ISTA 421/521 Introduction to Machine Learning

Lecture 13: Bayesian Olympics, Marginal Likelihood Model Selection

Clay Morrison

clayton@sista.arizona.edu

Gould-Simpson 819

Phone 621-6609

7 October 2014

Review

Let θ be some unobserved (population) parameter.
The function $\theta \mapsto f(x|\theta)$ is the likelihood function.
The *maximum likelihood* (ML) estimate of θ is then:

$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} f(x|\theta)$$

Now let's treat θ as a random variable itself.
Let g be a prior distribution over θ.
The posterior distribution of θ is defined, using Bayes' Theorem:

$$\theta \mapsto f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\vartheta\in\Theta} f(x|\vartheta)g(\vartheta)\mathrm{d}\vartheta}$$

The *maximum a posteriori* (MAP) estimate of θ is the **mode** of the posterior distribution.

$$\hat{\theta}_{\mathrm{MAP}} x = \arg\max_{\theta} \frac{f(x|\theta)g(\theta)}{\int_{\vartheta\in\Theta} f(x|\vartheta)g(\vartheta)\mathrm{d}\vartheta} = \arg\max_{\theta} f(x|\theta)g(\theta)$$

Note: The denominator (the **marginal likelihood**, **probability of the evidence**, **partition function**) does not depend on θ, so it plays no role in the optimization!

Note: When the prior g is uniform (a constant function), then the MAP estimate Of θ is the same as the ML estimate.

Review

We can select our parameters so that they maximize the marginal likelihood (type II maximum likelihood)

$$p(y_N|\alpha,\beta) = \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha,\beta)\;dr$$

Finally, the "full" Bayesian approach: keep the full posterior over your parameters and when you must make a decision, integrate over the uncertainty in the parameters.

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{10} \le 6|r)\right\} = \int_{r=0}^{r=1} P(Y_{10} \le 6|r)p(r|y_N)\;dr$$

Return (again) to the Olympics 100m

The Bayesian treatment…

- First, the model:

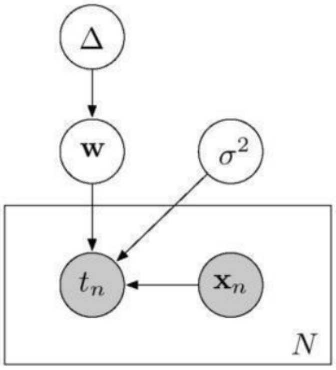
$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_k x_n^k + \epsilon_n$$

kth-order polynomial (Ch 1)
Gaussian distributed noise (Ch 2)

$$\epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \qquad \mathbf{t} = \mathbf{X}^\top \mathbf{w} + \epsilon$$

$$\begin{aligned} p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta) &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2,\Delta)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X},\sigma^2,\Delta)} \\ &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X},\sigma^2,\Delta)}. \\ &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)\;d\mathbf{w}} \end{aligned}$$



Predictions, Likelihood & Prior

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)\ d\mathbf{w}}$$

Can use $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ to make predictions:

$$p(t_{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \sigma^2, \Delta) = \int p(t_{new}|\mathbf{x}_{new}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)\ d\mathbf{w}$$

$$p(t_{new} < 9.5|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \sigma^2, \Delta) = \int p(t_{new} < 9.5|\mathbf{x}_{new}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)\, d\mathbf{w}$$

The Likelihood:

$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N)$ analogous to binomial likelihood in coin example

The Prior:

Want an exact posterior, so want prior that is conjugate to the Gaussian likelihood

$p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ analogous to beta prior in coin example

The Posterior

We know that a Gaussian prior is conjugate with a Gaussian likelihood, so the posterior is Gaussian!
Our goal is therefore to multiply the two and manipulate the prior and likelihood to get them into a single Gaussian form.

