STA 314: Statistical Methods for Machine Learning I

Lecture 11 - Bayesian Linear Regression, Probabilistic PCA

Chris J. Maddison

University of Toronto

Today

- Final exam does not include this nor next week's lecture.
- Continuing in our theme of probabilistic models for continuous variables.
  - Probabilistic interpretation of linear regression
  - Probabilistic interpretation of PCA
- (Optional) Bayesian model selection.

Completing the Square for Gaussians

- First, we're going to review a very powerful technique that will let us figure out the distribution of Gaussian random variables.
- It's a multivariate generalization of completing the square.
- The density of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ satifies:

$$\begin{aligned}\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \text{const} \\ &= -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}\end{aligned}$$

- Thus, if we know $\mathbf{w}$ is Gaussian with *unknown* mean and covariance, and we also know that

$$\log p(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^\top \mathbf{A}\mathbf{w} + \mathbf{w}^\top \mathbf{b} + \text{const}$$

for $\mathbf{A}$ positive definite, then we know that

$$\mathbf{w} \sim \mathcal{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1})$$

Bayesian Linear Regression

- We're going to be Bayesian about the parameters of the model.
  - This is in contrast with naïve Bayes and GDA: in those cases, we used Bayes' rule to infer the class, but used point estimates of the parameters.
  - By inferring a posterior distribution over the *parameters*, the model can know what it doesn't know.
- How can uncertainty in the predictions help us?
  - Smooth out the predictions by averaging over lots of plausible explanations (just like ensembles!)
  - Assign confidences to predictions
  - Make more robust decisions

Recap: Linear Regression

- Given a training set of inputs and targets $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}$
- Linear model:

$$y = \mathbf{w}^\top \psi(\mathbf{x})$$

- Vectorized, we have the design matrix $\mathbf{X}$ in input space and

$$\boldsymbol{\Psi} = \begin{bmatrix} - & \psi(\mathbf{x}^{(1)}) & - \\ - & \psi(\mathbf{x}^{(2)}) & - \\ & \vdots & \\ - & \psi(\mathbf{x}^{(N)}) & - \end{bmatrix}$$

and predictions

$$\mathbf{y} = \boldsymbol{\Psi}\mathbf{w}$$

Recap: Linear Regression

- Squared error loss:

$$L(\mathbf{y}, \mathbf{t}) = \frac{1}{2}\|\mathbf{y} - \mathbf{t}\|^2$$

- L_2 regularization:

$$\phi(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- Solution 1: solve analytically by setting the gradient to 0

$$\mathbf{w} = (\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \lambda \mathbf{I})^{-1} \boldsymbol{\Psi}^\top \mathbf{t}$$

- Solution 2: solve approximately using gradient descent

$$\mathbf{w} \leftarrow (1 - \alpha\lambda)\mathbf{w} - \alpha \boldsymbol{\Psi}^\top (\mathbf{y} - \mathbf{t})$$

Linear Regression as Maximum Likelihood

- We can give linear regression a probabilistic interpretation by assuming a Gaussian noise model:

$$t \mid \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}),\ \sigma^2)$$

- Linear regression is just maximum likelihood under this model:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \log p(t^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}, b) &= \frac{1}{N}\sum_{i=1}^{N} \log \mathcal{N}(t^{(i)}; \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}), \sigma^2) \\
&= \frac{1}{N}\sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}))^2}{2\sigma^2} \right) \right] \\
&= \text{const} - \frac{1}{2N\sigma^2}\sum_{i=1}^{N} (t^{(i)} - \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}))^2
\end{aligned}$$

Regularized Linear Regression as MAP Estimation

- We can view an L_2 regularizer as MAP inference with a Gaussian prior.
- Recall MAP inference:

$$\arg\max_{\mathbf{w}} \log p(\mathbf{w} \mid \mathcal{D}) = \arg\max_{\mathbf{w}} \left[\log p(\mathbf{w}) + \log p(\mathcal{D} \mid \mathbf{w})\right]$$

- We just derived the likelihood term $\log p(\mathcal{D} \mid \mathbf{w})$:

$$\log p(\mathcal{D} \mid \mathbf{w}) = \text{const} - \frac{1}{2N\sigma^2} \sum_{i=1}^{N} (t^{(i)} - \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}))^2$$

- Assume a Gaussian prior, $\mathbf{w} \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$:

$$\begin{aligned}
\log p(\mathbf{w}) &= \log \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{S}) \\
&= \log \left[\frac{1}{(2\pi)^{D/2} |\mathbf{S}|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{w}-\mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w}-\mathbf{m})\right) \right] \\
&= -\tfrac{1}{2}(\mathbf{w}-\mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w}-\mathbf{m}) + \text{const}
\end{aligned}$$

- Commonly, $\mathbf{m} = \mathbf{0}$ and $\mathbf{S} = \eta \mathbf{I}$, so

$$\log p(\mathbf{w}) = -\frac{1}{2\eta} \|\mathbf{w}\|^2 + \text{const.}$$

This is just L_2 regularization!

