ISTA 421/521 Introduction to Machine Learning

Lecture 8: Maximum Likelihood – Uncertainty in Parameters and Predictions

Clay Morrison

clayton@sista.arizona.edu

Gould-Simpson 819

Phone 621-6609

18 September 2014

Topics

- Review/complete Maximum Likelihood estimation
- Review: (What to expect from) Expectation
- Revisiting the generative picture
- Maximum Likelihood Estimation
  - Uncertainty in parameters
  - Uncertainty in predictions

Explicitly Modeling Uncertainty

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n, \ \ \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

Adding a R.V to our linear model: … to model noise/variance/uncertainty in the response (*t*).

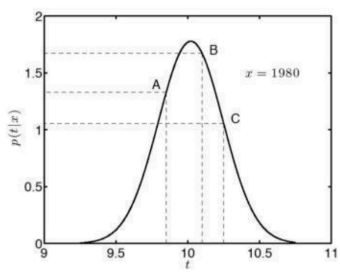
$$p(t_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

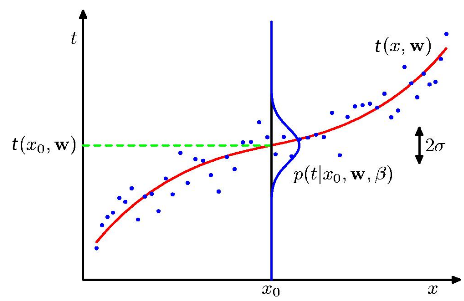
Now our "response" is itself a random variable. (Here we assume Gaussian/Normal noise.)

We assume (conditional) independence: given a particular input **x** and parameters **w**, the predicted variance at that point is indep. of any other point. However, we also assume the variance form (i.e., what type of distribution) and its parameters will have the same value at any point (so are constant, not a fn of **x** or **w**).

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

We can now use the data (input and target pairs) to give us a "likelihood" score of how well the model "fits" the mass/density of our model's uncertainty over the data. The more the data all fits under the bulk of the probability mass/density, the more it looks like our model could have generated the data (the more "likely" the data looks given our model).





Maximize the Likelihood

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

Since we are working with a product of Gaussians, which in turn include The exponential function (*e*), take the natural log (often just represented Generically as log(*L*))

$$\begin{aligned}
\log L &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n; \mathbf{w}))^2\right\}\right) \\
&= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n, \mathbf{w}))^2\right) \\
&= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n; \mathbf{w}))^2.
\end{aligned}$$

$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}.$

Maximize the Likelihood: w

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2$$
$$= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2$$

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n(t_n - \mathbf{x}_n^\mathsf{T}\mathbf{w})$$
$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n t_n - \mathbf{x}_n\mathbf{x}_n^\mathsf{T}\mathbf{w} = \mathbf{0}$$

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\mathsf{T}\\ \mathbf{x}_2^\mathsf{T}\\ \vdots\\ \mathbf{x}_N^\mathsf{T}\end{bmatrix} = \begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_N\end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix}t_1\\ t_2\\ \vdots\\ t_N\end{bmatrix}$$

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = \mathbf{0}$$

$$\begin{aligned}\frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) &= 0\\ \mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} &= 0\\ \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} &= \mathbf{X}^\mathsf{T}\mathbf{t}\\ \mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\end{aligned}$$

Maximize the Likelihood: σ

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2$$

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \mathbf{x}^\mathsf{T}\widehat{\mathbf{w}})^2 = 0$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\mathsf{T}\widehat{\mathbf{w}})^2$$

$$\begin{aligned}\sigma^2 &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})\\ &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}} + \widehat{\mathbf{w}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}})\end{aligned}$$

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} + \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})\\ &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} + \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})\\ &= \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})\end{aligned}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}})$$

