CSC 411 Lecture 19: Bayesian Linear Regression

Mengye Ren and Matthew MacKay

University of Toronto

Overview

- We've covered both parametric and nonparametric models for regression and classification.
  - Parametric models summarize the data into a model with a finite number of parameters. E.g., linear regression, logistic regression, neural nets, (linear) SVM, Naïve Bayes, GDA
  - Nonparametric models refer back to the data to make predictions. E.g., KNN
- The next two lectures are about Bayesian approaches to regression.
  - This lecture: Bayesian linear regression, a parametric model
  - Next lecture: Gaussian processes, a nonparametric model

Overview

- We're going to be Bayesian about the parameters of the model, i.e. model them as random variables
  - Do not confuse a Bayesian approach with using Bayes rule! i.e. naïve Bayes and GDA 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
  - Guide exploration (focus on areas you're uncertain about)
    - E.g., Bayesian optimization (see next tutorial)

Recap: Linear Regression

- Given a training set $\mathcal{D}$ of inputs and targets $\{(\mathbf{x}^{(n)}, t^{(n)})\}_{n=1}^{N}$, use linear model with fixed feature mapping ϕ

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

- Squared error cost (aka least-squares objective):

$$\frac{1}{2}\left[\sum_{n=1}^{N}(\mathbf{w}^\top \phi(\mathbf{x}^{(n)}) - t^{(n)})^2\right]$$

- L_2 regularization:

$$\frac{\lambda}{2}\|\mathbf{w}\|^2$$

- Let Φ be data matrix:

$$\Phi = \begin{bmatrix} — & \phi(\mathbf{x}^{(1)})^\top & — \\ & \vdots & \\ — & \phi(\mathbf{x}^{(N)})^\top & — \end{bmatrix}$$

Recap: Linear Regression

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

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

- Solution 2: solve approximately using gradient descent

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

Recap: Linear Regression

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

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

- Minimizing least squares objective is equivalent to maximizing likelihood under this model:

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

Recap: Linear Regression

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

$$\arg\max_{\mathbf{w}} \log p(\mathbf{w} \mid \mathcal{D}) = \arg\max_{\mathbf{w}} \left[\log p(\mathbf{w}) + \log p(\mathcal{D} \mid \mathbf{w}) - \underbrace{\log p(\mathcal{D})}_{\text{constant w.r.t. } \mathbf{w}} \right]$$

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

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

- Assume a Gaussian prior, $\mathbf{w} \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$. Commonly, $\mathbf{m} = \mathbf{0}$ and $\mathbf{S} = \eta \mathbf{I}$, so:

$$\begin{aligned} \log p(\mathbf{w}) &= \log \mathcal{N}(\mathbf{w}; \mathbf{0}, \eta \mathbf{I}) \\ &= \log \left(\frac{1}{(2\pi)^{D/2} \eta^{D/2}} \exp \left(-\frac{1}{2\eta} \mathbf{w}^T \mathbf{w} \right) \right) \\ &= -\frac{1}{2\eta} \|\mathbf{w}\|^2 + \text{const.} \end{aligned}$$

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:

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

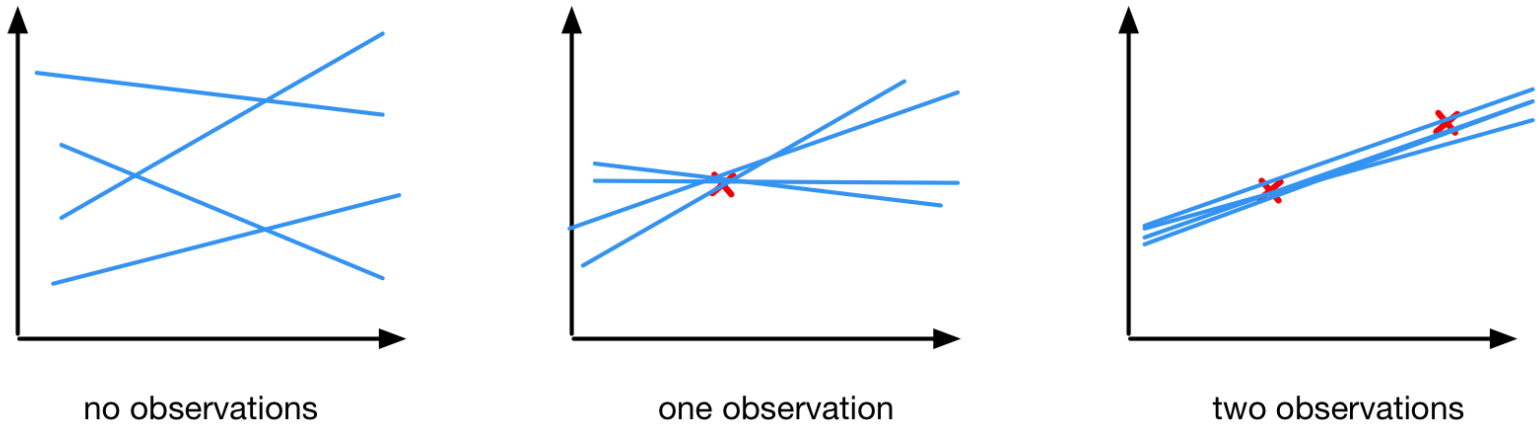
- Make predictions using the posterior predictive distribution:

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

- Doing this lets us quantify our uncertainty.

Bayesian Linear Regression

- **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.
  - We can visualize how $p(\mathbf{w}|\mathcal{D})$ changes with more data by sampling $\mathbf{w} \sim p(\mathbf{w}|\mathcal{D})$ and plotting $y = \mathbf{w}^T\mathbf{x}$:



- **Prior distribution:** $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \eta I)$
- **Likelihood:** $t \,|\, \mathbf{x}, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\top \phi(\mathbf{x}),\ \sigma^2)$
- η and σ^2 are hyperparameters

Bayesian Linear Regression: Posterior

- **Deriving the posterior distribution:**

$$
\begin{aligned}
\log p(\mathbf{w} \mid \mathcal{D}) &= \log p(\mathbf{w}) + \log p(\mathcal{D} \mid \mathbf{w}) + \text{const} \\
&= -\tfrac{1}{2\eta}\mathbf{w}^\top \mathbf{w} - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbf{w}^T \phi(\mathbf{x}^{(n)}) - t^{(n)})^2 + \text{const} \\
&= -\tfrac{1}{2\eta}\mathbf{w}^\top \mathbf{w} - \frac{1}{2\sigma^2}\|\Phi\mathbf{w} - \mathbf{t}\|^2 + \text{const} \\
&= -\tfrac{1}{2\eta}\mathbf{w}^\top \mathbf{w} - \frac{1}{2\sigma^2}(\Phi\mathbf{w} - \mathbf{t})^\top(\Phi\mathbf{w} - \mathbf{t}) + \text{const} \\
&= -\tfrac{1}{2\eta}\mathbf{w}^\top \mathbf{w} - \frac{1}{2\sigma^2}\left(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - 2\mathbf{t}^\top\Phi\mathbf{w} + \mathbf{t}^\top\mathbf{t}\right) + \text{const} \\
&= -\tfrac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}) + \text{const} \qquad \text{(complete the square!)}
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{\mu} &= \sigma^{-2}\boldsymbol{\Sigma}\Phi^\top\mathbf{t} \\
\boldsymbol{\Sigma}^{-1} &= \sigma^{-2}\Phi^\top\Phi + \eta^{-1}I
\end{aligned}
$$

Bayesian Linear Regression: Posterior

$$\log p(\mathbf{w} \mid \mathcal{D}) = -\tfrac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}) + \text{const}$$

where

$$\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top \mathbf{t}, \qquad \boldsymbol{\Sigma}^{-1} = \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \eta^{-1}\boldsymbol{I}$$

- Hence:

$$\begin{aligned} p(\mathbf{w} \mid \mathcal{D}) &= \exp(-\tfrac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}))\exp(\text{const}) \\ &\propto \exp(-\tfrac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})) \end{aligned}$$

- This is a multivariate Gaussian distribution, i.e.