Recap: Full Bayesian Inference

- Recall: full Bayesian inference makes predictions by averaging over all likely explanations under the posterior distribution.
- Compute posterior using Bayes' Rule:

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

- Make predictions using the posterior predictive distribution:

$$p(t \mid \mathbf{x}, \mathcal{D}) = \int p(\mathbf{w} \mid \mathcal{D})\, p(t \mid \mathbf{x}, \mathbf{w})\, \mathrm{d}\mathbf{w}$$

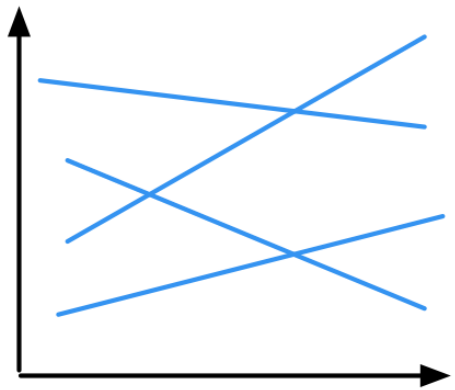
- Doing this lets us quantify our uncertainty.

Bayesian Linear Regression

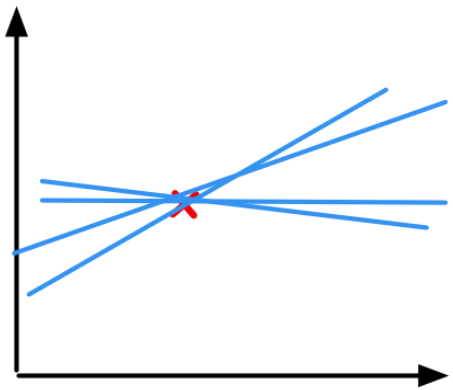
- **Prior distribution:** $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{S})$
- **Likelihood:** $t \mid \mathbf{x}, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}),\ \sigma^2)$
- Assuming fixed/known $\mathbf{S}$ and σ^2 is a big assumption. More on this later.

Bayesian Linear Regression

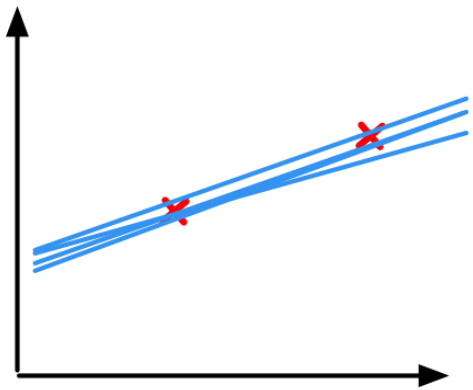
- Bayesian linear regression considers various plausible explanations for how the data were generated.
- It makes predictions using all possible regression weights, weighted by their posterior probability.
- Here are samples from the prior $p(\mathbf{w})$ and posteriors $p(\mathbf{w} \mid \mathcal{D})$



no observations



one observation



two observations

Bayesian Linear Regression: Posterior

- **Deriving the posterior distribution:**

$$\begin{aligned}
\log p(\mathbf{w} \mid \mathcal{D}) &= \log p(\mathbf{w}) + \log p(\mathcal{D} \mid \mathbf{w}) + \text{const} \\
&= -\tfrac{1}{2}\mathbf{w}^\top \mathbf{S}^{-1}\mathbf{w} - \frac{1}{2\sigma^2}\|\boldsymbol{\Psi}\mathbf{w} - \mathbf{t}\|^2 + \text{const} \\
&= -\tfrac{1}{2}\mathbf{w}^\top \mathbf{S}^{-1}\mathbf{w} - \frac{1}{2\sigma^2}\left(\mathbf{w}^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi}\mathbf{w} - 2\mathbf{t}^\top \boldsymbol{\Psi}\mathbf{w} + \mathbf{t}^\top \mathbf{t}\right) + \text{const} \\
&= -\tfrac{1}{2}\mathbf{w}^\top \left(\sigma^{-2}\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \mathbf{S}^{-1}\right)\mathbf{w} - \frac{1}{\sigma^2}\mathbf{t}^\top \boldsymbol{\Psi}\mathbf{w} + \text{const} \quad \text{(complete the square!)}
\end{aligned}$$

Thus $\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Psi}^\top \mathbf{t}$$

$$\boldsymbol{\Sigma} = \left(\sigma^{-2}\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \mathbf{S}^{-1}\right)^{-1}$$