Simplify further by plugging in

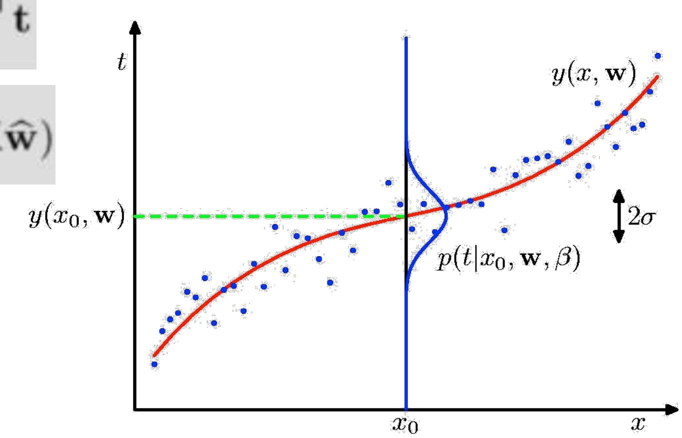
$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

Maximum Likelihood

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\hat{\mathbf{w}})$$



Predictive distribution

The equations are unique

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\hat{\mathbf{w}})$$

Hessian Matrix

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_K} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_K \partial w_1} & \frac{\partial^2 f}{\partial w_K \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_K^2} \end{bmatrix}$$

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2$$

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w})$$

$$\frac{\partial^2 \log L}{\partial \mathbf{w}\partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

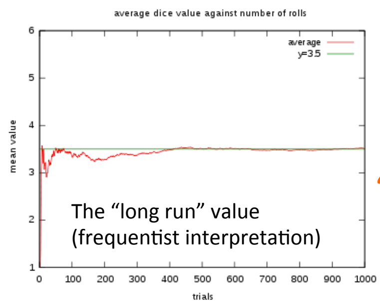
This matrix is definite negative, so maximum likelihood solution (with linear model with additive Gaussian noise) is unique. $\mathbf{x}^\mathsf{T}\mathbf{H}\mathbf{x} < 0$

Expectation

- The **expected value** of a random variable is the weighted (by probability) average of all possible values

$$\mathbf{E}_{P(x)}\left\{f(x)\right\} = \sum_x f(x)P(x)$$

$$\mathbf{E}_{p(\mathbf{x})}\left\{f(\mathbf{x})\right\} = \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}x$$



"Belief" in possible worlds
(Bayesian interpretation)

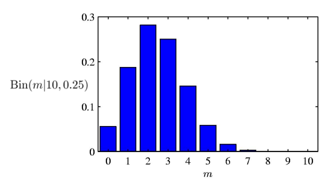
Mean:

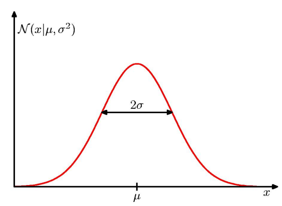
$$\mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\right\} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x})$$

"Variance":

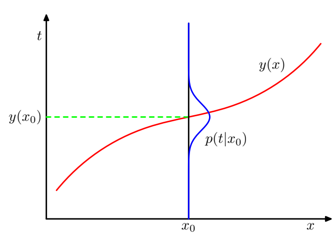
$$\operatorname{cov}\left\{\mathbf{x}\right\} = \mathbf{E}_{P(\mathbf{x})}\left\{(\mathbf{x}-\mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\right\})(\mathbf{x}-\mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\right\})^\top\right\}$$

$$\operatorname{cov}\left\{\mathbf{x}\right\} = \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^\top\right\} - \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\right\}\mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\right\}^\top$$



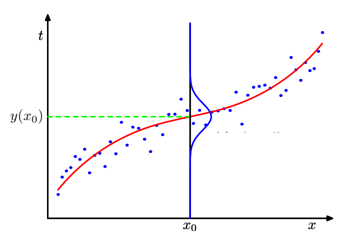


Explicitly Modeling Uncertainty: The Generative Picture

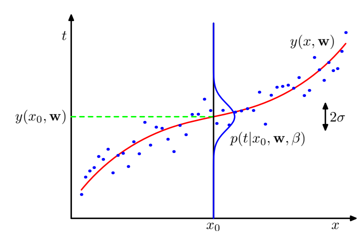


The generating process…

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \quad ; \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$



