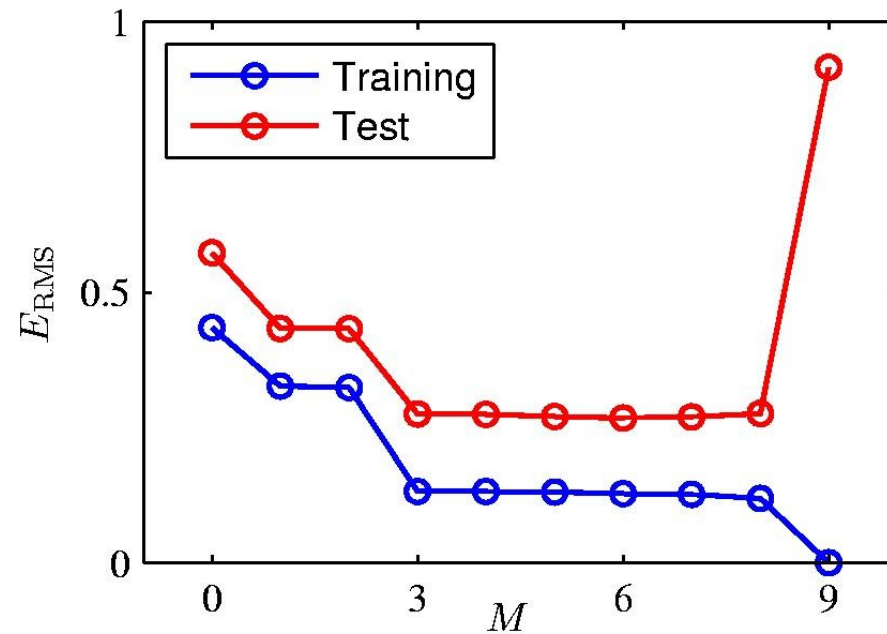


Linear regression - part 2

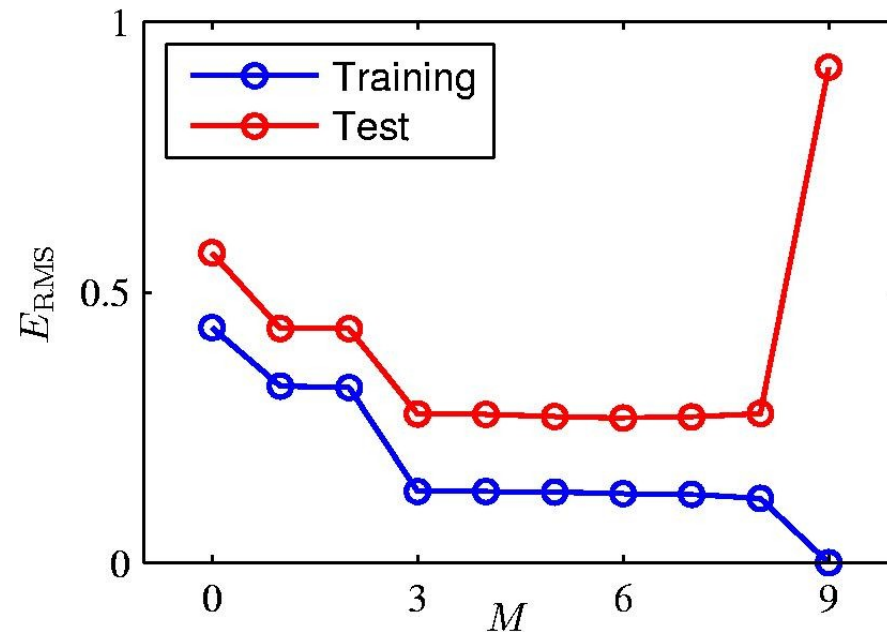


Machine Learning; Fri Apr 20, 2007

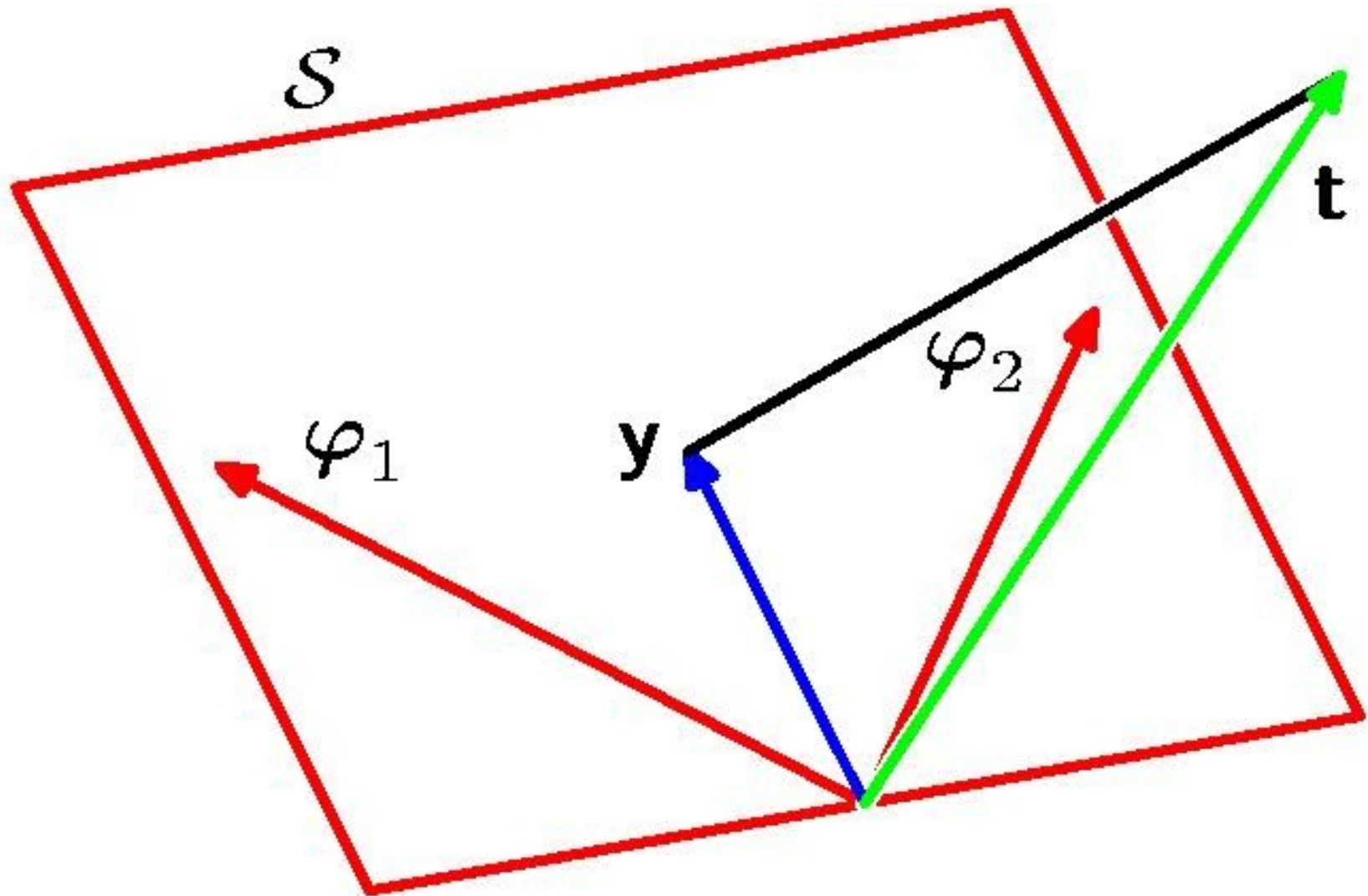
Motivation

Problem: Over-fitting is always a problem when we fit data to generic models.

With nested models, the ML parameters will **never** prefer a simple model over a more complex model...



Maximum likelihood problems



Bayesian model selection

We can take a more Bayesian approach and select model based on posterior model probabilities:

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{D} | \mathcal{M}_i)p(\mathcal{M}_i)$$

Bayesian model selection

We can take a more Bayesian approach and select model based on posterior model probabilities:

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{D} | \mathcal{M}_i)p(\mathcal{M}_i)$$

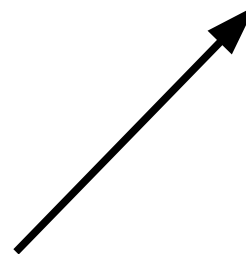
The normalizing factor is the same for all models:

$$p(\mathcal{M}_i | \mathcal{D}) = \frac{p(\mathcal{D} | \mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})}$$

Bayesian model selection

We can take a more Bayesian approach and select model based on posterior model probabilities:

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{D} | \mathcal{M}_i)p(\mathcal{M}_i)$$

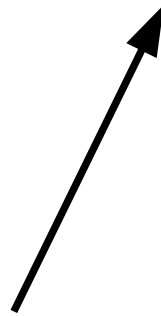


The prior captures our preferences in the models.