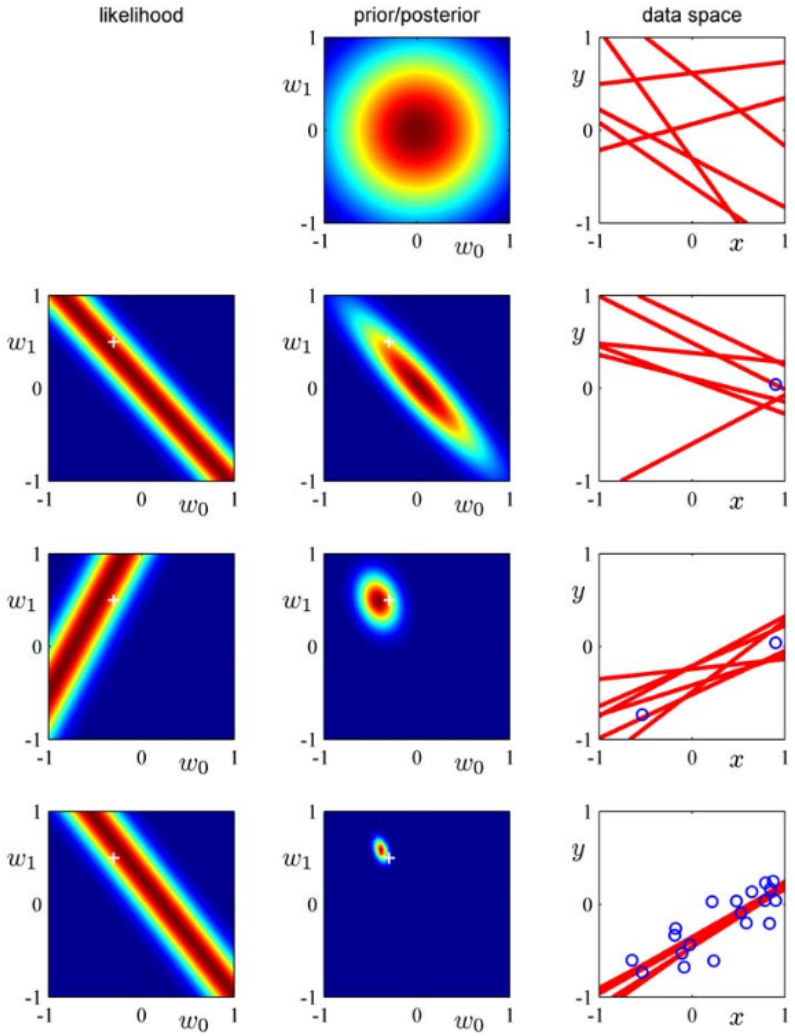
Bayesian Linear Regression: Posterior

- Since a Gaussian prior leads to a Gaussian posterior, this means the Gaussian distribution is the conjugate prior for linear regression!
- Compare $\boldsymbol{\mu}$ the closed-form solution for linear regression:

$$\mathbf{w} = (\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \lambda \mathbf{I})^{-1} \boldsymbol{\Psi}^\top \mathbf{t}$$

  This is the mean of the posterior, assuming that $\mathbf{S} = \lambda^{-1}\mathbf{I}$ and $\sigma = 1$.
- λ^{-1} is the standard deviation of the prior. As this becomes infinite, the mean of the posterior converges to the maximum likelihood solution.

Bayesian Linear Regression

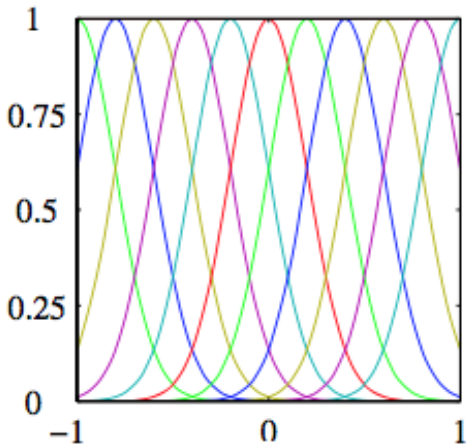


— Bishop, Pattern Recognition and Machine Learning

Bayesian Linear Regression

- Example with radial basis function (RBF) features

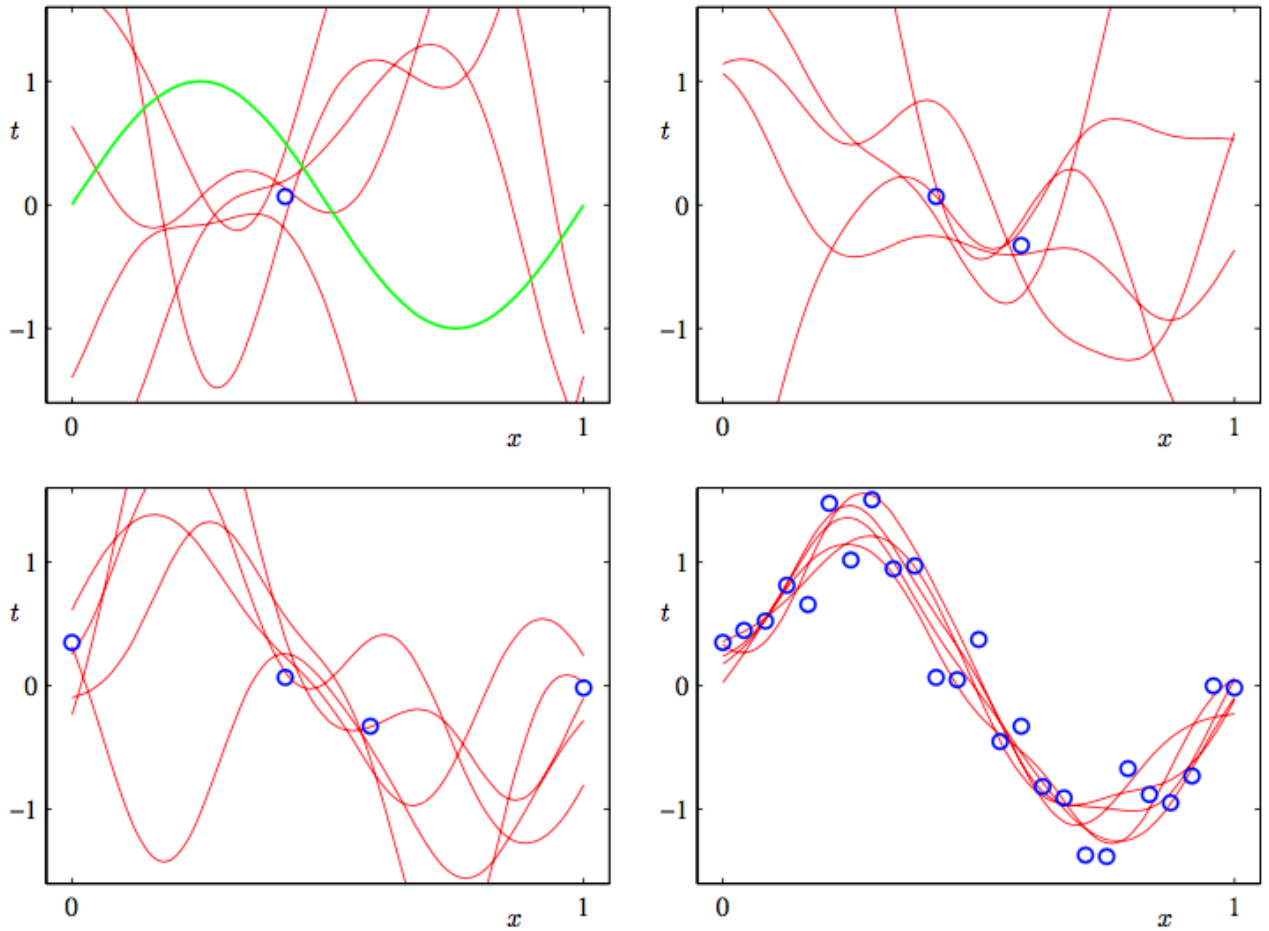
$$\psi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$$