… generates data …



… that we fit a model to

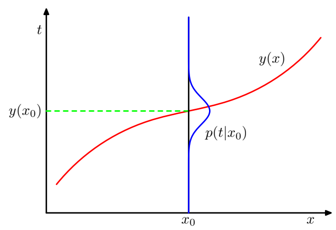
$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\sigma^2) \\ &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top\mathbf{x}_n,\sigma^2) \end{aligned}$$

estimated parameters

$$\begin{aligned} p(\hat{\mathbf{t}}|\mathbf{X},\hat{\mathbf{w}},\hat{\sigma^2}) &= \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n,\hat{\mathbf{w}},\hat{\sigma^2}) \\ &= \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top\mathbf{x}_n,\hat{\sigma^2}) \end{aligned}$$

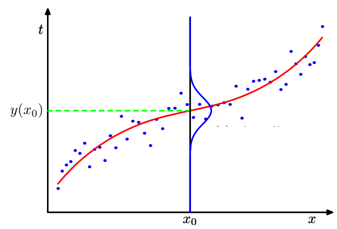
prediction

Explicitly Modeling Uncertainty: The Generative Picture



The generating process…

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \quad ; \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$



… generates data …

If we generate new data from the generating process, we get different parameters $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{\sigma}}^2$, and a different predictive distribution $p(\hat{\mathbf{t}}|\ldots)$

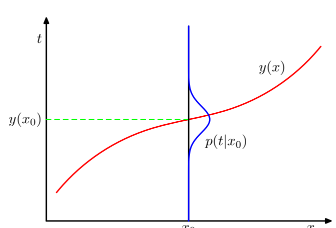
$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$

estimated parameters

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

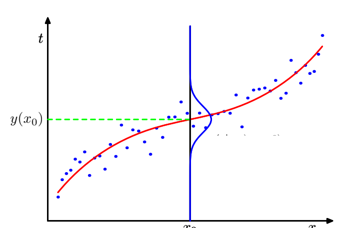
prediction

Explicitly Modeling Uncertainty: The Generative Picture



The generating process…

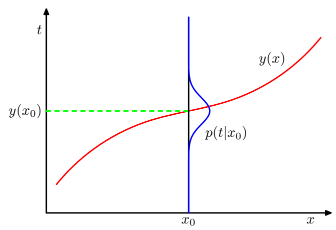
$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \quad ; \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$



… generates data …

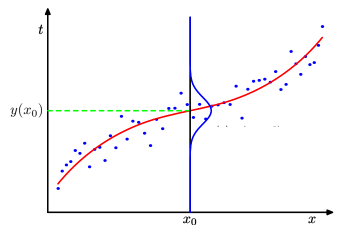
Suppose we generate 10,000 datasets (each slightly different) but from the same generating process…

Explicitly Modeling Uncertainty: The Generative Picture



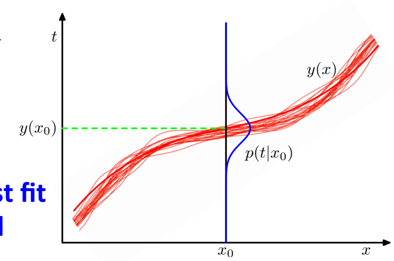
The generating process…

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$



… generates data …

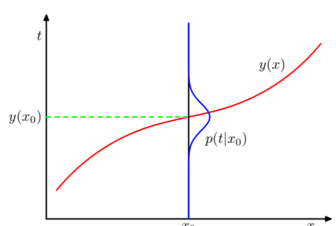
Suppose we generate 10,000 datasets (each slightly different) but from the same generating process…



Each curve here is a best fit to one of the generated datasets.

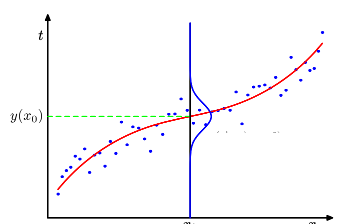
Each is an estimate of the true generating process, as mediated by the dataset

Explicitly Modeling Uncertainty: The Generative Picture



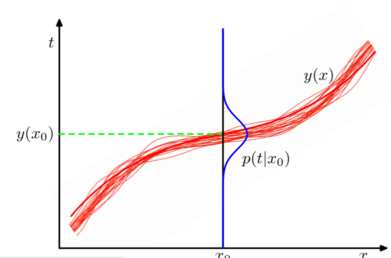
The generating process…

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$



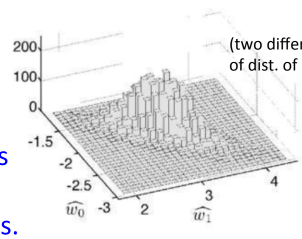
… generates data …

Suppose we generate 10,000 datasets (each slightly different) but from the same generating process…

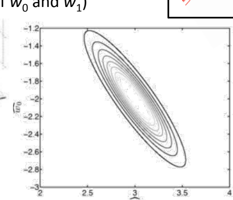


We can estimate the **uncertainty in parameter values** by looking at the ***empirical distribution*** of different parameter values we get from each of the different dataset estimates.

