MODEL SELECTION

Sergey Nikolenko

Harbour Space University, Barcelona, Spain
March 20, 2017

CURSE OF DIMENSIONALITY

CURSE OF DIMENSIONALITY

- k-NN might yield much better results than a linear model, especially once we have chosen a good k.
- Maybe we won't need anything else?
- Let's see how k-NN behaves in high dimension (which is very realistic).

CURSE OF DIMENSIONALITY

- Let us look for nearest neighbors for a point in a unit hypercube. Suppose that the original distribution was uniform.
- To cover share α of test example, we have to cover (in expectation) a share α of the volume, and the expected length of the side of a hypercube neighborhood in dimension p will be $e_p(\alpha) = \alpha^{1/p}$.
- E.g., in dimension 10 $e_{10}(0.1) = 0.8$, $e_{10}(0.01) = 0.63$, i.e., to cover 1% of the volume we have to take a neighborhood of length more than $\frac{1}{2}$ w.r.t. each coordinate!
- This is bad for k-NN computationally too: it's hard to reject with a small number of coordinates, and fast algorithms don't work well.

CURSE OF DIMENSIONALITY

- The second problem from the curse of dimensionality: consider N points uniformly distributed in a unit ball of dimension p.
- The mean distance to zero is

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p},$$

  e.g., in dimension 10 for $N = 500$ $d \approx 0.52$, i.e., more than $\frac{1}{2}$.
- Most points are closer to the "sides" of the support than to other points, which is bad for k-NN: we extrapolate instead of interpolating.

CURSE OF DIMENSIONALITY

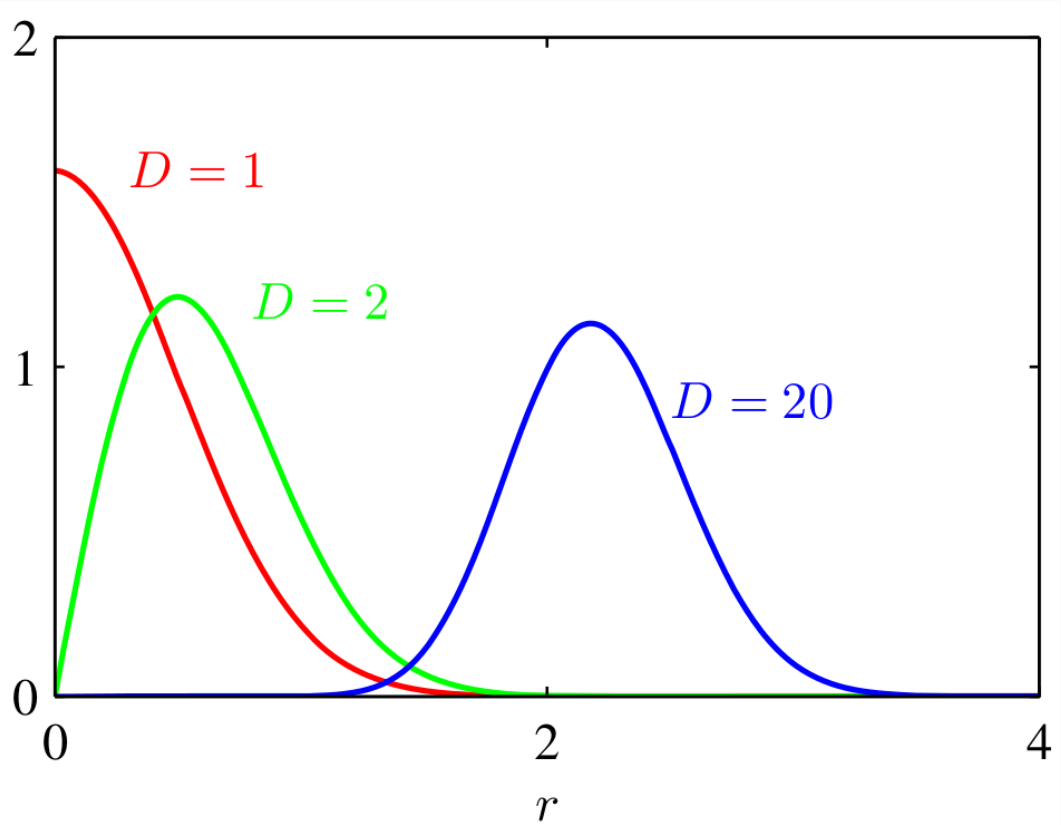
- Third phenomenon: problems in optimization.
- To approximately optimize a function of d variables on a grid with step ϵ, we will need approx. $\left(\frac{1}{\epsilon}\right)^d$ function computations.
- Numerical integration: to integrate a function up to ϵ, we will need $\left(\frac{1}{\epsilon}\right)^d$ computations.