Bayesian model selection

We can take a more Bayesian approach and select model based on posterior model probabilities:

$$p(\mathcal{M}_i | \mathcal{D}) \propto p(\mathcal{D} | \mathcal{M}_i)p(\mathcal{M}_i)$$



The likelihood captures the data's preferences in models.

The marginal likelihood

The likelihood of the model is the integral over all the models parameters:

$$p(\mathcal{D} | \mathcal{M}_i) = \int p(\mathcal{D} | \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} | \mathcal{M}_i) d\mathbf{w}$$

The marginal likelihood

The likelihood of the model is the integral over all the models parameters:

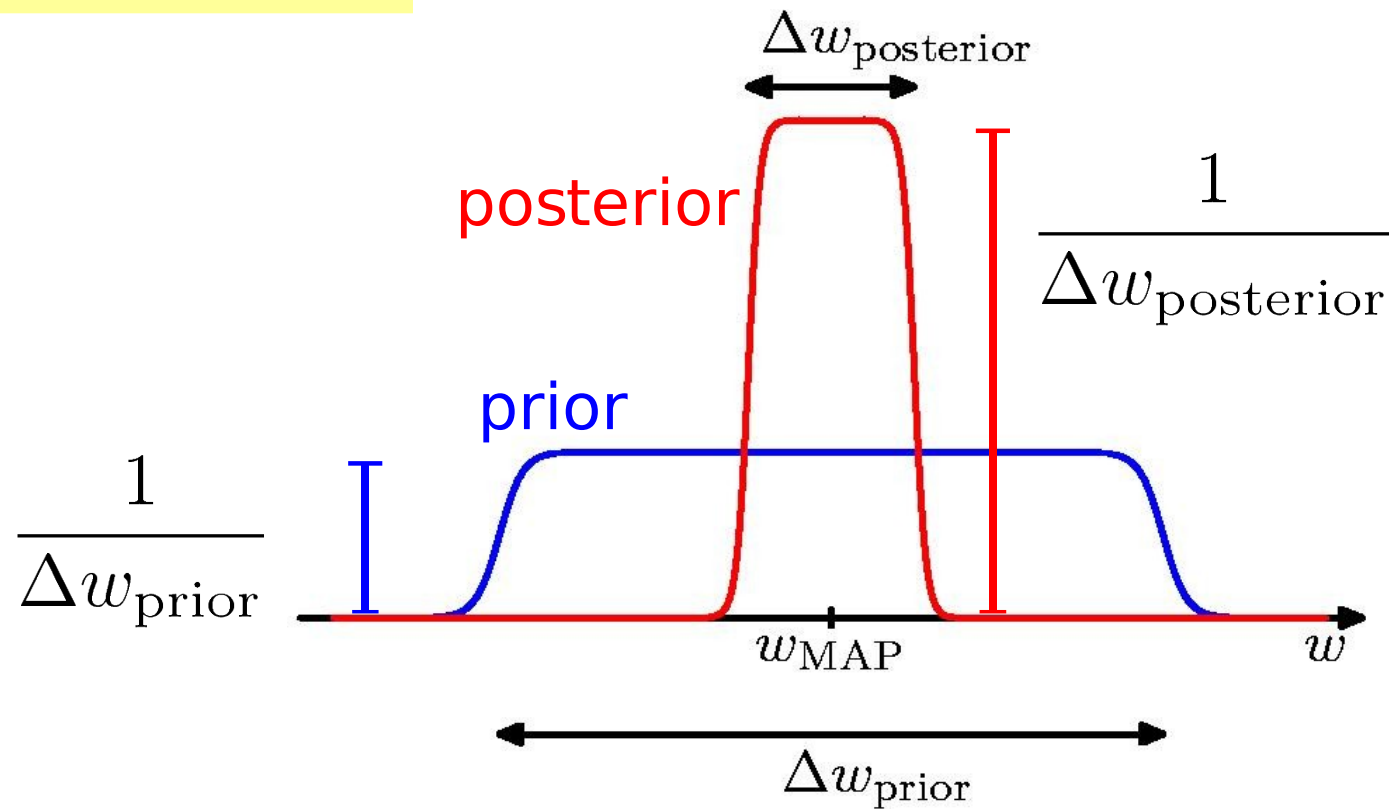
$$p(\mathcal{D} | \mathcal{M}_i) = \int p(\mathcal{D} | \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} | \mathcal{M}_i) d\mathbf{w}$$

which is also the normalizing factor for the posterior:

$$p(\mathbf{w} | \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} | \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} | \mathcal{M}_i)}{p(\mathcal{D} | \mathcal{M}_i)}$$