E.g., here is the empirical distribution of w_0 and w_1…

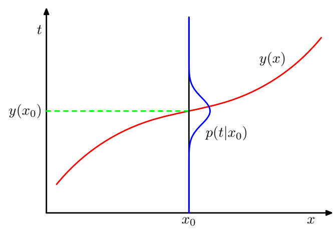


(two different views of dist. of w_0 and w_1)



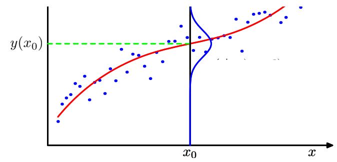
Explicitly Modeling Uncertainty: The Generative Picture

The uncertainty in the parameter estimates is a function of the uncertainty in the data, which in turn is a function of the uncertainty of the original generating process.

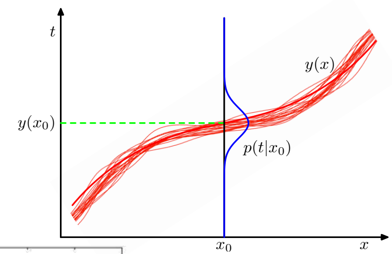


The generating process…

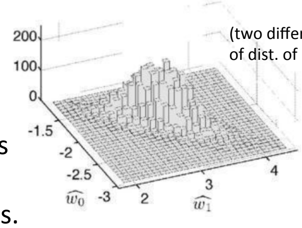
$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \quad ; \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$



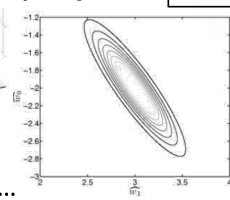
… generates data …



We can estimate the **uncertainty in parameter values** by looking at the ***empirical distribution*** of different parameter values we get from each of the different dataset estimates.



(two different views of dist. of w_0 and w_1)

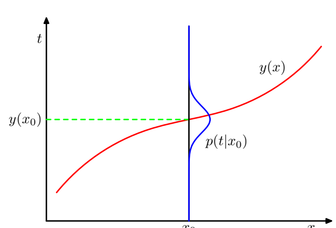


E.g., here is the empirical distribution of w_0 and w_1…

Explicitly Modeling Uncertainty: The Generative Picture

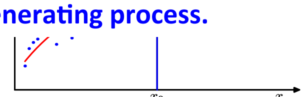
The uncertainty in the parameter estimates is a function of the uncertainty in the data, which in turn is a function of the uncertainty of the original generating process.

Each dataset provides some information about the generating process.

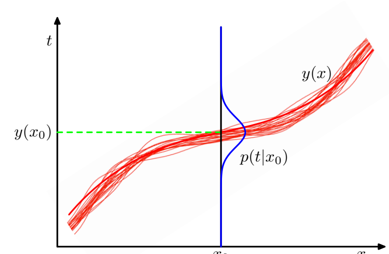


The generating process…

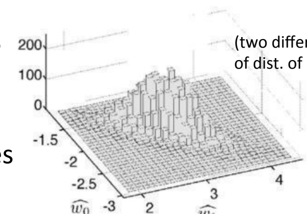
$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \quad ; \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$



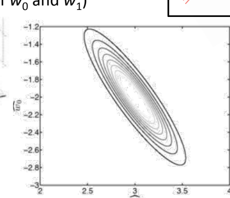
… generates data …



We can estimate the **uncertainty in parameter values** by looking at the ***empirical distribution*** of different parameter values we get from each of the different dataset estimates.