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

Bayesian Linear Regression: Posterior

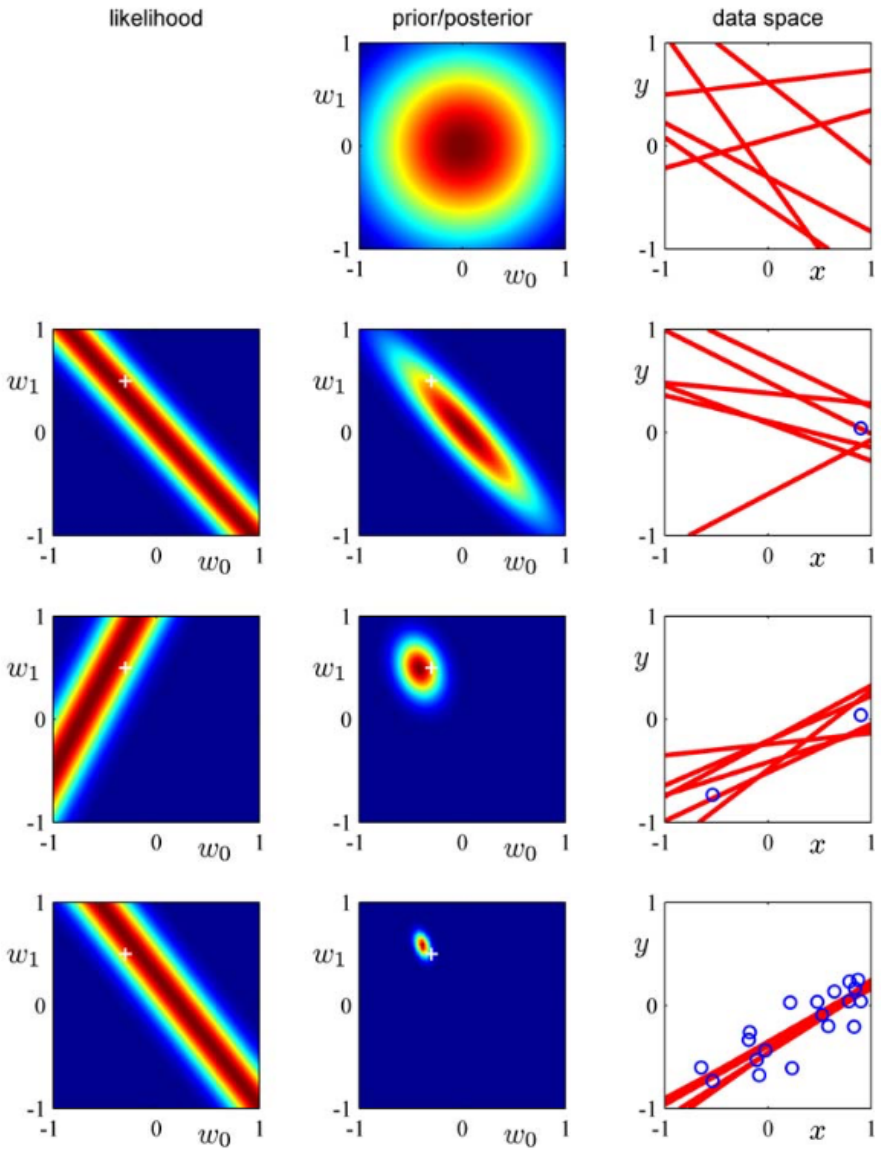
- Just showed:

$$\mathbf{w} \,|\, \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\Sigma} \boldsymbol{\Phi}^\top \mathbf{t}$$
$$\boldsymbol{\Sigma}^{-1} = \sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \eta^{-1} I$$

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

$$\boldsymbol{\mu} = \sigma^{-2} (\sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \eta^{-1} I)^{-1} \boldsymbol{\Phi}^\top \mathbf{t}$$
$$\mathbf{w} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \mathbf{I})^{-1} \boldsymbol{\Phi}^\top \mathbf{t}$$