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mu_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})$$

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) &\propto p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
&= \frac{1}{(2\pi)^{N/2}|\sigma^2\mathbf{I}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{t}-\mathbf{X}\mathbf{w})\right) \\
&\quad \times \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_0|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w})\right)\exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right) \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w}) + (\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right)\right\}. \\
&\propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \frac{1}{\sigma^2}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{w} - 2\boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}
\end{aligned}$$

Ignore any term that doesn't involve **w**

The Posterior

$$\propto \exp\left\{-\frac{1}{2}\left(\overset{\text{linear}}{-\frac{2}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X}\mathbf{w}} + \overset{\text{quadratic}}{\frac{1}{\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}} + \overset{\text{quadratic}}{\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}} - \overset{\text{linear}}{2\boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}}\right)\right\}$$

The form we want…

$$\begin{aligned} p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) &= \mathcal{N}(\boldsymbol{\mu}_\mathbf{w},\boldsymbol{\Sigma}_\mathbf{w}) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})\right) \\ &\propto \exp\left\{-\frac{1}{2}\left(\underset{\text{quadratic}}{\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}} - \underset{\text{linear}}{2\boldsymbol{\mu}_\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}}\right)\right\} \end{aligned}$$

Combine the quadratic terms…

$$\begin{aligned} \mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} &= \frac{1}{\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} + \mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\ &= \mathbf{w}^\mathsf{T}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)\mathbf{w} \end{aligned}$$

$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

Combine the linear terms…

$$\begin{aligned} -2\boldsymbol{\mu}_\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} &= -\frac{2}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X}\mathbf{w} - 2\boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\ \boldsymbol{\mu}_\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} &= \frac{1}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X}\mathbf{w} + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\ \boldsymbol{\mu}_\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1} &= \frac{1}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X} + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1} \\ \boldsymbol{\mu}_\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{w}^{-1}\boldsymbol{\Sigma}_\mathbf{w} &= \left(\frac{1}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X} + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_\mathbf{w} \\ \boldsymbol{\mu}_\mathbf{w}^\mathsf{T} &= \left(\frac{1}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X} + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_\mathbf{w} \end{aligned}$$

$$\boldsymbol{\Sigma}_\mathbf{w}^\mathsf{T} = \boldsymbol{\Sigma}_\mathbf{w}$$

$$\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

The Posterior

In summary:

$$\begin{aligned} p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) &= \mathcal{N}(\boldsymbol{\mu}_\mathbf{w},\boldsymbol{\Sigma}_\mathbf{w}) \\ \boldsymbol{\Sigma}_\mathbf{w} &= \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \\ \boldsymbol{\mu}_\mathbf{w} &= \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) \end{aligned}$$

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + N\lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

$$\mu_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_\mathbf{0}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_\mathbf{0}^{-1}\mu_\mathbf{0}\right)$$

Given that the posterior is a Gaussian, the single most likely value of **w** is the mean of the posterior, $\mu_\mathbf{w}$

This is the **maximum a posteriori** (MAP) estimate of **w**
… and is the maximum of (the product of the likelihood and the prior):

$$p(\mathbf{w},\mathbf{t}|\mathbf{X},\sigma^2,\Delta)$$

Recall that the squared loss considered in Chapter 1 is very similar to the Gaussian likelihood.
Computing the most likely posterior (when the likelihood is Gaussian) is equivalent to using regularized least squares!
Can help provide intuition about effect of prior: inverse of prior covariance.

Consider 1st-Order Polynomial

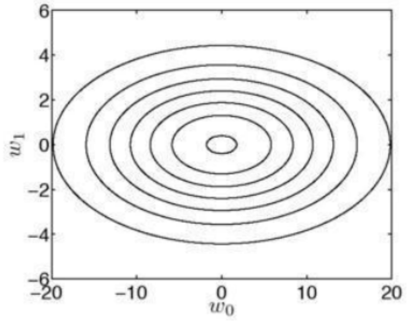
- With a 1st-order polynomial, we can visualize the two-dimensions of the parameters **w**.
- Choose our priors:

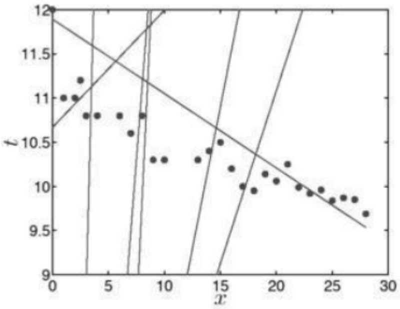
variance for w_0 (intercept term)

$$\mu_\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{\Sigma_0} = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$$

variance for w_1 (slope term)

Zero's indicate prior independence
Does not preclude posterior dependence!