CURSE OF DIMENSIONALITY

- Dense sets become very sparse. E.g., to get the density created in dimension 1 with $N = 100$ points we will need 100^{10} points in dimension 10.
- The behaviour of functions also becomes more complicated as dimension grows: to construct regressions in high dimension with the same accuracy one might need exponentially more points than in low dimension.
- While a, say, linear model does not have any such effects, it's not subject to the curse of dimensionality.

CURSE OF DIMENSIONALITY

- One more example: a normally distributed value will be concentrated in a thin shell.



Exercise. Convert the density of a Gaussian into polar coordinates and check this statement.

EQUIVALENT KERNEL

EQUIVALENT KERNEL

- In linear regression we had

$$p(t \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \mu_N^\top \phi(\mathbf{x}), \sigma_N^2),$$
$$\text{where } \sigma_N^2 = \frac{1}{\beta} + \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}).$$

- Let us rewrite the mean of the posterior as (recall that $\mu_N = \beta \Sigma_N \Phi^\top \mathbf{t}$):

$$y(\mathbf{x}, \mu_N) = \mu_N^\top \phi(\mathbf{x}) = \beta \phi(\mathbf{x})^\top \Sigma_N \Phi^\top \mathbf{t} = \\ = \sum_{n=1}^{N} \beta \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}_n) t_n.$$