Bayesian Linear Regression



— Bishop, Pattern Recognition and Machine Learning

Bayesian Linear Regression: Posterior

- Aside: why are likelihood contours lines?

$$c = \mathcal{N}(t|w_0 x_0 + w_1 x_1, \sigma^2)$$
$$\implies c = d \exp\left(-\frac{1}{2\sigma^2}(t - w_0 x_0 - w_1 x_1)^2\right)$$
$$\implies \sqrt{-2\sigma^2 \log(c/d)} = t - w_0 x_0 - w_1 x_1$$

- Set $e = \sqrt{-2\sigma^2 \log(c/d)}$:

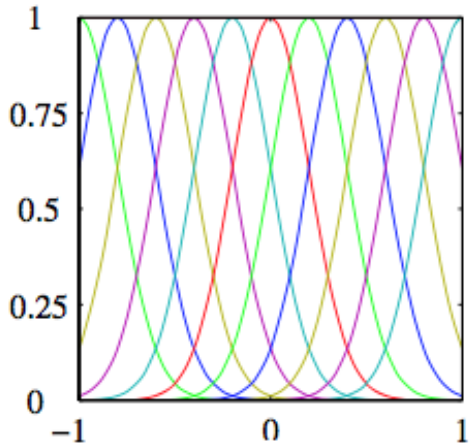
$$e = t - w_0 x_0 - w_1 x_1$$
$$\implies w_1 = \frac{1}{x_1}(t - w_0 x_0 - e)$$