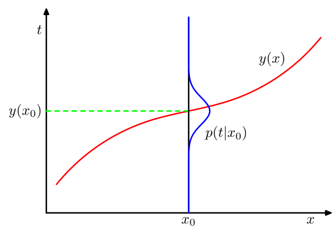


(two different views of dist. of w_0 and w_1)



E.g., here is the empirical distribution of w_0 and w_1…

Explicitly Modeling Uncertainty: The Generative Picture



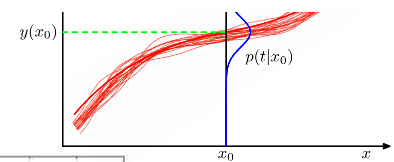
The generating process…

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \ \ ; \ \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

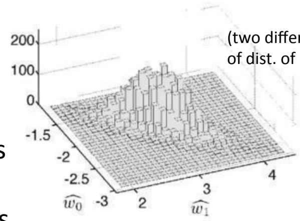
The uncertainty in the parameter estimates is a function of the uncertainty in the data, which in turn is a function of the uncertainty of the original generating process.

Each dataset provides some information about the generating process.

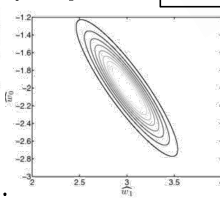
The maximum likelihood fit to the data is therefore an estimate of the inherent uncertainty of the generating process.



We can estimate the **uncertainty in parameter values** by looking at the ***empirical distribution*** of different parameter values we get from each of the different dataset estimates.

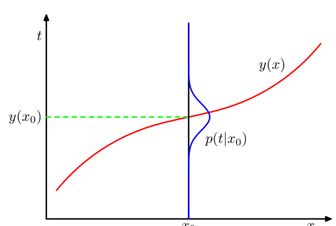


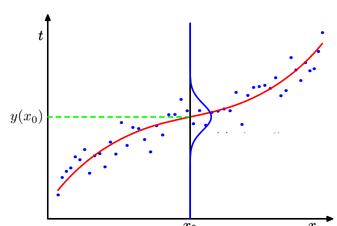
(two different views of dist. of w_0 and w_1)

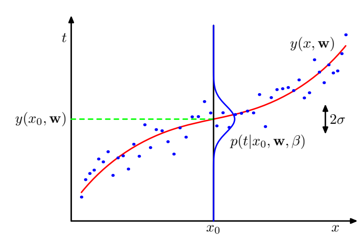


E.g., here is the empirical distribution of w_0 and w_1…

Explicitly Modeling Uncertainty: The Generative Picture







The generating process…

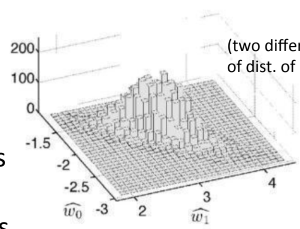
$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \ \ ; \ \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Can we estimate the uncertainty in our parameters from a single dataset?

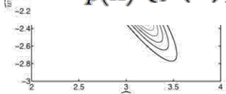
YES! … by considering the ***expected values*** of the parameters with respect to the generating process distribution!

$$\mathbf{E}_{p(\mathbf{x})}\left\{f(\mathbf{x})\right\} = \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}x$$

We can estimate the **uncertainty in parameter values** by looking at the ***empirical distribution*** of different parameter values we get from each of the different dataset estimates.



(two different of dist. of w_0 a



E.g., here is the empirical distribution of w_0 and w_1…

Side Note

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n\mathbf{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n,\sigma^2)$$

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

In the following, easier to deal with multivariate Gaussian rather than product of individual Gaussians

Deriving the Expectation of $\hat{w}$ w.r.t. the generating distribution

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

$\hat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$ Substitute in $\hat{w}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\int \mathbf{t}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t}$$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\}$$

$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} = \mathbf{X}\mathbf{w}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\} = \mathbf{w}$$

Our estimate $\hat{w}$ of w is inherently **unbiased**!