— Bishop, Pattern Recognition and Machine Learning

Bayesian Linear Regression

Functions sampled from the posterior:



— Bishop, Pattern Recognition and Machine Learning

Bayesian Linear Regression

- The posterior just gives us distribution over the parameter space, but if we want to make predictions, the natural choice is to use the posterior predictive distribution.
- Posterior predictive distribution:

$$p(t \mid \mathbf{x}, \mathcal{D}) = \int \underbrace{p(t \mid \mathbf{x}, \mathbf{w})}_{\mathcal{N}(t\,;\,\mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}),\sigma)} \underbrace{p(\mathbf{w} \mid \mathcal{D})}_{\mathcal{N}(\mathbf{w}\,;\,\boldsymbol{\mu},\boldsymbol{\Sigma})} \, \mathrm{d}\mathbf{w}$$

- Another interpretation: $t = \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma)$ is independent of $\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Bayesian Linear Regression

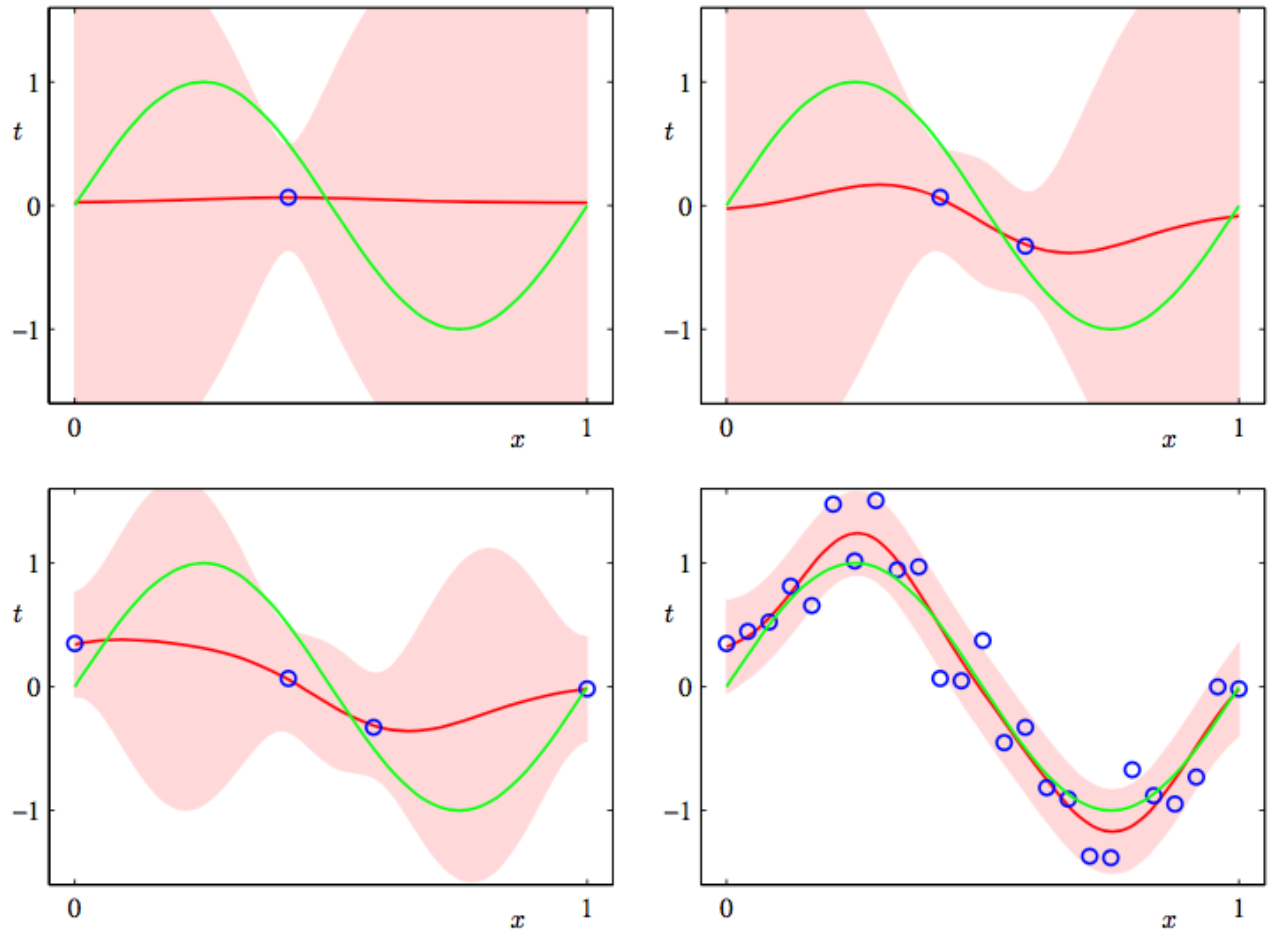
- Another interpretation: $t = \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma)$ is independent of $\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- By the linear combination rules for Gaussian random variables, t is a Gaussian distribution with parameters