Prior and Posterior as we add Data

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \mathbf{\Sigma}_\mathbf{w})$$

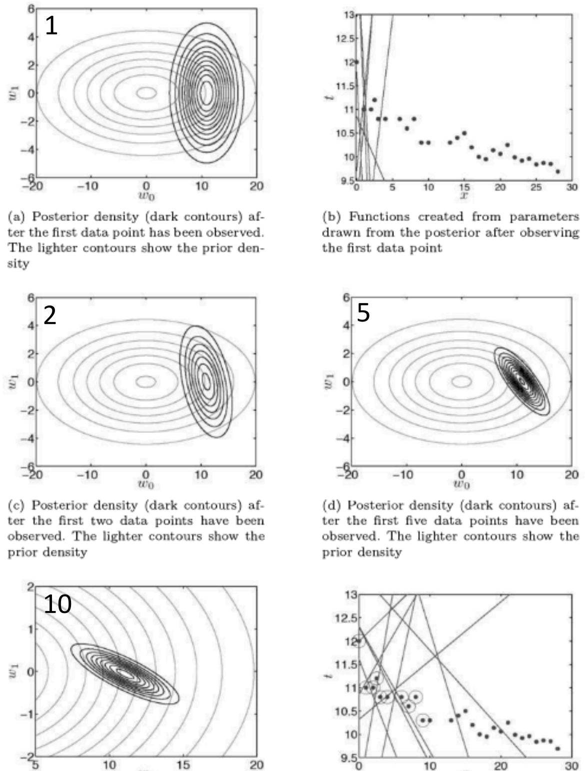
$$\mathbf{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_\mathbf{w} = \mathbf{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

assume $\sigma^2 = 10$
(actually pretty high, but for illustration)

1: Lots of info about intercept,
No info about slope

2, 5, 10: Get's more dense!
Starts to tilt: dependency between **w**$_0$ and **w**$_1$
(based entirely on evidence)



(a) Posterior density (dark contours) after the first data point has been observed. The lighter contours show the prior density

(b) Functions created from parameters drawn from the posterior after observing the first data point

(c) Posterior density (dark contours) after the first two data points have been observed. The lighter contours show the prior density

(d) Posterior density (dark contours) after the first five data points have been observed. The lighter contours show the prior density

(e) Posterior density (dark contours) after the first 10 data points have been observed. The lighter contours show the prior density. (Note that we have zoomed in)

(f) Functions created from parameters drawn from the posterior after observing the first 10 data points (these data points are highlighted)

Prior and Posterior as we add Data

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})$$

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

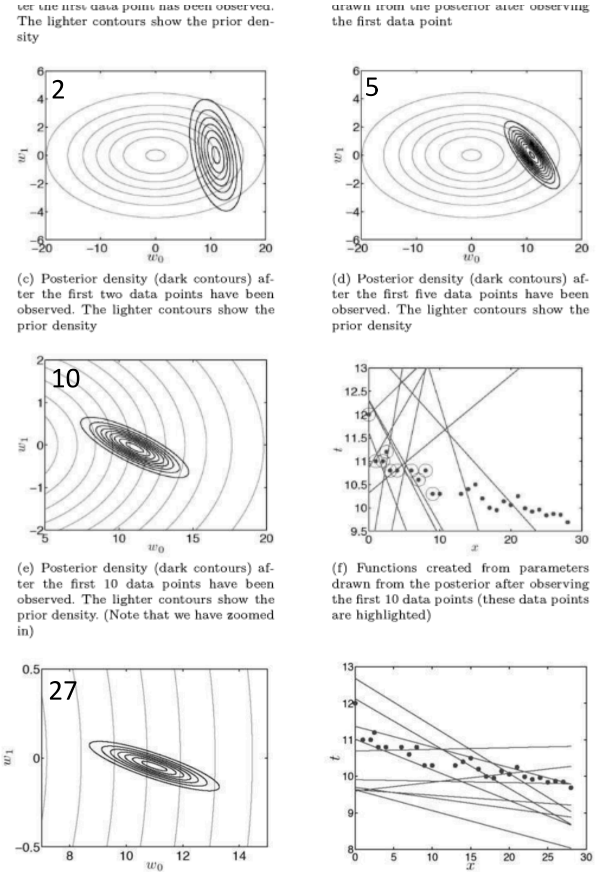
assume $\sigma^2 = 10$

(actually pretty high, but for illustration)

1: Lots of info about intercept,
No info about slope

2, 5, 10: Get's more dense!
Starts to tilt: dependency between $\mathbf{w}_0$ and $\mathbf{w}_1$
(based entirely on evidence)

