Institute for Advanced Studies
in Basic Sciences
Gava Zang, Zanjan, Iran

Regression with Linear Models

By Dr. Parvin Razzaghi (p.razzaghi@iasbs.ac.ir)

Winter 2024

Slides credit

- Professor:
  - Dr. Parvin Razzaghi
- Original Slides:
  - Amir-massoud Farahman
  - Eric Eaton
- Link:
  - https://B2n.ir/n25709
- University:
  - University of Toronto and Vector Institute
- Acknowledgment:
  - Credit for slides goes to many members of the ML Group at the U of T, and beyond, including (recent past): Roger Grosse, Murat Erdogdu, Richard Zemel, Juan Felipe Carrasquilla, Emad Andrews.
  - For any comments or suggestions, please contact: p.razzaghi@iasbs.ac.ir

ℓ_2 or (L^2) Regularization

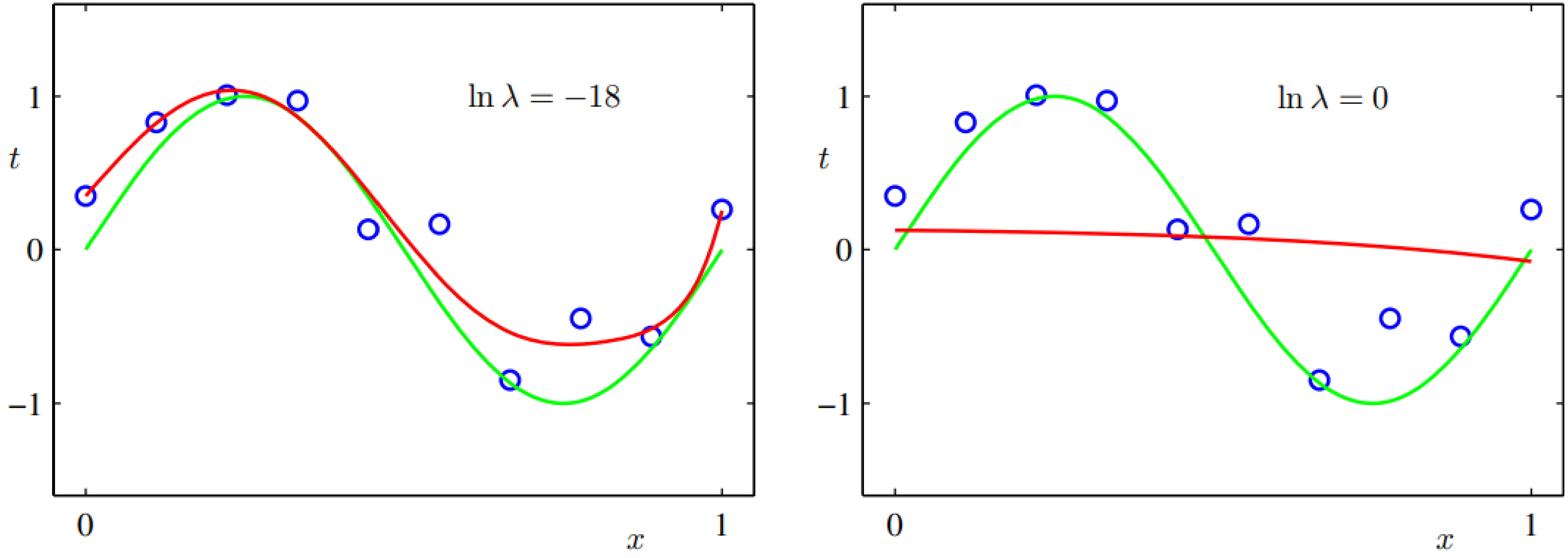
- The regularized cost function

$$\mathcal{J}_{reg}(w) = \mathcal{J}(w) + \lambda \mathcal{R}(w) = \mathcal{J}(w) + \frac{\lambda}{2} \sum_j w_j^2.$$

- The basic idea is that "simpler" functions have weights $\boldsymbol{w}$ with smallerℓ_2-norm and we prefer them to functions with larger ℓ_2-norm.
  - Intuition: Large weights makes the function f have more abrupt changes as a function of the input **x**; it will be less smooth.
- If you fit training data poorly, $\mathcal{J}$ is large. If the fitted weights have high values, $\mathcal{R}$ is large.
- Large λ penalizes weight values more.
- Here λ is a hyperparameter that we can tune with a validation set.