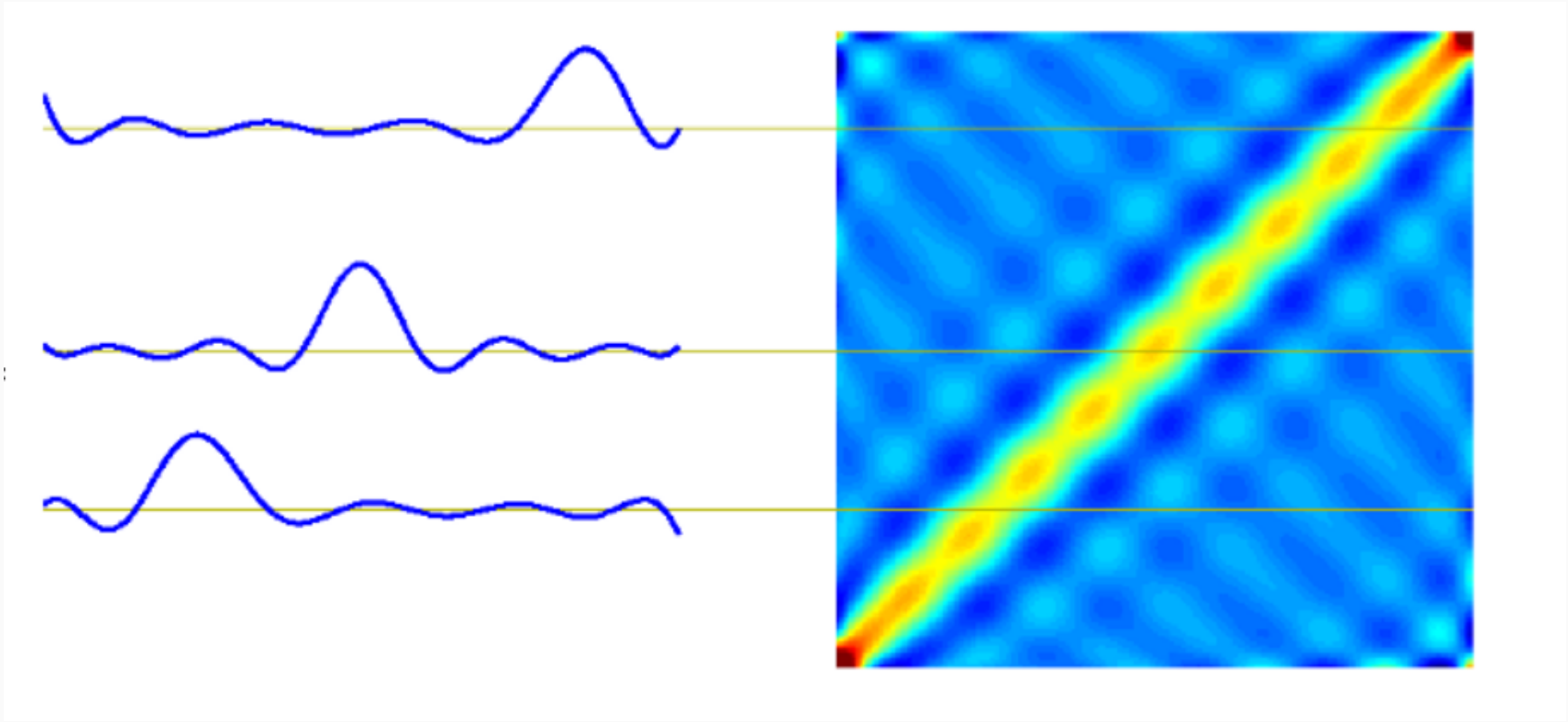
EQUIVALENT KERNEL

- $y(\mathbf{x}, \mu_N) = \sum_{n=1}^{N} \beta\phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}_n) t_n$.
- And the prediction can be rewritten as

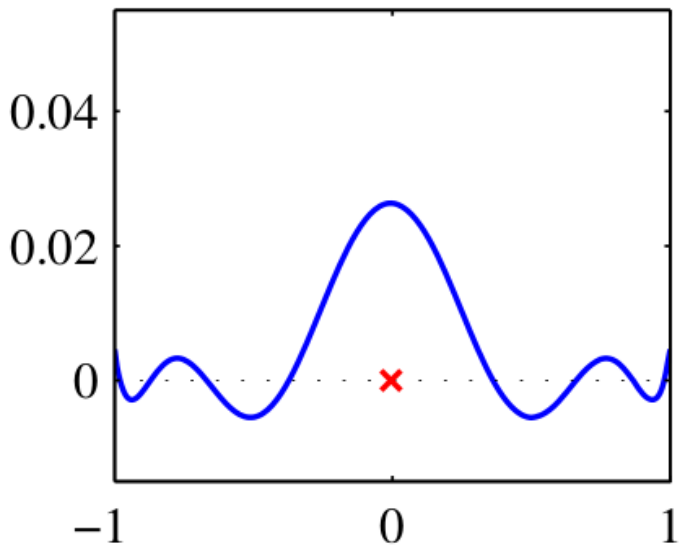
$$y(\mathbf{x}, \mu_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n.$$

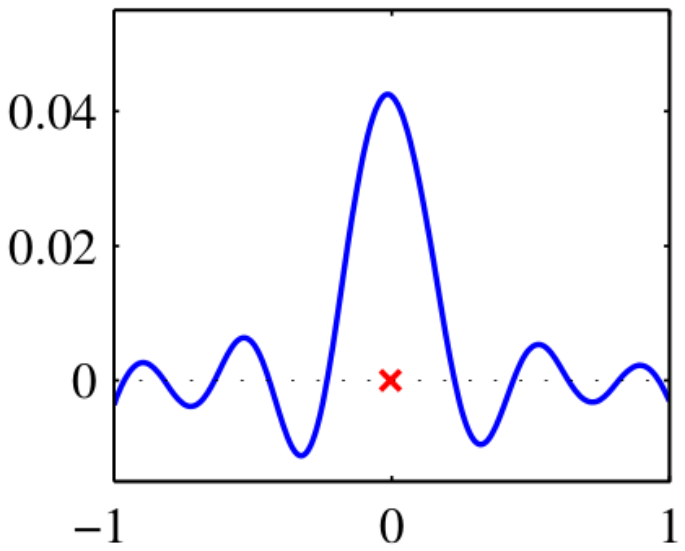
- I.e., we predict the next point as a linear combination of values in known points.
- Function $k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}')$ is called the *equivalent kernel*.

EQUIVALENT KERNEL



EQUIVALENT KERNEL





EQUIVALENT KERNEL

- Equivalent kernel $k(\mathbf{x}, \mathbf{x}')$ is localized around $\mathbf{x}$ as a function of $\mathbf{x}'$, i.e., every point has the largest influence nearby and then less and less (but it's not monotone!).
- We could simply define the kernel from the outset and predict with it without any ϕ functions — this is an important idea for the future.

Exercise. Prove that $\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1$.