Implicit over-fitting penalty

Assume this is the shape of prior and posterior

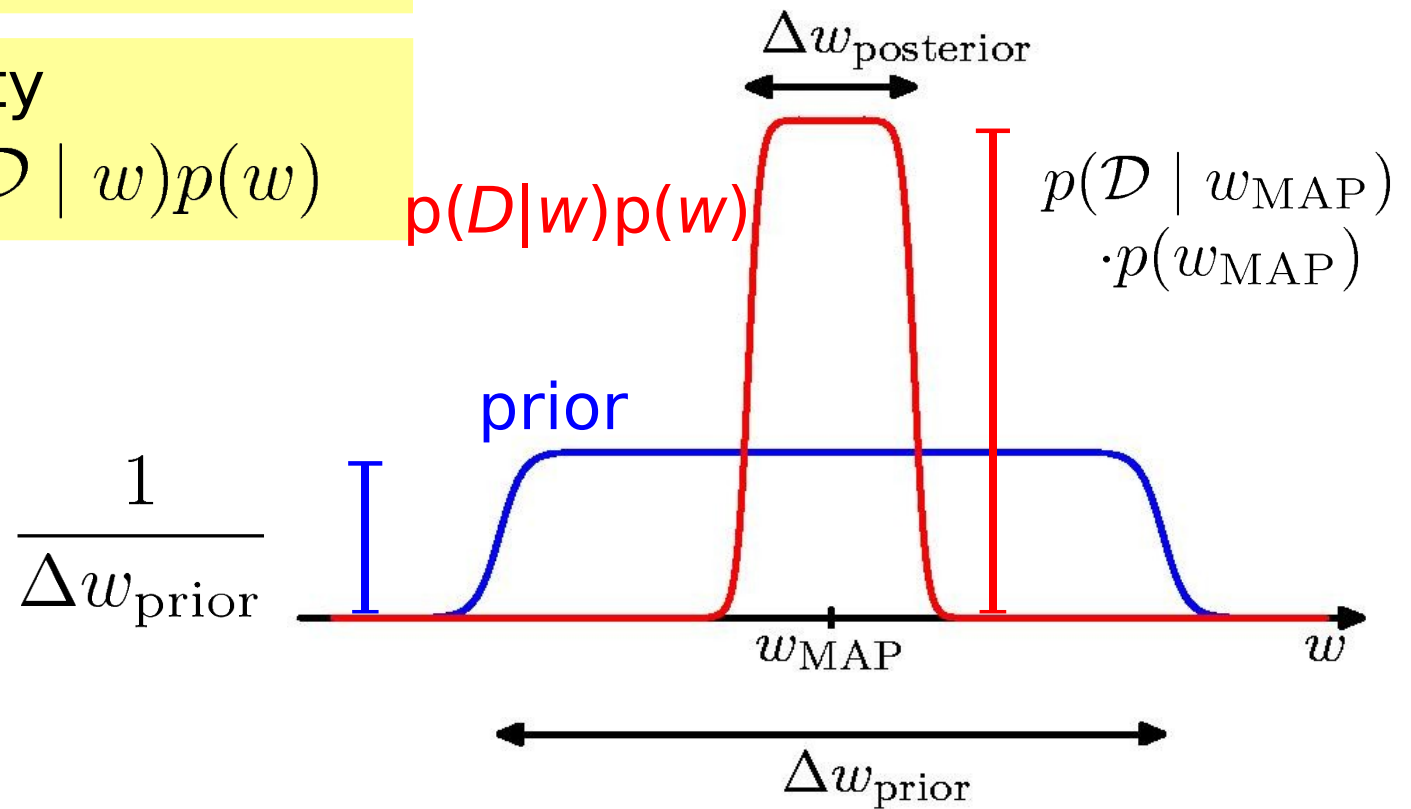


Implicit over-fitting penalty

Assume this is the shape of prior and posterior

By proportionality

$$p(w | \mathcal{D}) \propto p(\mathcal{D} | w)p(w)$$



Implicit over-fitting penalty

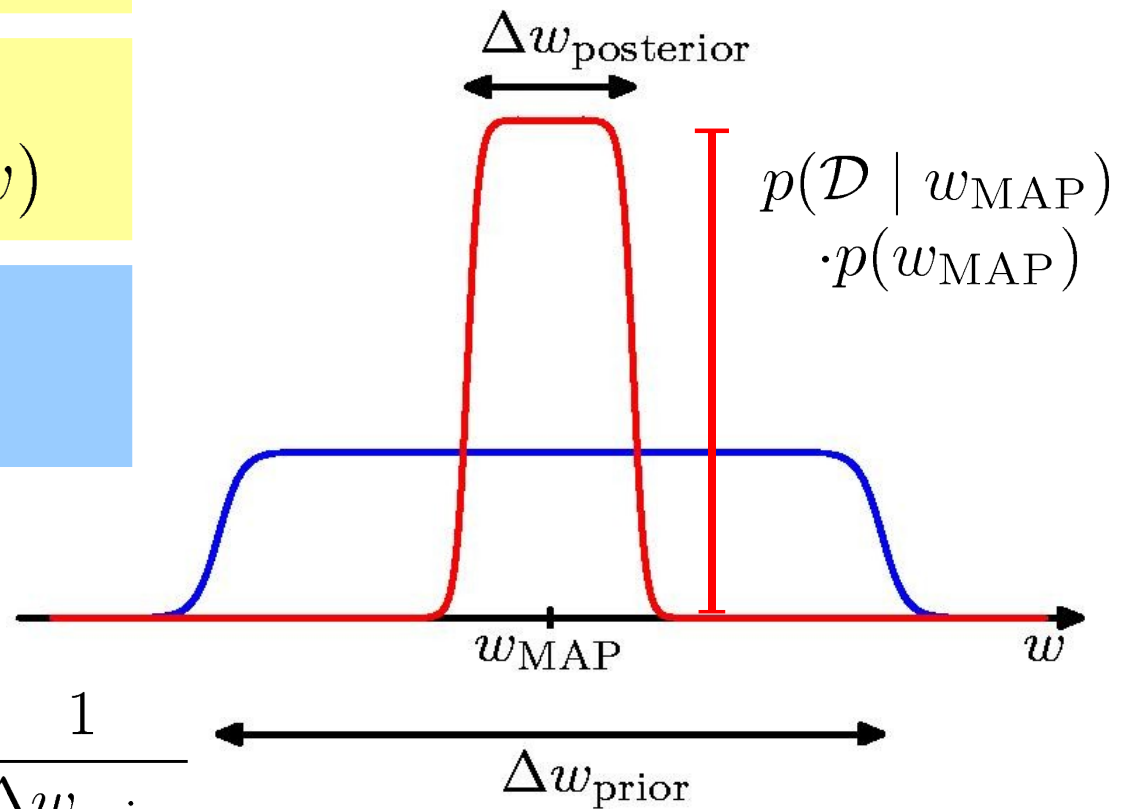
Assume this is the shape of prior and posterior

By proportionality

$$p(w \mid \mathcal{D}) \propto p(\mathcal{D} \mid w)p(w)$$

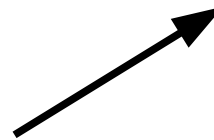
Integral approximately
“width” times “height”

$$\int p(\mathcal{D} \mid w)p(w) dw \approx \Delta w_{\text{posterior}} \cdot p(\mathcal{D} \mid w_{\text{MAP}}) \cdot \frac{1}{\Delta w_{\text{prior}}}$$



Implicit over-fitting penalty

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D} \mid w_{\text{MAP}}) + \log \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$



Increasingly negative as posterior becomes “pointy” compared to prior

Close fitting to data is implicitly penalized, and the marginal likelihood is a trade-off between maximizing the posterior and minimizing this penalty.

Implicit over-fitting penalty

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D} \mid w_{\text{MAP}}) + M \log \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$



Penalty increases with number of parameters M

Close fitting to data is implicitly penalized, and the marginal likelihood is a trade-off between maximizing the posterior and minimizing this penalty.

On average we prefer the true model

This doesn't mean we always prefer the simplest model!