Regularized linear regression

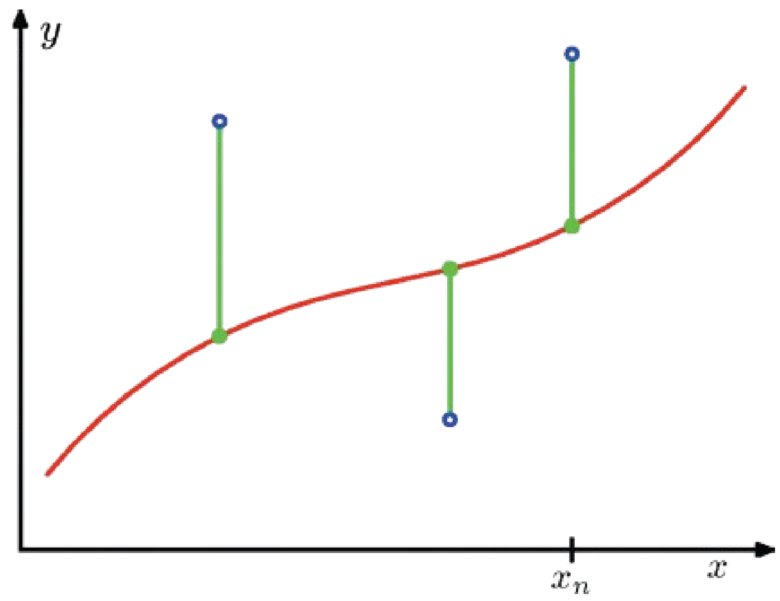
- M=9



Probabilistic Interpretation of the Squared Error

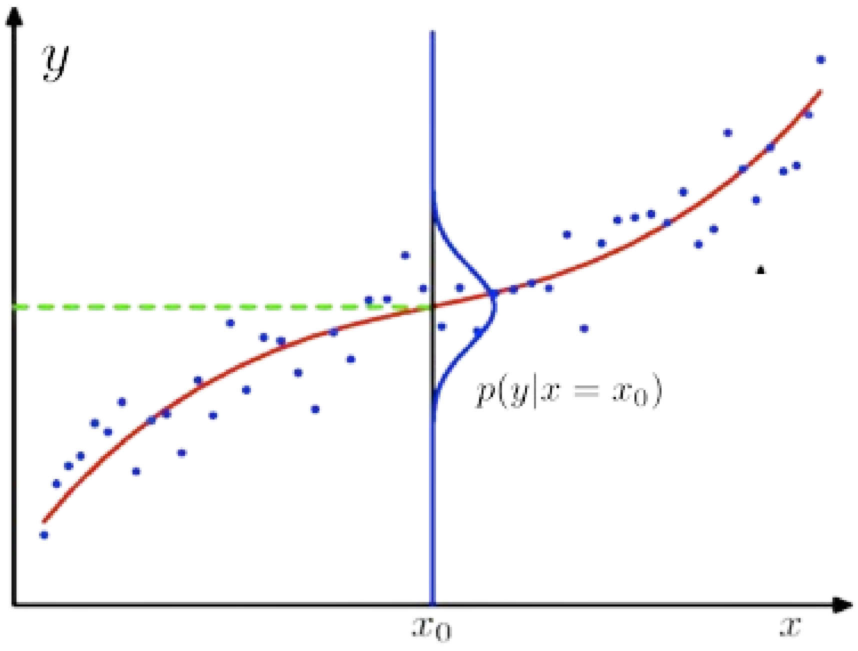
- For the least squares: we minimize the sum of the squares of the errors between the predictions for each data point $\boldsymbol{x}^{(i)}$and the corresponding target values$t^{(i)}$, i.e.,

$$\underset{(\boldsymbol{w},\boldsymbol{w}_0)}{\text{minimize}} \sum_{i=1}^{n} (\boldsymbol{w}^T, \boldsymbol{x}^{(i)} + b - t^{(i)})^2$$



- $t \approx \boldsymbol{x}^T\boldsymbol{w} + b, (\boldsymbol{w}, b) \in \mathbb{R}^D \times \mathbb{R}$
- We measure the quality of the fit using the squared error loss. Why?
- Even though the squared error loss is intuitive, we did not justify it.
- We provide a probabilistic perspective here.
- There are other justifications too; we get to them in the Bias-Variance decomposition lecture.

Probabilistic Interpretation of the Squared Error



- Suppose that our model arose from a statistical model ($b = 0$ for simplicity):

$$y^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \varepsilon^{(i)}$$

where $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ is independent of the input $\boldsymbol{x}^{(i)}$.

- Thus, $y^{(i)}|x^{(i)} \sim p\left(y\middle|\boldsymbol{x}^{(i)}, \boldsymbol{w}\right) = \mathcal{N}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)}, \sigma^2\right)$.

Probabilistic Interpretation of the Squared Error: Maximum Likelihood Estimation

- Suppose that the input data $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$ are given and the outputs are independently drawn from

$$t^{(i)} \sim p\left(y \middle| \boldsymbol{x}^{(i)}, \boldsymbol{w}\right).$$

- with an unknown parameter **w**. So the dataset is

$$D = \left\{\left(\boldsymbol{x}^{(1)}, t^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(N)}, t^{(N)}\right)\right\}.$$

Probabilistic Interpretation of the Squared Error: Maximum Likelihood Estimation

- The likelihood function is $\Pr(D|\boldsymbol{w})$.
- The maximum likelihood estimation (MLE) is based on the "principle" suggesting that we have to find a parameter $\hat{\mathbf{w}}$ that maximizes the likelihood, i.e.,

$$\hat{w} \leftarrow \underset{w}{\operatorname{argmax}} \Pr(D|\boldsymbol{w}) .$$

Maximum likelihood estimation: after observing the data samples $(\boldsymbol{x}^{(t)}, t^{(i)})$ for $i = 1, 2, \ldots, N$, we should choose **w** that maximizes the likelihood.