- We find w_1 is a linear function of w_0

Bayesian Linear Regression

- Example with radial basis function (RBF) features

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

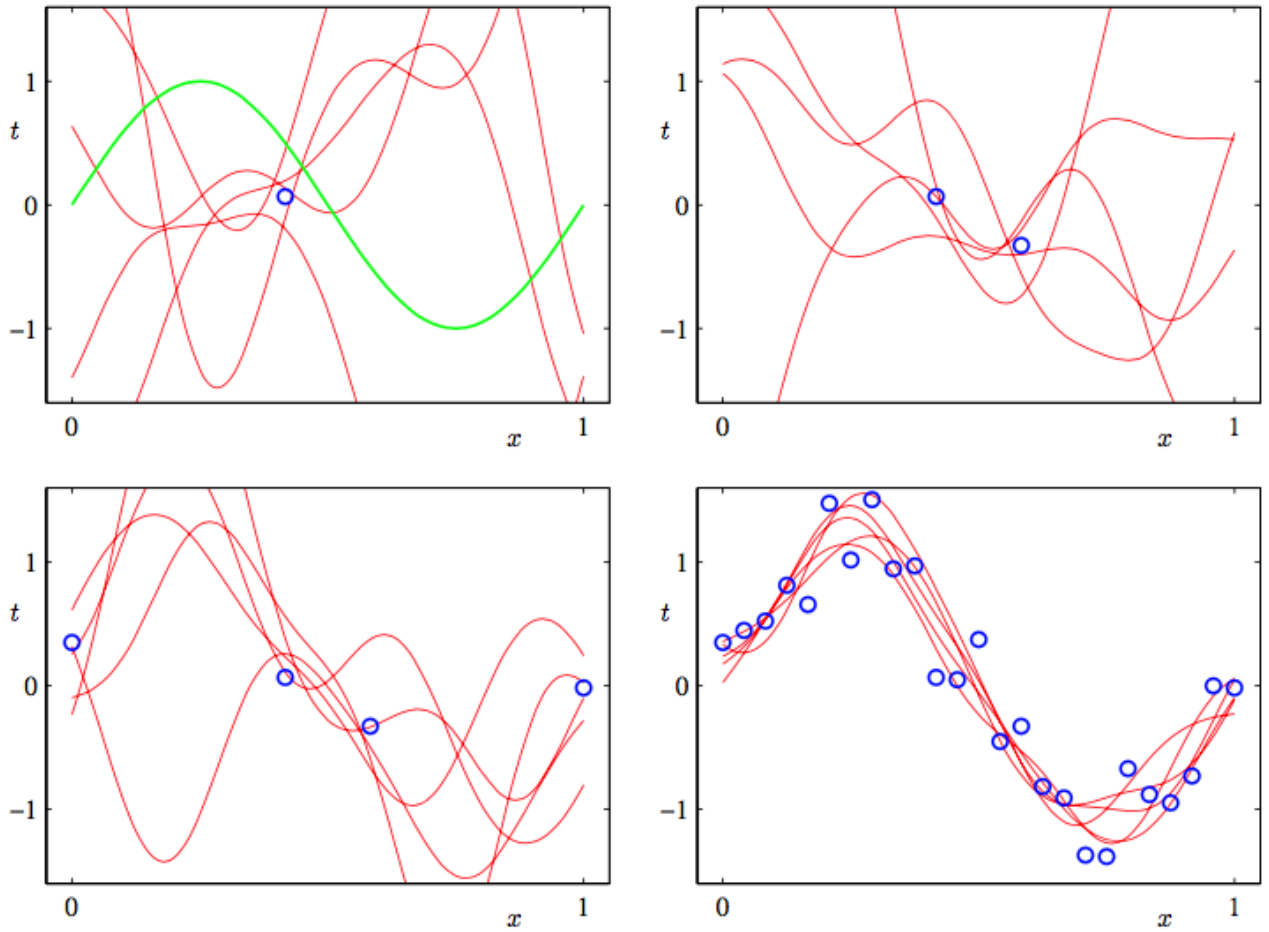


— Bishop, Pattern Recognition and Machine Learning

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_J(x) \end{pmatrix}$$

Bayesian Linear Regression

Functions sampled from the posterior:



— Bishop, Pattern Recognition and Machine Learning

Bayesian Linear Regression

- Posterior predictive distribution:

$$p(t \mid \mathbf{x}, \mathcal{D}) = \int \underbrace{p(t \mid \mathbf{x}, \mathbf{w})}_{\mathcal{N}(t\,;\,\mathbf{w}^\top \phi(\mathbf{x}), \sigma^2)} \underbrace{p(\mathbf{w} \mid \mathcal{D})}_{\mathcal{N}(\mathbf{w}\,;\,\boldsymbol{\mu}, \boldsymbol{\Sigma})} \, \mathrm{d}\mathbf{w}$$