27: Posterior distribution now
Still lots of variance in sample due to $\sigma^2 = 10$


ter the first data point has been observed. The lighter contours show the prior density

drawn from the posterior after observing the first data point

(c) Posterior density (dark contours) after the first two data points have been observed. The lighter contours show the prior density

(d) Posterior density (dark contours) after the first five data points have been observed. The lighter contours show the prior density

(e) Posterior density (dark contours) after the first 10 data points have been observed. The lighter contours show the prior density. (Note that we have zoomed in)

(f) Functions created from parameters drawn from the posterior after observing the first 10 data points (these data points are highlighted)

Prior and Posterior as we add Data

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}})$$

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

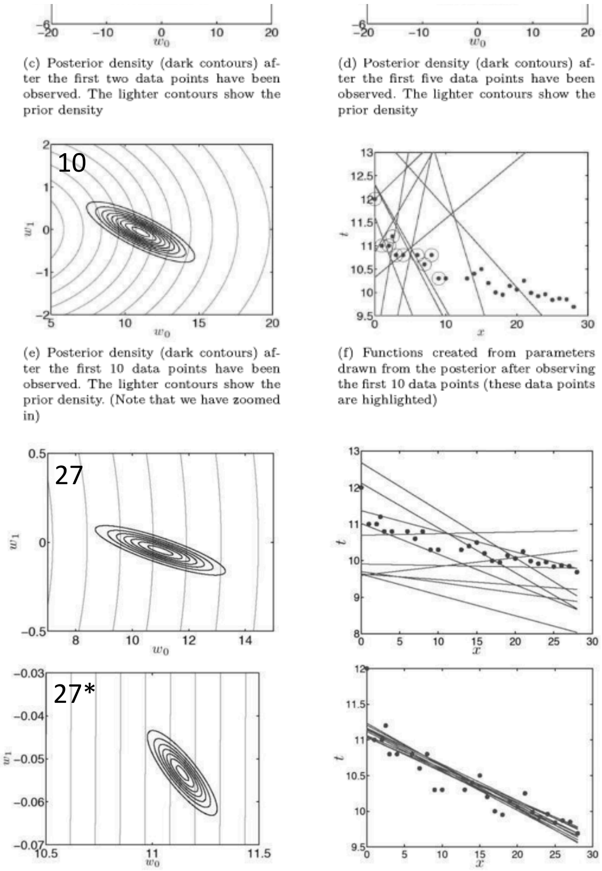
assume $\sigma^2 = 10$

(actually pretty high, but for illustration)

1: Lots of info about intercept,
No info about slope

2, 5, 10: Get's more dense!
Starts to tilt: dependency between $\mathbf{w}_0$ and $\mathbf{w}_1$
(based entirely on evidence)

27*: Very tight posterior distribution
Now adjust to $\sigma^2 = 0.05$


(c) Posterior density (dark contours) after the first two data points have been observed. The lighter contours show the prior density

(d) Posterior density (dark contours) after the first five data points have been observed. The lighter contours show the prior density

(e) Posterior density (dark contours) after the first 10 data points have been observed. The lighter contours show the prior density. (Note that we have zoomed in)

(f) Functions created from parameters drawn from the posterior after observing the first 10 data points (these data points are highlighted)

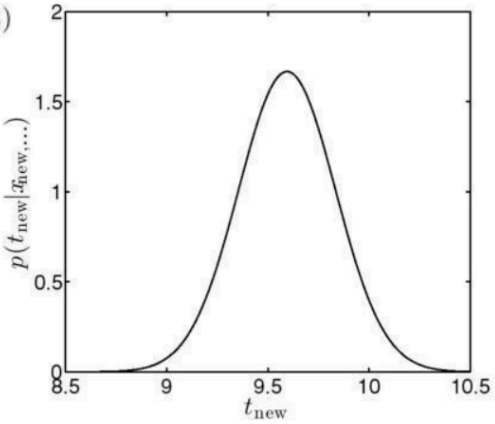
Making Predictions

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathbf{E}_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)}\left\{p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)\right\}$$
$$= \int p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2)\ d\mathbf{w}$$

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathsf{T}}\mathbf{w}, \sigma^2)$$

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\mu}_{\mathbf{w}}, \sigma^2 + \mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{x}_{\text{new}})$$
$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$
$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathcal{N}(9.5951, 0.0572)$$


2
1.5
1
0.5
0
8.5
9
9.5
10
10.5
$p(t_{\text{new}}|x_{\text{new}},\ldots)$
t_{new}