Probabilistic Interpretation of the Squared Error: Maximum Likelihood Estimation

- For independent samples, the likelihood function of samples D is the product of their likelihoods

$$p\left(t^{(1)}, t^{(2)}, \dots, t^{(N)} \middle| \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}, w\right) = \prod_{i=1}^{N} p\left(t^{(i)}, \boldsymbol{x}^{(i)}, w\right) = L(w).$$

- Product of N terms is not easy to minimize.
- Taking log reduces it to a sum. Two objectives are equivalent since log is strictly increasing.
- Maximizing the likelihood is equivalent to minimizing the negative log-likelihood:

Probabilistic Interpretation of the Squared Error: Maximum Likelihood Estimation

- Maximizing the likelihood is equivalent to minimizing the negative log-likelihood:

$$\ell(w) = -\log L(w) = -\log \prod_{i=1}^{N} p\left(t^{(i)} \middle| \boldsymbol{x}^{(i)}; \boldsymbol{w}\right) = -\sum_{i=1}^{n} \log p\left(t^{(i)} \middle| \boldsymbol{x}^{(i)}; \boldsymbol{w}\right).$$

Probabilistic Interpretation of the Squared Error: Maximum Likelihood Estimation

Maximum Likelihood Estimator (MLE)

After observing $z^{(i)} = \left(\boldsymbol{x}^{(i)}, t^{(i)}\right)$ for $i = 1, \ldots, N$ independent and identically distributed (i.i.d.) samples from $p(z, \mathbf{w})$, MLE is

$$\boldsymbol{w}^{MLE} = \underset{w}{\operatorname{argmin}}\, l(\boldsymbol{w}) = -\sum_{i=1}^{N} \log p\left(t^{(i)} \middle| \boldsymbol{x}^{(i)}; \boldsymbol{w}\right).$$

Probabilistic Interpretation of the Squared Error: From MLE to Squared Error

- Suppose that our model arose from a statistical model:

$$y^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \varepsilon^{(i)}$$

where $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ is independent of anything else.

- $p\left(t^{(i)}, \boldsymbol{x}^{(i)}, w\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2\}$

- $\log p\left(t^{(i)}, \boldsymbol{x}^{(i)}, w\right) = \frac{1}{2\sigma^2}(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2 - \log(\sqrt{2\pi\sigma^2})$

Probabilistic Interpretation of the Squared Error: From MLE to Squared Error

- The MLE solution is

$$\boldsymbol{w}^{MLE} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{w}) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(t^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2 + C$$

- As C and σ do not depend on $\boldsymbol{w}$, they do not contribute to the minimization.

$\boldsymbol{w}^{MLE} = \boldsymbol{w}^{LS}$ when we work with Gaussian densities.

Probabilistic Interpretation of the Squared Error: From MLE to Squared Error

- Suppose that our model arose from a statistical model:

$$y^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \varepsilon^{(i)}$$

where $\varepsilon^{(i)}$ comes from the Laplace distribution, that is, the distribution of $\varepsilon^{(i)}$ has density.

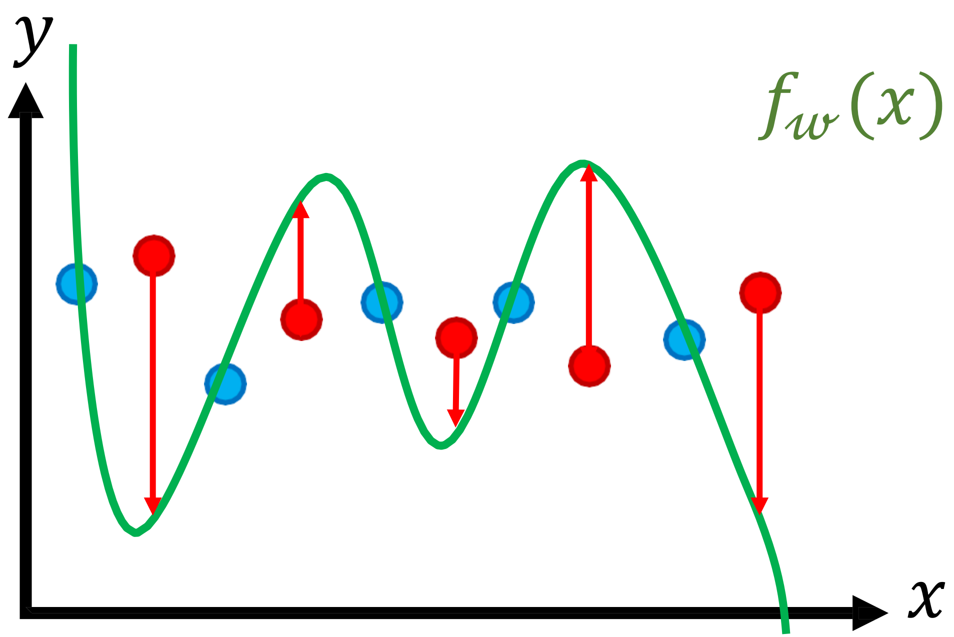
$$\frac{1}{2b} \exp\left(\frac{\left|y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)}\right|}{2b}\right)$$