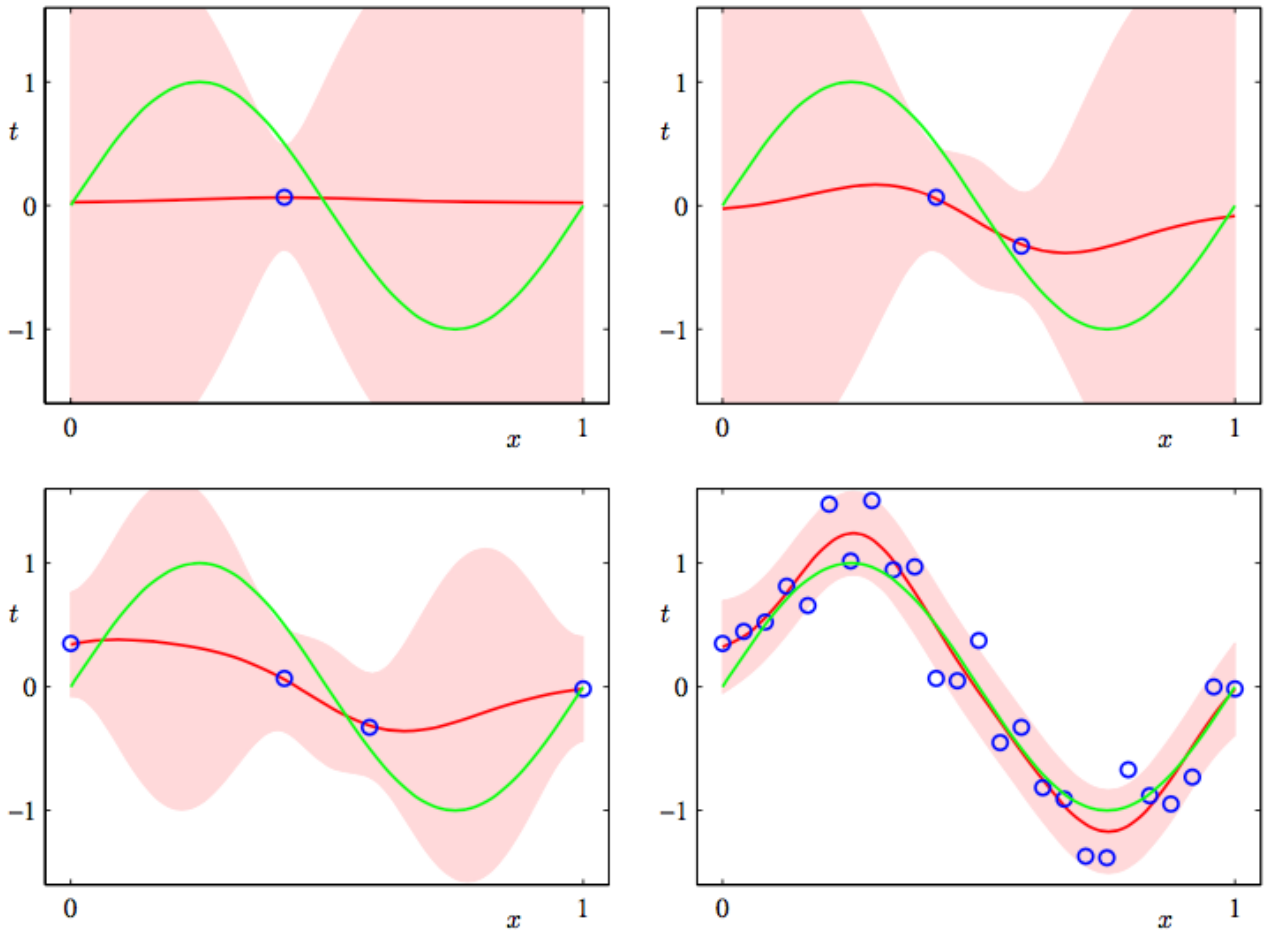
- Another interpretation: $t = \mathbf{w}^\top \phi(\mathbf{x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of $\mathbf{w}$.
- By the linear combination rules for Gaussian random variables, t is a Gaussian distribution with parameters

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

- 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:



— Bishop, Pattern Recognition and Machine Learning

Bayesian Decision Theory

- What do we actually do with the posterior predictive distribution $p(t \mid \mathbf{x}, \mathcal{D})$?
- Often, we want to make a decision. We can formulate this as minimizing an expected loss under the posterior predictive distribution. This is known as **decision theory**.
- Simple example: we have an entire distribution over targets $p(t \mid \mathbf{x}, \mathcal{D})$. How should we choose a single prediction to make?
- One criterion: choose single prediction y to minimize the expected squared error loss.

$$\arg\min_y \mathbb{E}_{p(t \mid \mathbf{x}, \mathcal{D})}[(y - t)^2] = \mathbb{E}_{p(t \mid \mathbf{x}, \mathcal{D})}[t] = \mu_{\text{pred}}$$

  - Same derivation as bias/variance from Lecture 4
- Another criterion: minimize the expected absolute value loss. You can show that you should pick the median of $p(t \mid \mathbf{x}, \mathcal{D})$

Optional material

Occam's Razor (optional)

- 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

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):