$$\mu_{\text{pred}} = \boldsymbol{\mu}^\top \boldsymbol{\psi}(\mathbf{x})$$
$$\sigma^2_{\text{pred}} = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\Sigma} \boldsymbol{\psi}(\mathbf{x}) + \sigma^2$$

- Hence, the posterior predictive distribution is $\mathcal{N}(t\,;\, \mu_{\text{pred}}, \sigma^2_{\text{pred}})$.

Bayesian Linear Regression

Here we visualize confidence intervals based on the posterior predictive mean and variance at each point:



Overview: Probabilistic PCA

- The formulation of PCA that we saw earlier in the course was motivated heuristically.
- We will show that it can be expressed as the maximum likelihood estimate of a certain probabilistic model.

Recall: PCA

- Data set $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$
- Each input vector $\mathbf{x}^{(i)} \in \mathbb{R}^D$ is approximated as $\hat{\boldsymbol{\mu}} + \mathbf{U}\mathbf{z}^{(i)}$,

$$\mathbf{x}^{(i)} \approx \tilde{\mathbf{x}}^{(i)} = \hat{\boldsymbol{\mu}} + \mathbf{U}\mathbf{z}^{(i)}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_i \mathbf{x}^{(i)}$ is the data mean, $\mathbf{U} \in \mathbb{R}^{D \times K}$ is the orthogonal basis for the principal subspace, and $\mathbf{z}^{(i)} \in \mathbb{R}^K$ is the code vector

$$\mathbf{z}^{(i)} = \mathbf{U}^\top(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})$$

- $\mathbf{U}$ is chosen to minimize the reconstruction error

$$\mathbf{U}^* = \arg\min_{\mathbf{U}} \sum_i \|\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}} + \mathbf{U}\mathbf{U}^\top(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})\|^2$$

Probabilistic PCA

- To formulate probabilistic PCA, let's start with a latent variable model.
- Similar to the Gaussian mixture model, but we will assume continuous, Gaussian latents:

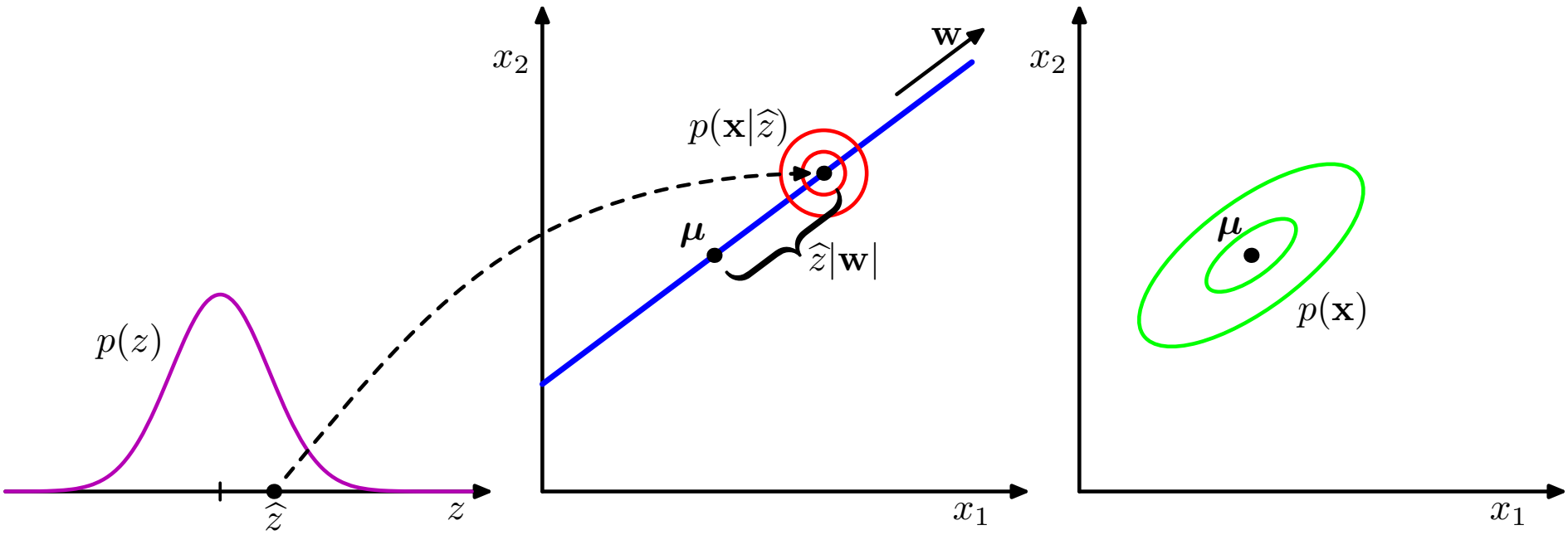
$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I})$$