Probabilistic Interpretation of the Squared Error: From MLE to Squared Error

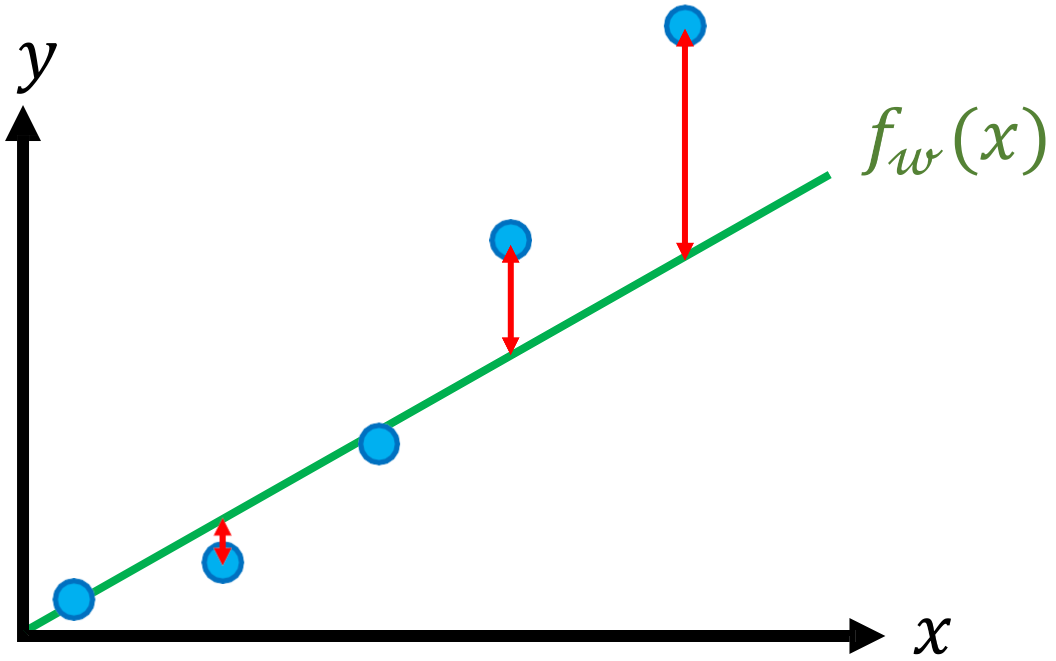
- Q: What is the loss in the MLE?
  - Choice 1: $\frac{1}{N}\sum_{i=1}^{N}|t^{(i)} - w^T x^{(i)}|^{1/2}$
  - Choice 2: $\frac{1}{N}\sum_{i=1}^{N}(t^{(i)} - w^T x^{(i)})$
  - Choice 3: $\frac{1}{N}\sum_{i=1}^{N}|t^{(i)} - w^T x^{(i)}|$
  - Choice 4: $\frac{1}{N}\left|\sum_{i=1}^{N} t^{(i)} - w^T x^{(i)}\right|$
- Q: Can you think of an application area with non-Gaussian probabilistic model?