BAYESIAN MODEL COMPARISON

BAYESIAN MODEL COMPARISON

- As the number of parameters increases, overfitting begins.
- How do we choose a model without overfitting? How can we compare models with different number of parameters?
- Bayesian approach: let's just compare $p(\mathcal{M} \mid D)$. :)

BAYESIAN MODEL COMPARISON

- Suppose we have a set of models $\{\mathcal{M}_i\}_{i=1}^{L}$.
- A model is a probability distribution over D.
- And we can estimate the posterior

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

BAYESIAN MODEL COMPARISON

- If we know the posterior, we can make a prediction:

$$p(t \mid \mathbf{x}, D) = \sum_{i=1}^{L} p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i \mid D).$$

- *Model selection* is when we approximate the prediction by choosing the most probable model (a posteriori).

BAYESIAN MODEL COMPARISON

- If the models are defined parametrically with $\mathbf{w}$, we have

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

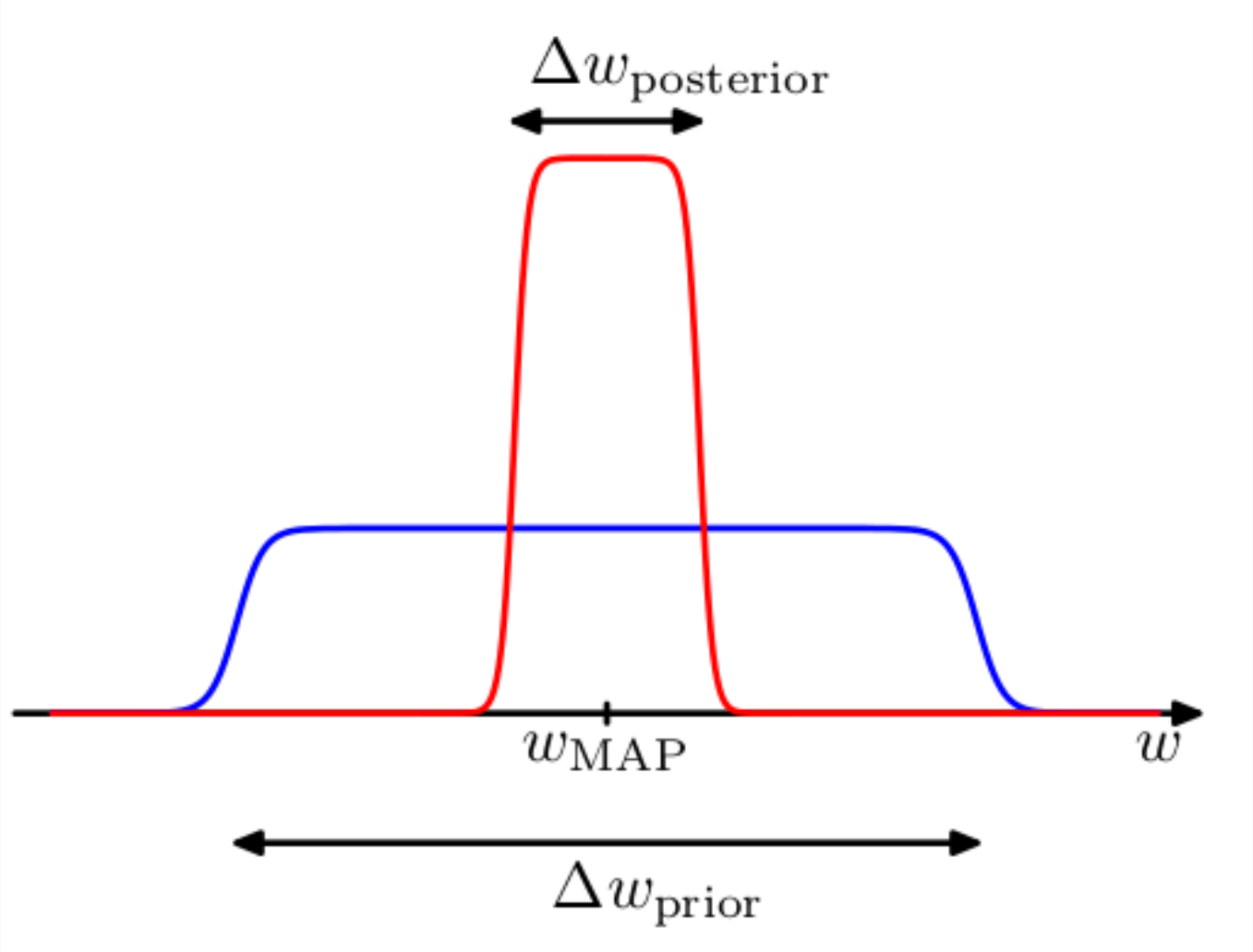
- This is the probability to generate D if we choose model parameters with its prior and then sample the data.
- Exactly the denominator from the Bayes' theorem:

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

BAYESIAN MODEL COMPARISON

- Suppose that the model has a single parameter w, and the posterior is a sharp peak around w_{MAP} of width $\Delta w_{\mathrm{posterior}}$.
- Then we can approximate $p(D) = \int p(D \mid w)p(w)dw$ as the value in the maximum times the width.
- Let's also assume that the prior distribution is flat, $p(w) = \frac{1}{\Delta w_{\mathrm{prior}}}$.