- Note: this is a naive Bayes model, because $p(\mathbf{x} \mid \mathbf{z})$ factorizes with respect to the dimensions of $\mathbf{x}$.
- What sort of data does this model produce?

Probabilistic PCA

- $\mathbf{z}$ is a random coordinate within the affine space centered at $\boldsymbol{\mu}$ and spanned by the columbs of $\mathbf{W}$.
- To get the random variable $\mathbf{x}$, we samples a standard Normal $\mathbf{z}$ and then add a small amount of isotropic noise to $\mathbf{Wz} + \boldsymbol{\mu}$.

Probabilistic PCA



— Bishop, Pattern Recognition and Machine Learning

Probabilistic PCA : Maximum Likelihood

- To perform maximum likelihood in this model, we need to maximize the following:

$$\max_{\mathbf{W},\boldsymbol{\mu},\sigma^2} \log p(\mathbf{x} \mid \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \max_{\mathbf{W},\boldsymbol{\mu},\sigma^2} \log \int p(\mathbf{x} \mid \mathbf{z}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{z}) \, d\mathbf{z}$$

- This was hard for the Gaussian mixture model, but in this case it's easy.
- $p(\mathbf{x} \mid \mathbf{W}, \boldsymbol{\mu}, \sigma^2)$ will be Gaussian (confirm this) and so we just need to compute and Cov$[\mathbf{x}]$ and $\mathbb{E}[\mathbf{x}]$.

Probabilistic PCA : Maximum Likelihood

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \epsilon] = \boldsymbol{\mu}$$

$$\begin{aligned}
\text{Cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{W}\mathbf{z} + \epsilon)(\mathbf{W}\mathbf{z} + \epsilon)^\top] \\
&= \mathbb{E}[(\mathbf{W}\mathbf{z}\mathbf{z}^\top\mathbf{W}^\top] + \text{Cov}[\epsilon\epsilon^\top] \\
&= \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}
\end{aligned}$$

Probabilistic PCA : Maximum Likelihood

- Thus, the likelihood of the data under this model is given by

$$-\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu})$$

  where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.
- It's a bit involved to derive the maximum likelihood solution, so we will skip it, but Tipping and Bishop (Probabilistic PCA, 1999) show that this is maximized at the following stationary points.

Probabilistic PCA : Maximum Likelihood

$$\hat{\boldsymbol{\mu}}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}$$

Probabilistic PCA : Maximum Likelihood

$$\hat{\mathbf{W}}_{\text{MLE}} = \hat{\mathbf{U}}_{\text{MLE}}(\hat{\mathbf{L}}_{\text{MLE}} - \hat{\sigma}^2_{\text{MLE}}\mathbf{I})^{\frac{1}{2}}\mathbf{R}$$

where $\hat{\mathbf{U}}_{\text{MLE}}$ is the matrix whose columns are the K unit eigenvectors of the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ that have the largest eigenvalues, $\hat{\mathbf{L}}_{\text{MLE}} \in \mathbb{R}^{K\times K}$ is the diagonal matrix whose elements are the corresponding eigenvalues, and $\mathbf{R}$ is any orthogonal matrix.

Probabilistic PCA : Maximum Likelihood

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{D-K} \sum_{i=K+1}^{D} \lambda_i$$

where λ_i is the ith largest eigenvalue of the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ of the data. In otherwords, the average variance of the discarded subspace.

Probabilistic PCA : Maximum Likelihood

- That seems complex, to get an intuition about how this model behaves when it is fit to data, lets consider the MLE density.
- Recall that the marginal distribution on $\mathbf{x}$ in our fitted model is a Gaussian with mean

$$\hat{\boldsymbol{\mu}}_{\text{MLE}}$$

and covariance

$$\hat{\mathbf{W}}_{\text{MLE}}\hat{\mathbf{W}}_{\text{MLE}}^{\top} + \hat{\sigma}^2_{\text{MLE}}\mathbf{I} = \hat{\mathbf{U}}_{\text{MLE}}(\hat{\mathbf{L}}_{\text{MLE}} - \hat{\sigma}^2_{\text{MLE}}\mathbf{I})\hat{\mathbf{U}}_{\text{MLE}}^{\top} + \hat{\sigma}^2_{\text{MLE}}\mathbf{I}$$

- The covariance gives us a nice intuition about the type of model this forms.