Bias-Variance Tradeoff

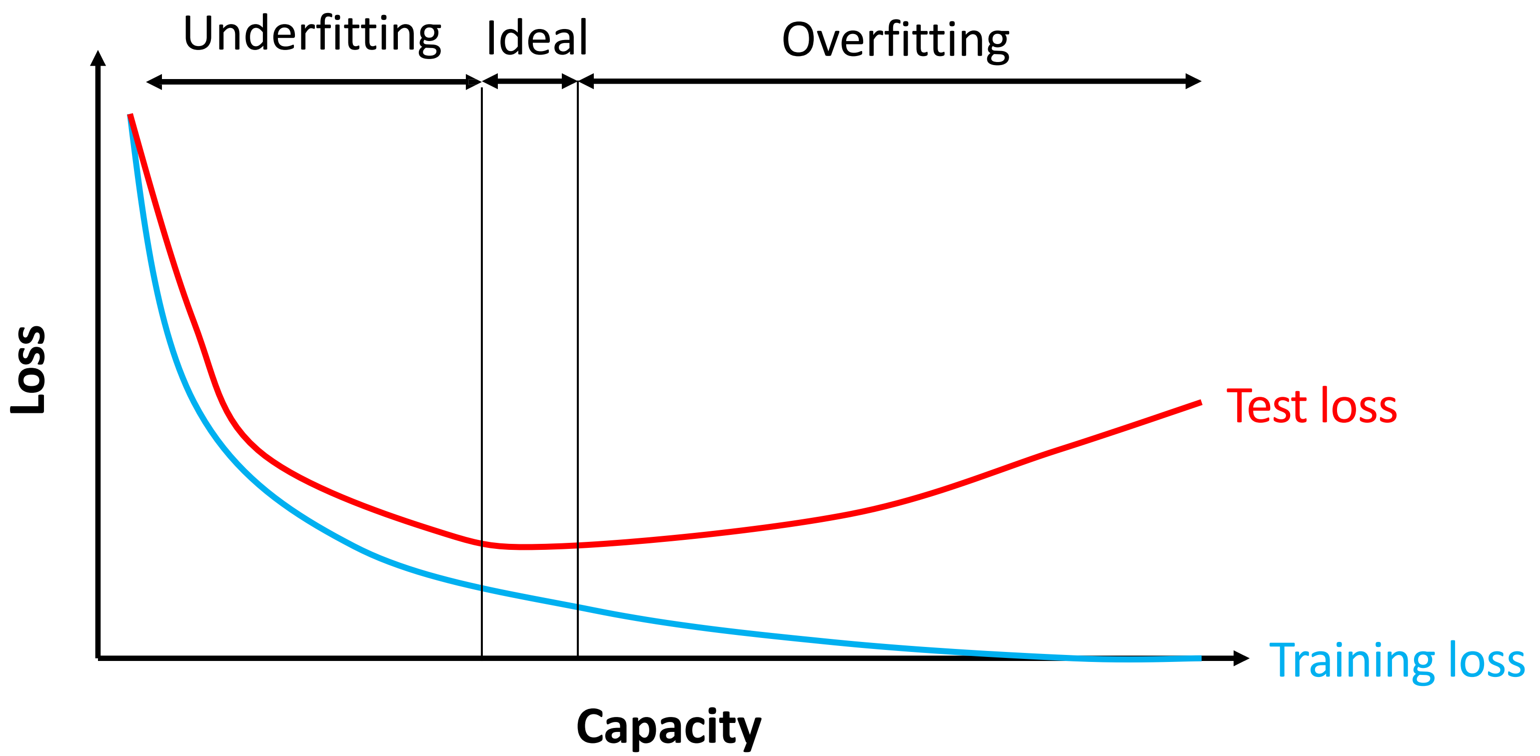
- **Overfitting (high variance)**
  - High capacity model capable of fitting complex data
  - Insufficient data to constrain it



- **Underfitting (high bias)**
  - Low capacity model that can only fit simple data
  - Sufficient data but poor fit



Bias-Variance Tradeoff



The Bias-Variance Decomposition$_{(1)}$

- Recall the *expected squared loss*,

$$\mathbb{E}[L] = \int \{y(x) - h(x)\}^2 p(x)dx + \underbrace{\iint \{h(x) - t\}^2 P(x,t)dxdt}$$

where

$$h(x) = \mathbb{E}[t|x] = \int tp(t|x)dt.$$

The second term of $E[L]$ corresponds to the noise inherent in the random variable t.

Q: What about the first term?

The Bias-Variance Decomposition$_{(2)}$

- Suppose we were given multiple data sets, each of size N. Any particular data set, D, will give a particular function $y(x; D)$. We then have

$$\{y(x;\mathcal{D} - h(x)\}^2$$

$$= \{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})] - h(x)\}^{\ 2}$$

$$= \{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})] - h(x)\}^2$$

$$+2\{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})] - h(x)\}$$

The Bias-Variance Decomposition$_{(3)}$

- Taking the expectation over D yields

$$\mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D}) - h(x)\}^2]$$

$$= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})] - h(x)\}^2}_{\text{(bias)}^2} + \underbrace{\mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}^2]}_{\text{variance}}$$

The Bias-Variance Decomposition$_{(4)}$

- Thus we can write

$$expected\ loss = (bias)^2 + variance + noise$$

where

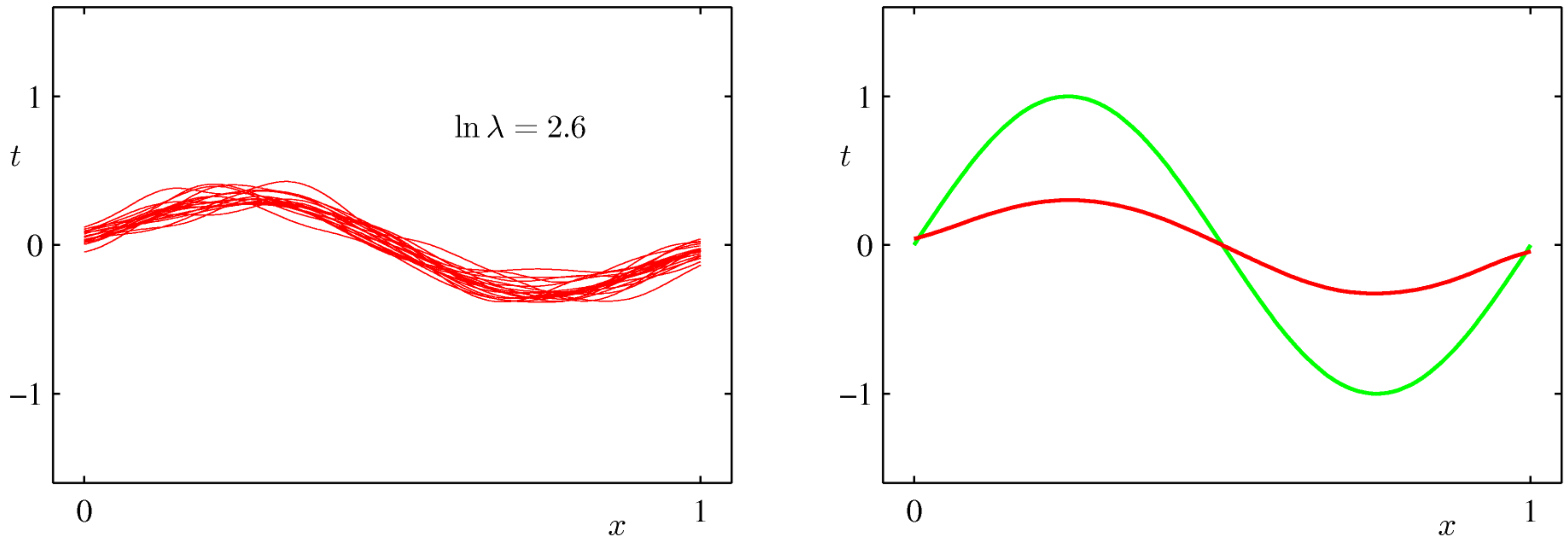
$$(bias)^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})] - h(x)\}^2 p(x) dx$$