APPROXIMATING $p(d)$



APPROXIMATING $p(d)$

- Then we get

$$p(D) = \int p(D \mid w)p(w)dw \approx p\left(D \mid w_{\mathrm{MAP}}\right) \frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}},$$

$$\ln p(D) \approx \ln p\left(D \mid w_{\mathrm{MAP}}\right) + \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right).$$

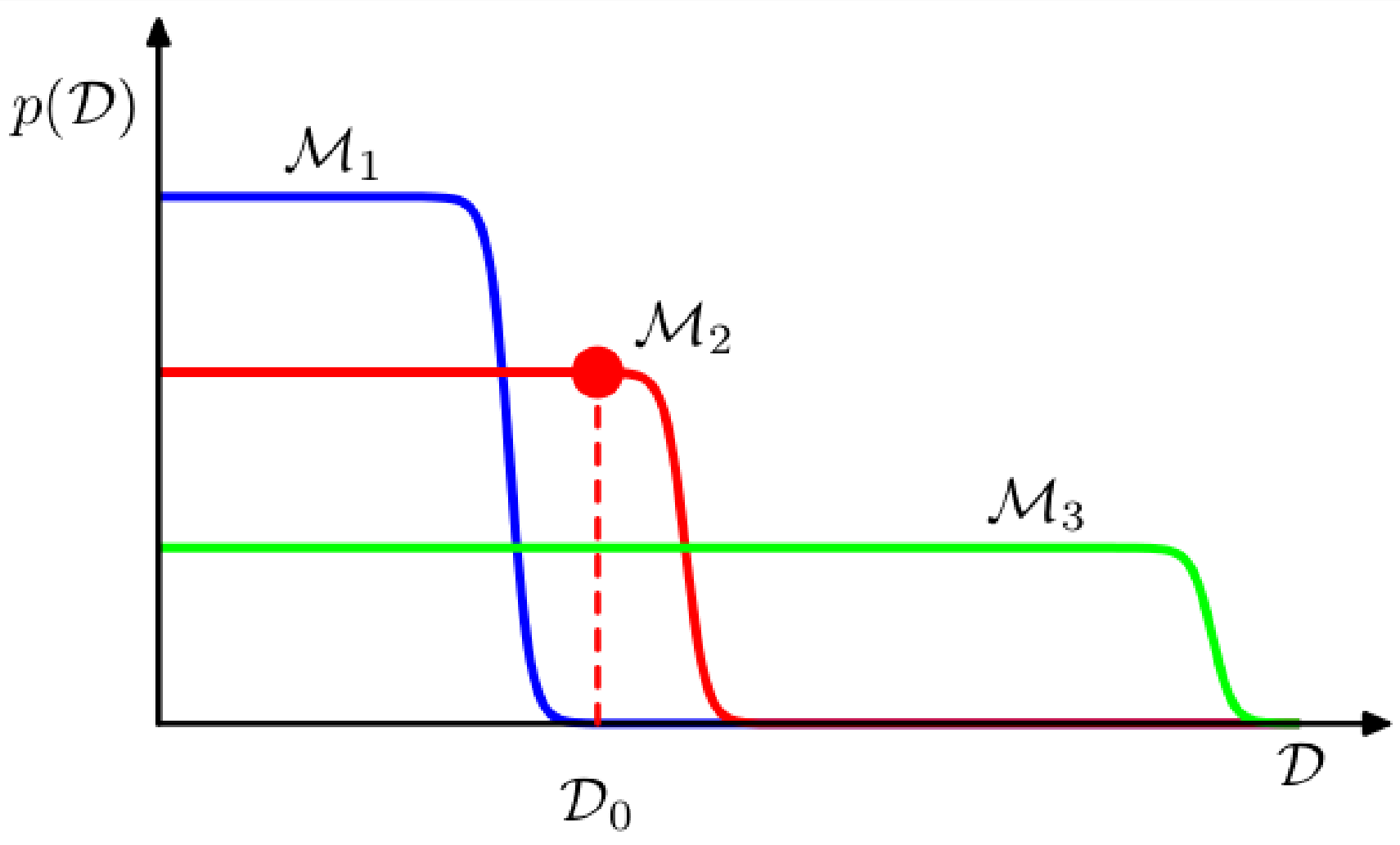
- This means that we add a penalty for "too narrow" posteriors... that is, precisely the penalty for overfitting!
- For a model of M parameters, if we assume that they have identical $\Delta w_{\mathrm{posterior}}$ we get

$$\ln p(D) \approx \ln p\left(D \mid w_{\mathrm{MAP}}\right) + M \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right).$$