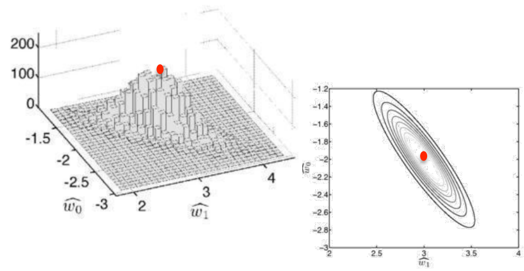
This also means any variance in the estimate is encapsulated in its **(co)variance (matrix)**

Deriving the Expectation of $\hat{w}$ w.r.t. the generating distribution

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

$\hat{\mathrm{w}} = (\mathrm{X}^{\mathsf{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathsf{T}}\mathrm{t}$ Substitute in $\hat{w}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$$



Our estimate $\hat{w}$ of w is inherently **unbiased**!

This also means any variance in the estimate is encapsulated in its **(co)variance (matrix)**

Let's be clear on what we mean by "Maximum likelihood of the mean of a linear Gaussian model is unbiased"

- Here we mean: if we take repeated samples of size *N* from a Normal generator distribution with true parameters **w**, then the collection of maximum likelihood estimated $\hat{\mathbf{w}}$'s will be "centered" (in the sense of the arithmetic mean) around the true **w**.
- Not to be confused with:
  - "better" estimation by collecting more data (it is true that more data makes better estimation here, but that's a sample size effect)
  - "UMVU" estimator: *Uniformly minimum variance unbiased estimator*. This is a stronger claim: an UMVU estimator is also an unbiased estimator that is "closest" to the true parameters, for a given sample size.
- And, there could be:
  - Other *biased* estimators that tend to be closer

Derive the Covariance of $\hat{w}$

w.r.t. the generating distribution

$$\text{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^\mathsf{T}\right\} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\,\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^\mathsf{T}$$

$\text{cov}\{\widehat{\mathbf{w}}\}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\ \mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\} = \mathbf{w}$$

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t})^\mathsf{T}\right\} \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\}\,\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}.\end{aligned}$$

$$\text{cov}\{\mathbf{t}\} = \sigma^2\mathbf{I} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\} - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}^\mathsf{T}$$

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}\,\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}^\mathsf{T} + \sigma^2\mathbf{I} \\ &= \mathbf{X}\mathbf{w}(\mathbf{X}\mathbf{w})^\mathsf{T} + \sigma^2\mathbf{I} \\ &= \mathbf{X}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T} + \sigma^2\mathbf{I}.\end{aligned}$$

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^\mathsf{T}\right\} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\ &\quad +\sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\ &= \mathbf{w}\mathbf{w}^\mathsf{T} + \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}.\end{aligned}$$

$$\begin{aligned}\text{cov}\{\widehat{\mathbf{w}}\} &= \mathbf{w}\mathbf{w}^\mathsf{T} + \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} - \mathbf{w}\mathbf{w}^\mathsf{T} \\ &= \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\end{aligned}$$

$\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$ ← Remember!

$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1}$$

The Fisher Information

$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1}$$

$$\mathcal{I} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{-\frac{\partial^2 \log p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right\}$$

$$\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

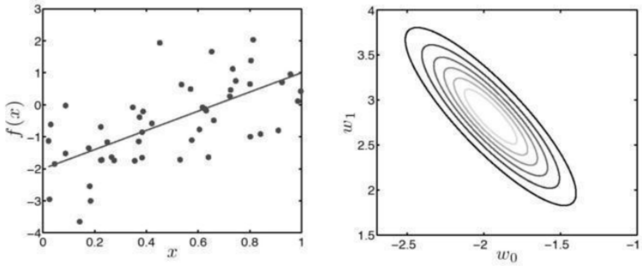
$$\mathcal{I} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right\}$$

$$\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}.$$

The elements of ***I*** tell us how much information (the more negative the value, the more information present) the data provides about the particular parameter (diagonal elements) or pairs of parameters (off-diagonal elements)

Example

$$\mathcal{I} = \begin{bmatrix} 50.0000 & 24.3311 \\ 24.3311 & 15.8953 \end{bmatrix}, \ \mathsf{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.0784 & -0.1200 \\ -0.1200 & 0.2466 \end{bmatrix}$$



$$\mathcal{I} = \begin{bmatrix} 1.2500 \times 10^3 & 0.6083 \times 10^3 \\ 0.6083 \times 10^3 & 0.3974 \times 10^3 \end{bmatrix}, \ \mathsf{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.0031 & -0.0048 \\ -0.0048 & 0.0099 \end{bmatrix}$$