$$variance = \int \mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}^2] p(x) dx$$

$$noise = \iint \{h(x) - t\}^2 p(x,t) dx dt$$

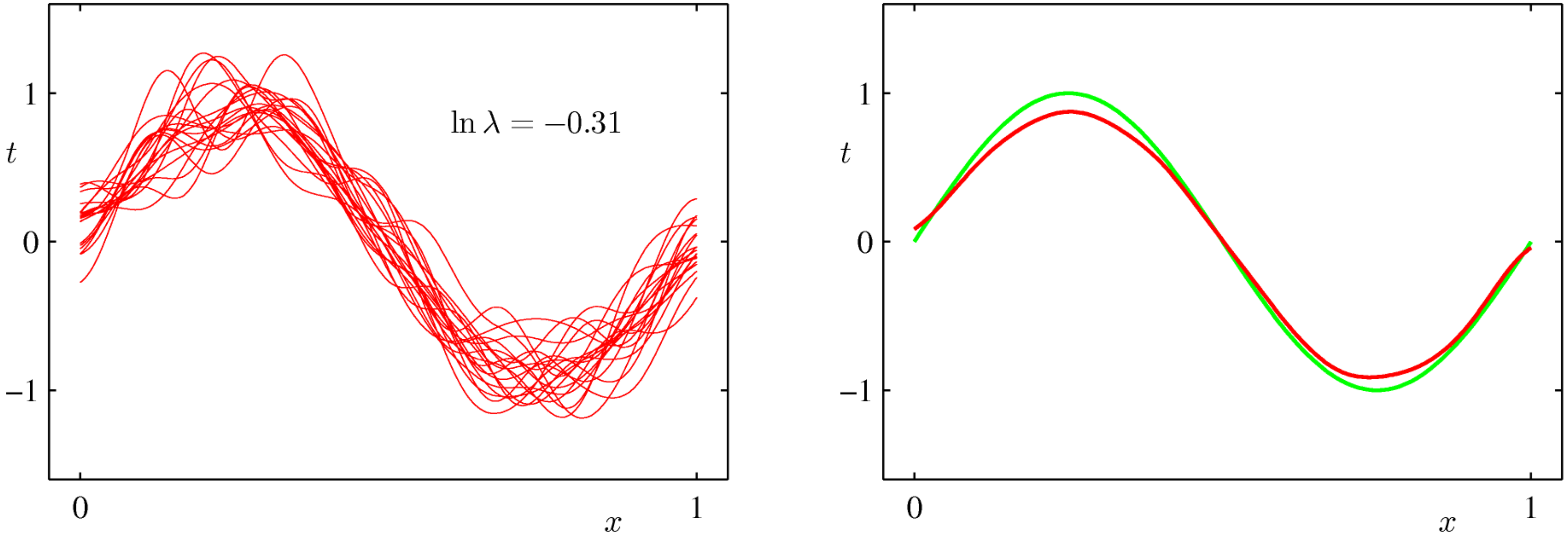
The Bias-Variance Decomposition(5)

- Example: 25 data sets from the sinusoidal, varying the degree of regularization,