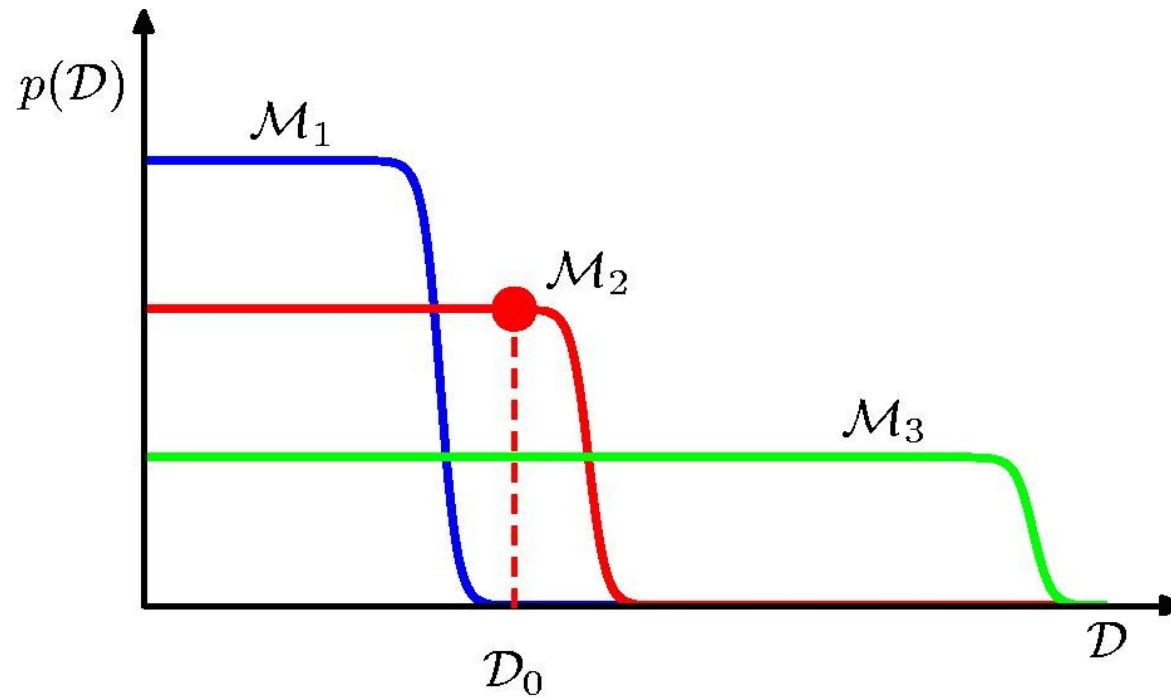
One can show
$$\int p(\mathcal{D} | \mathcal{M}_1) \ln \frac{p(\mathcal{D} | \mathcal{M}_1)}{p(\mathcal{D} | \mathcal{M}_2)} d\mathcal{D} \geq 0$$

with zero only when $\mathcal{M}_1 = \mathcal{M}_2$

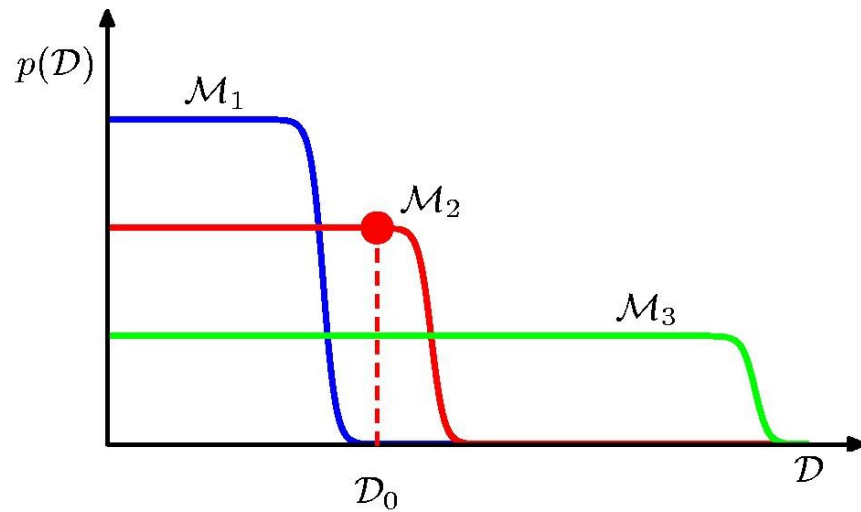
i.e. on average the right model is the preferred model.

Close fitting to data is implicitly penalized, and the marginal likelihood is a trade-off between maximizing the posterior and minimizing this penalty.

**On average we prefer
the true model**



Summary



- Over-fitting is an inherent problem in ML estimation
- Bayesian methods avoid the maximization caused over-fitting problem
 - (but is still vulnerable to model mis-specification)