Probabilistic PCA : Maximum Likelihood

- Consider centering the data and checking the variance along one of the unit eigenvectors $\mathbf{u}_i$, which are the eigenvectors forming the columns of $\hat{\mathbf{U}}_{\text{MLE}}$:

$$\begin{aligned}\text{Var}(\mathbf{u}_i^\top(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\text{MLE}})) &= \mathbf{u}_i^\top \text{Cov}[\mathbf{x}]\mathbf{u}_i \\ &= \mathbf{u}_i^\top \hat{\mathbf{U}}_{\text{MLE}}(\hat{\mathbf{L}}_{\text{MLE}} - \hat{\sigma}^2_{\text{MLE}}\mathbf{I})\hat{\mathbf{U}}_{\text{MLE}}^\top \mathbf{u}_i + \hat{\sigma}^2_{\text{MLE}} \\ &= \lambda_i - \hat{\sigma}^2_{\text{MLE}} + \hat{\sigma}^2_{\text{MLE}} = \lambda_i\end{aligned}$$

- Now, consider centering the data and checking the variance along any unit vector orthogonal to the subspace spanned by $\hat{\mathbf{U}}_{\text{MLE}}$:

$$\begin{aligned}\text{Var}(\mathbf{u}_i^\top(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\text{MLE}})) &= \mathbf{u}_i^\top \hat{\mathbf{U}}_{\text{MLE}}(\hat{\mathbf{L}}_{\text{MLE}} - \hat{\sigma}^2_{\text{MLE}}\mathbf{I})\hat{\mathbf{U}}_{\text{MLE}}^\top \mathbf{u}_i + \hat{\sigma}^2_{\text{MLE}} \\ &= \hat{\sigma}^2_{\text{MLE}}\end{aligned}$$

- In other words, the model captures the variance along the principle axes and approximates the variance in all remaining directions with a single variance.

Probably easier to visualize after implementing it.

How does it relate to PCA?

- The posterior mean is given by

$$\mathbb{E}[\mathbf{z} \mid \mathbf{x}] = \left(\hat{\mathbf{W}}_{\text{MLE}}^\top \hat{\mathbf{W}}_{\text{MLE}} + \hat{\sigma}^2_{\text{MLE}} \mathbf{I}\right)^{-1} \hat{\mathbf{W}}_{\text{MLE}}^\top (\mathbf{x} - \hat{\boldsymbol{\mu}}_{\text{MLE}})$$

- So, if we don't fit σ^2 and instead take it to 0 we get

$$\mathbb{E}[\mathbf{z} \mid \mathbf{x}] \overset{\sigma^2 \to 0}{\to} \left(\hat{\mathbf{W}}_{\text{MLE}}^\top \hat{\mathbf{W}}_{\text{MLE}}\right)^{-1} \hat{\mathbf{W}}_{\text{MLE}}^\top (\mathbf{x} - \hat{\boldsymbol{\mu}}_{\text{MLE}})$$

- Can show that this is a projection onto an affine space spanned by the columns of $\hat{\mathbf{U}}_{\text{MLE}}$.

Why Probabilistic PCA (PPCA)?

- Fitting a full-covariance Gaussian model of data requires $D(D+1)/2 + D$ parameters. With PPCA we model only the K most significant correlations and this only requires $\mathcal{O}(D)$ parameters as long as K is small.
- Basis of Bayesian treatement of PCA, which gives us a Bayesian method for determining the dimensionality of the principal subspace (i.e. K).
- Existence of likelihood functions allows direct comparison with other probabilistic models.
- Can use PPCA as a class-conditional density (as in GDA) to reduce the requirement to fit and store $\mathcal{O}(D^2)$ parameters.

Recap: Gaussian models that we covered

- Gaussian discriminant analysis.
  - Gaussian class-conditional generative model $p(\mathbf{x} \mid t)$ used for classification.
- Gaussian mixture model.
  - Gaussian latent variable model $p(\mathbf{x}) = \sum_z p(\mathbf{x}, z)$ used for clustering.
- Bayesian linear regression.
  - Gaussian discriminative model $p(t \mid \mathbf{x})$ used for regression with a Bayesian analysis for the weights.
- Probabilistic PCA.
  - Gaussian latent variable model $p(\mathbf{x}) = \int_z p(\mathbf{x}, z)$ used for dimensionality reduction.

Optional material: Bayesian model selection

Occam's Razor (optional)

- Consider selecting models from a Bayesian perspective.
- Related to Occam's Razor: "Entities should not be multiplied beyond necessity."
  - Named after the 14th century British theologian William of Occam
- Huge number of attempts to formalize mathematically
  - See Domingos, 1999, "The role of Occam's Razor in knowledge discovery" for a skeptical overview.

    `https://homes.cs.washington.edu/~pedrod/papers/dmkd99.pdf`
- Common misinterpretation: your prior should favor simple explanations
- Better interpretation: by averaging over many hypothesis, Bayesian model selection naturally prefers simpler models.