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

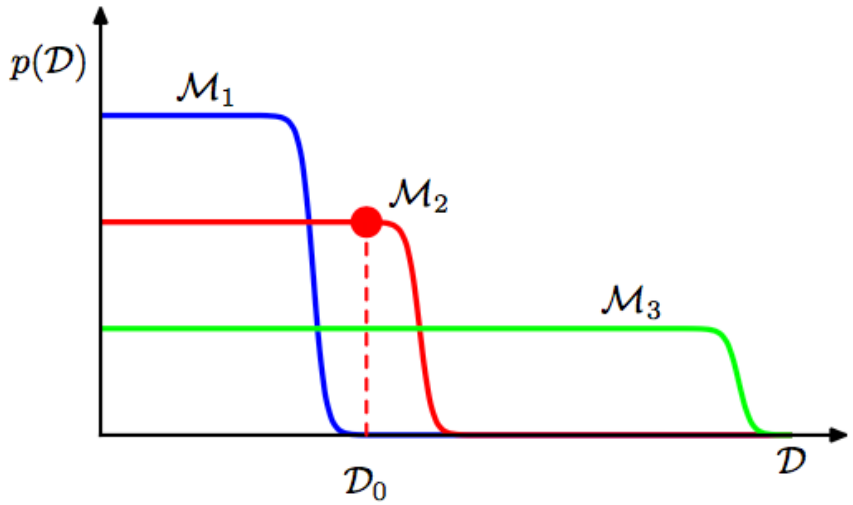
- Which of these terms do you think is more important?
- The evidence is also called **marginal likelihood** since it requires marginalizing out the parameters:

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

- If we're comparing a handful of hypotheses, $p(\mathcal{H}_i)$ isn't very important, so we can compare them based on marginal likelihood.

Occam's Razor (optional)

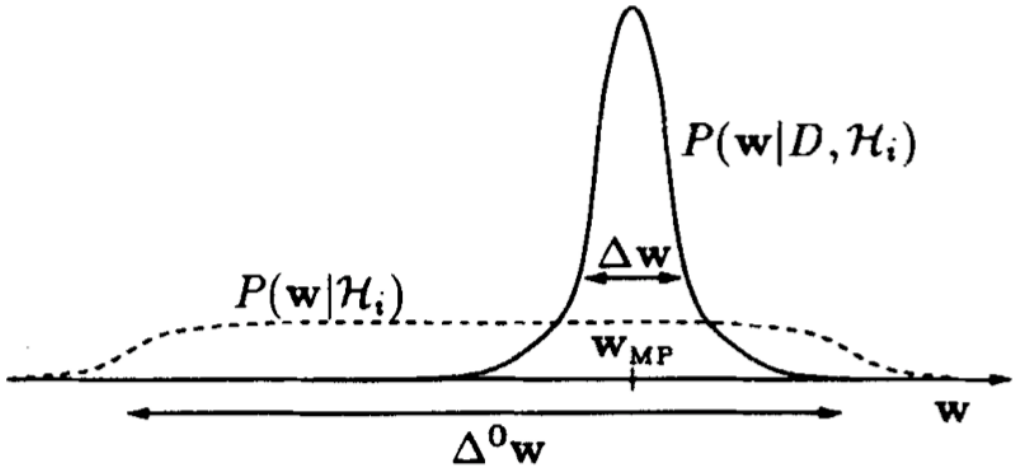
- Suppose M_1, M_2, and M_3 denote a linear, quadratic, and cubic model.
- M_3 is capable of explaning more datasets than M_1.
- But its distribution over $\mathcal{D}$ must integrate to 1, so it must assign lower probability to ones it can explain.



— Bishop, Pattern Recognition and Machine Learning

Occam's Razor (optional)

- How does the evidence (or marginal likelihood) penalize complex models?



- Approximating the integral:

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

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

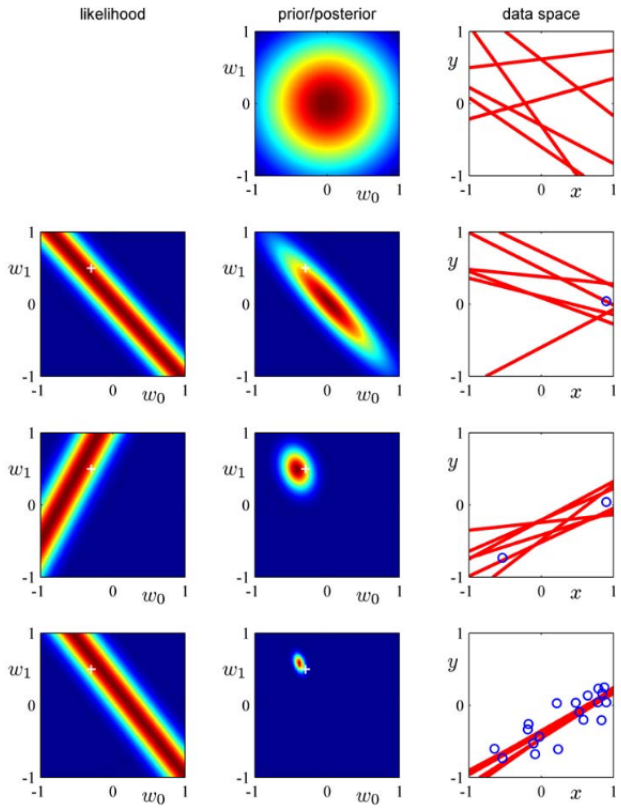
Occam's Razor (optional)

- Multivariate case:

$$p(\mathcal{D} \mid \mathcal{H}_i) \simeq \underbrace{p(\mathcal{D} \mid \mathbf{w}_{\text{MAP}}, \mathcal{H}_i)}_{\text{best-fit likelihood}} \underbrace{p(\mathbf{w}_{\text{MAP}} \mid \mathcal{H}_i)\, |\mathbf{A}|^{-1/2}}_{\text{Occam factor}},$$

  where $\mathbf{A} = \nabla^2_{\mathbf{w}} \log p(\mathcal{D} \mid \mathbf{w}, \mathcal{H}_i)$

- The determinant appears because we're taking the volume.
- The more parameters in the model, the higher dimensional the parameter space, and the faster the volume decays.



— Bishop, Pattern Recognition and Machine Learning

Occam's Razor (optional)

- Analyzing the asymptotic behavior:

$$
\begin{aligned}
\mathbf{A} &= \nabla^2_{\mathbf{w}} \log p(\mathcal{D} \mid \mathbf{w}, \mathcal{H}_i) \\
&= \sum_{j=1}^{N} \underbrace{\nabla^2_{\mathbf{w}} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}, \mathcal{H}_i)}_{\triangleq A_i} \\
&\approx N \, \mathbb{E}[A_i] \\
\text{log Occam factor} &= \log p(\mathbf{w}_{\text{MAP}} \mid \mathcal{H}_i) + \log |\mathbf{A}|^{-1/2} \\
&\approx \log p(\mathbf{w}_{\text{MAP}} \mid \mathcal{H}_i) + \log |N \, \mathbb{E}[A_i]|^{-1/2} \\
&= \log p(\mathbf{w}_{\text{MAP}} \mid \mathcal{H}_i) - \frac{1}{2} \log |\mathbb{E}[A_i]| - \frac{D \log N}{2} \\
&= \text{const} - \frac{D \log N}{2}
\end{aligned}
$$

- **Bayesian Information Criterion (BIC)**: penalize the complexity of your model by $\frac{1}{2} D \log N$.

Occam's Razor (optional)

- Summary

$$p(\mathcal{H}_i \mid \mathcal{D}) \propto p(\mathcal{H}_i)\, p(\mathcal{D} \mid \mathcal{H}_i)$$
$$p(\mathcal{D} \mid \mathcal{H}_i) \simeq p(\mathcal{D} \mid \mathbf{w}_{\mathrm{MAP}}, \mathcal{H}_i)\, p(\mathbf{w}_{\mathrm{MAP}} \mid \mathcal{H}_i)\, |\mathbf{A}|^{-1/2}$$

  Asymptotically, with lots of data, this behaves like

$$\log p(\mathcal{D} \mid \mathcal{H}_i) = \log p(\mathcal{D} \mid \mathbf{w}_{\mathrm{MAP}}, \mathcal{H}_i) - \frac{1}{2} D \log N.$$

- Occam's Razor is about integration, not priors (over hypotheses).