ANOTHER VIEW

- In other words: let's see what kinds of datasets can be generated by a certain model.
- A simple model (e.g., linear) generates similar datasets, "few" different datasets, and so has high $p(D \mid \mathcal{M})$.
- A complicated model (e.g., degree 9 poly) generates "many" different datasets, and so has low $p(D \mid \mathcal{M})$.
- But a complicated model can express datasets that a simple one cannot; so in total we should choose a "middle ground"

APPROXIMATING $p(d)$



THE CORRECT ANSWER IS BETTER

- Sanity check: we have introduced strange-looking penalties; but will the true correct answer $p(D \mid \mathcal{M}_{\mathrm{true}})$ be actually optimal in this sense?
- For a specific dataset, not necessarily.
- But averageing over all datasets sampled from the true distribution $p(D \mid \mathcal{M}_{\mathrm{true}})$...

THE CORRECT ANSWER IS BETTER

- ...we get

$$\mathrm{E}\left[\ln \frac{p(D \mid \mathcal{M}_{\mathrm{true}})}{p(D \mid \mathcal{M})}\right] = \int p(D \mid \mathcal{M}_{\mathrm{true}}) \ln \frac{p(D \mid \mathcal{M}_{\mathrm{true}})}{p(D \mid \mathcal{M})} dD.$$

- This is called *Kullback-Leibler divergence* between distributions $p(D \mid \mathcal{M}_{\mathrm{true}})$ and $p(D \mid \mathcal{M})$.

Exercise. Prove that the Kullback-Leibler divergence is always nonnegative, i.e., $p(D \mid \mathcal{M}_{\mathrm{true}}) \geq p(D \mid \mathcal{M})$ for every $\mathcal{M}$.

MODEL COMPARISON WITH LAPLACE APPROXIMATION

- But we can do better than just a flat plateau! Let's try Laplace approximation for this case.
- Again, to compare models $\{\mathcal{M}_i\}_{i=1}^{L}$, we evaluate with test set D the posterior

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

- If the models are parametric then $p(D \mid \mathcal{M}_i) = \int p(D \mid \theta, \mathcal{M}_i)p(\theta \mid \mathcal{M}_i)d\theta$.
- This is the probability to generate D if we choose model parameters according to its prior, the Bayes theorem's denominator:

$$p(\theta \mid \mathcal{M}_i, D) = \frac{p(D \mid \theta, \mathcal{M}_i)p(\theta \mid \mathcal{M}_i)}{p(D \mid \mathcal{M}_i)}.$$

Model comparison with Laplace approximation

- Let's approximate with a Gaussian; integrating, we get

$$Z = \int f(\mathbf{z}) d\mathbf{z} \approx \int f(\mathbf{z}_0) e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A}(\mathbf{z}-\mathbf{z}_0)} d\mathbf{z} = f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}.$$

- And we have $Z = p(D)$, $f(\theta) = p(D \mid \theta)p(\theta)$.

MODEL COMPARISON WITH LAPLACE APPROXIMATION

- We get

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) + \ln P(\theta_{\mathrm{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|.$$

- $\ln P(\theta_{\mathrm{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|$ is called the *Occam's factor*.
- $\mathbf{A} = -\nabla\nabla \ln p(D \mid \theta_{\mathrm{MAP}}) p(\theta_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\theta_{\mathrm{MAP}} \mid D)$.

MODEL COMPARISON WITH LAPLACE APPROXIMATION

- We get

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) + \ln P(\theta_{\mathrm{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|.$$

- If the Gaussian prior $p(\theta)$ is sufficiently wide, and $\mathbf{A}$ has full rank, we can roughly approximate as (prove it!)

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) - \frac{1}{2} M \ln N,$$

  where M is the number of parameters, N is the number of points in D, and we have omitted additive constants.

- This is called the *Bayesian information criterion* (BIC), or *Schwarz criterion*.

THANK YOU!

Thank you for your attention!