Occam's Razor (optional)

- Suppose you have a finite set of models, or hypotheses $\{\mathcal{H}_i\}_{i=1}^M$ (e.g. polynomials of different degrees)
- Posterior inference over models (Bayes' Rule):

$$p(\mathcal{H}_i \mid \mathcal{D}) \propto \underbrace{p(\mathcal{H}_i)}_{\text{prior}} \underbrace{p(\mathcal{D} \mid \mathcal{H}_i)}_{\text{evidence}}$$

- The evidence is also called marginal likelihood since it requires marginalizing out the parameters:

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

Occam's Razor (optional)

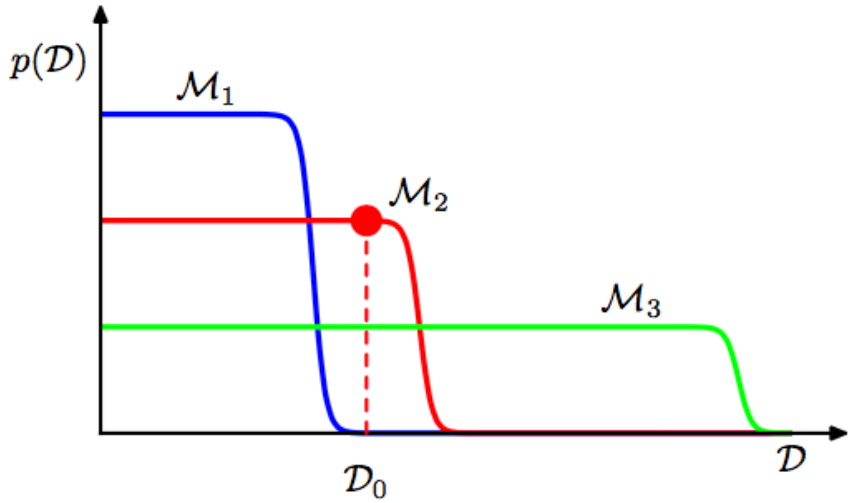
- $p(\mathcal{H}_i)$ is typically uniform, so we can compare them based on marginal likelihood.
- Bayesian model selection:

$$\mathcal{H}^* = \arg\max_i p(\mathcal{D} \mid \mathcal{H}_i)$$

- What types of models does this procedure prefer?

Occam's Razor (optional)

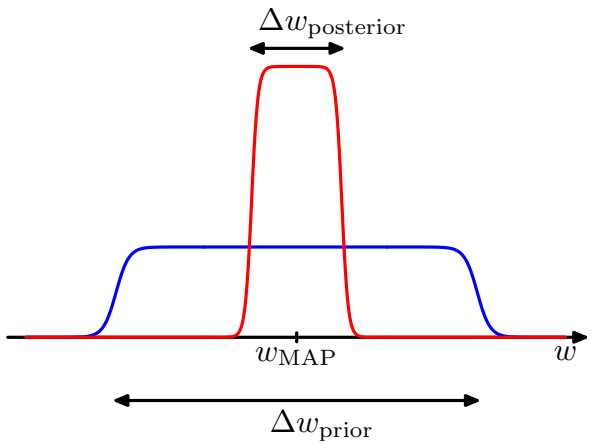
- Suppose M_1, M_2, and M_3 denote a linear, quadratic, and cubic model.
- M_3 is capable of explaning more datasets than M_1.
- But its distribution over $\mathcal{D}$ must integrate to 1, so it must assign lower probability to ones it can explain.



— Bishop, Pattern Recognition and Machine Learning

Occam's Razor (optional)

- How does the evidence penalize complex models?



- Approximating the integral for $\mathbf{w} \in \mathbb{R}$:

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

$$\simeq \underbrace{p(\mathcal{D} \mid \mathbf{w}_{\text{MAP}}, \mathcal{H}_i)}_{\text{best-fit likelihood}} \underbrace{\frac{\Delta \mathbf{w}_{\text{posterior}}}{\Delta \mathbf{w}_{\text{prior}}}}_{\text{Occam factor}}$$

Occam's Razor (optional)

- Let's investigate

$$\log p(\mathcal{D} \mid \mathcal{H}_i) = \log p(\mathcal{D} \mid \mathbf{w}_{\mathrm{MAP}}, \mathcal{H}_i) + \log \frac{\Delta \mathbf{w}_{\text{posterior}}}{\Delta \mathbf{w}_{\text{prior}}}$$

- First term represents fit to the data given the most probable parameter values.
- Second term is a penalty, because it is negative ($\Delta \mathbf{w}_{\text{posterior}} < \Delta \mathbf{w}_{\text{prior}}$).
- Thus if the posterior is very peaked and confident about the data, this penalty term will be very negative.

Occam's Razor (optional)

- $\mathbf{w} \in \mathbb{R}^M$ we have

$$\log p(\mathcal{D} \mid \mathcal{H}_i) = \log p(\mathcal{D} \mid \mathbf{w}_{\mathrm{MAP}}, \mathcal{H}_i) + M \log \frac{\Delta \mathbf{w}_{\text{posterior}}}{\Delta \mathbf{w}_{\text{prior}}}$$

- So the more parameters we have, the higher the penalty.
- Optimal model complexity is determined by a tradeoff.
- In Bayesian model selection, we naturally prefer simpler models that model the data well.