
  (slightly worse because we don't know which element is best)
- in our case, Bayes predicts significantly better than the best element in the model!
  (in terms of logarithmic loss; not in terms of, say, squared loss)

Bayes is 'too good'!



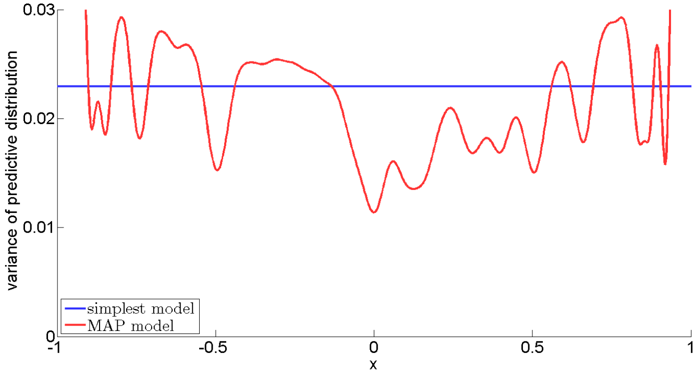
Cumulative logarithmic loss of Bayesian predictive distribution, for $\eta = 1$

How is this possible?

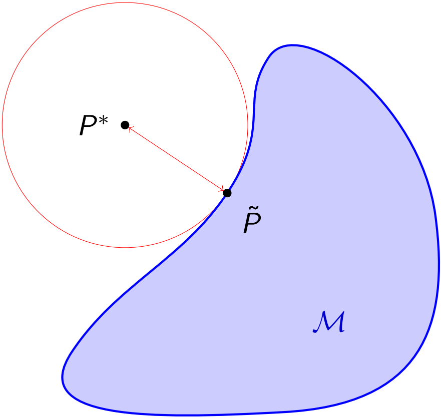
Possible because all elements of model predict with Gaussian distributions, while Bayesian predictive can be infinite mixture of these Gaussians

How is this possible?

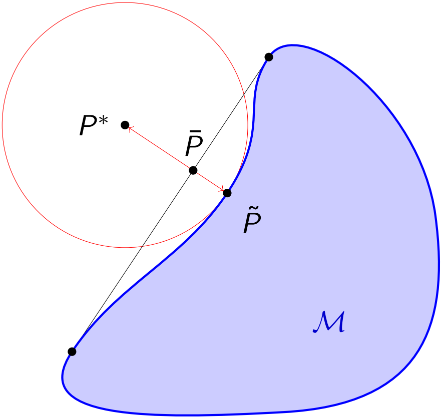
Possible because all elements of model predict with Gaussian distributions, while Bayesian predictive can be infinite mixture of these Gaussians



Bad and good misspecification



Bad and good misspecification



Conclusion

- Standard Bayes may fail to concentrate, even on fairly innocent data
- Generalized Bayes does concentrate, *if* you know the right learning rate
- SafeBayes learns the learning rate!

Thank you!