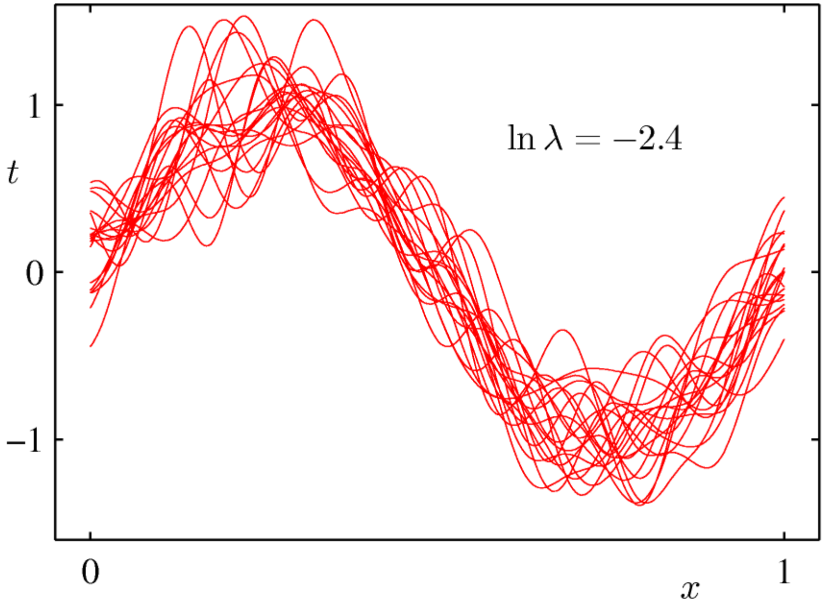
The Bias-Variance Decomposition(6)

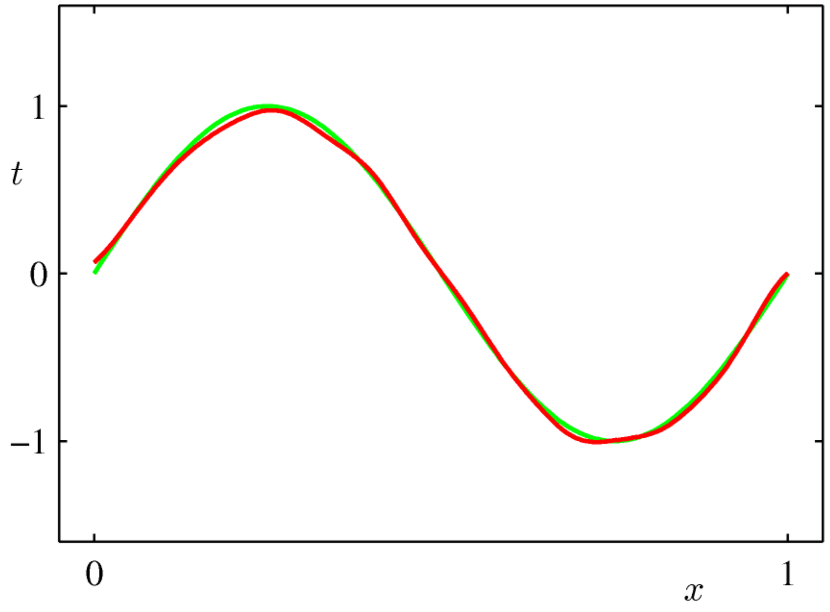
- Example: 25 data sets from the sinusoidal, varying the degree of regularization,



The Bias-Variance Decomposition(7)

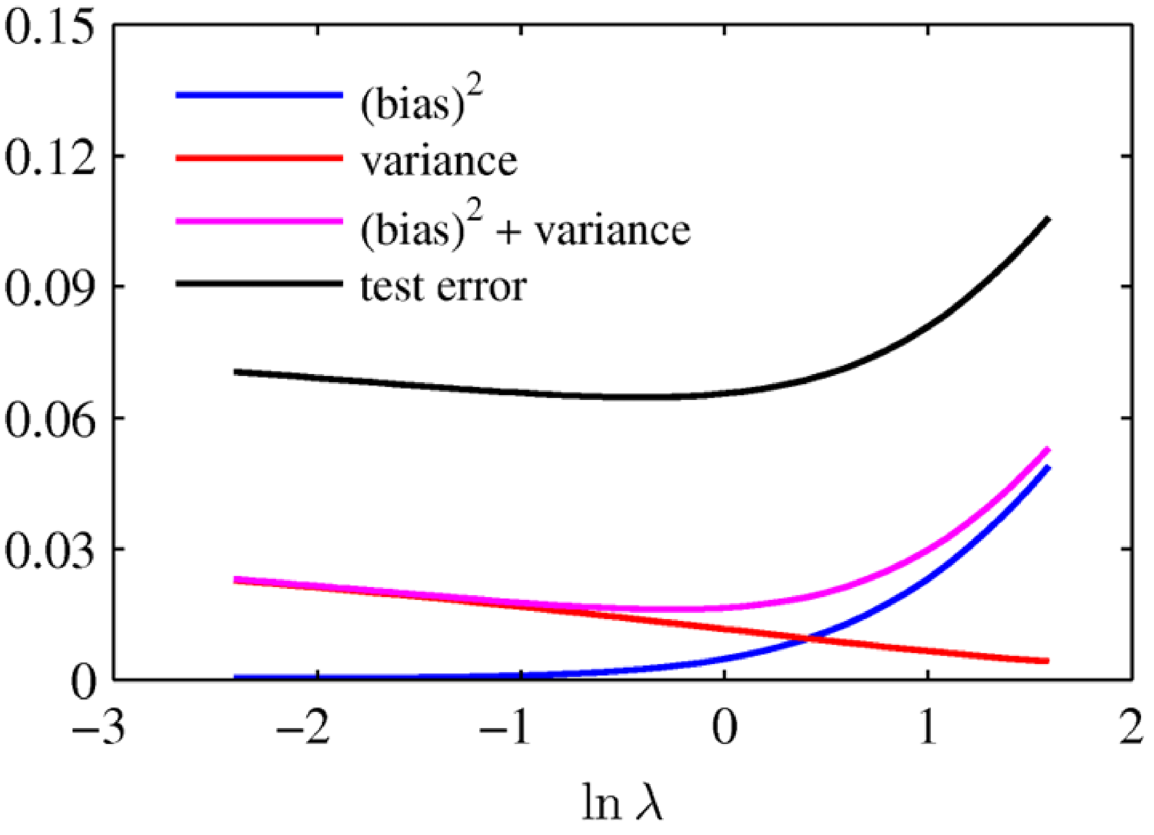
- Example: 25 data sets from the sinusoidal, varying the degree of regularization,





The Bias-Variance Trade-off

- From these plots, we note that an over-regularized model (large ¸) will have a high bias, while an under-regularized model (small ¸) will have a high variance.



Bayesian Linear Regression$_{(1)}$

- Define a conjugate prior over w

$$p(w) = \mathcal{N}(w|m_0, S_0)$$

Combining this with the likelihood function and using results for marginal and conditional Gaussian distributions, gives the posterior

where

$$p(w|t) = \mathcal{N}(w|m_N, S_N)$$

$$m_N = S_N(S_0^{-1}m_0 + \beta\boldsymbol{\Phi}^T\boldsymbol{t})$$

$$S_N^{-1} = S_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

Bayesian Linear Regression(2)

- A common choice for the prior is

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$$

for which

$$m_N = \beta \boldsymbol{S}_N \boldsymbol{\Phi}^T \boldsymbol{t}$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$$

Next we consider an example …

Predictive Distribution$_{(1)}$

- Predict t for new values of x by integrating over w:

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\mathbf{t}, \alpha, \beta) d\boldsymbol{w}$$

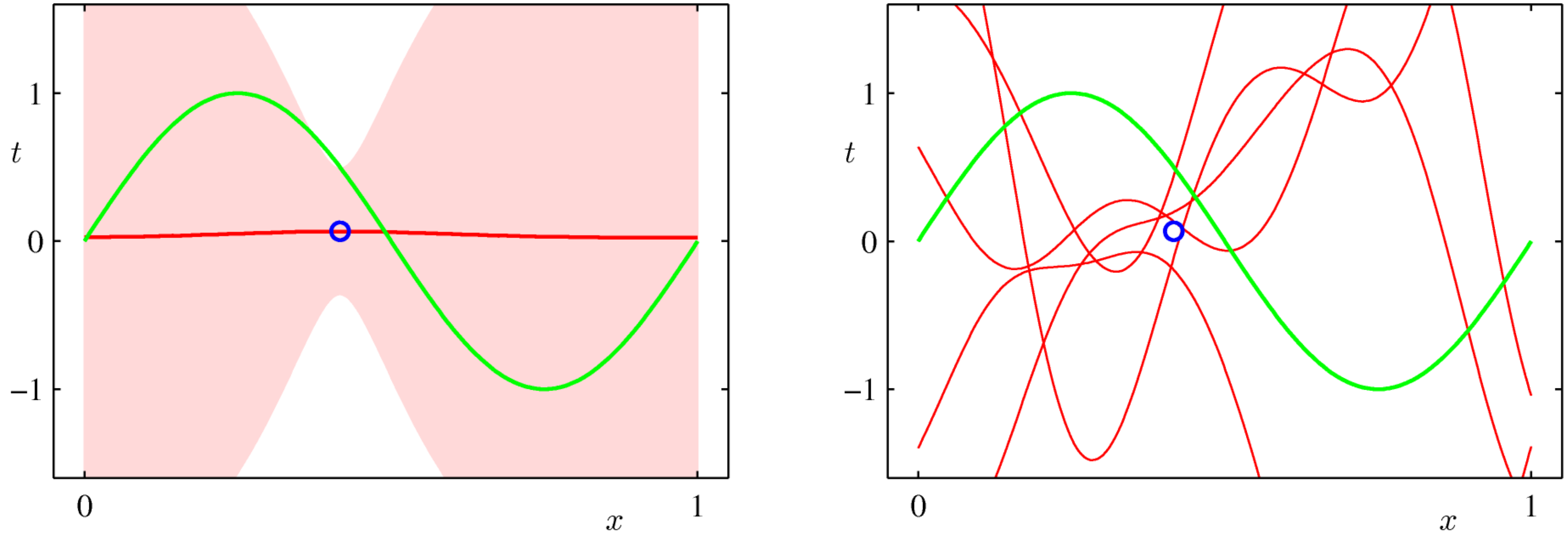
$$= \mathcal{N}(t|\boldsymbol{m}_N^T \phi(x), \sigma_N^2(x))$$

where

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$$

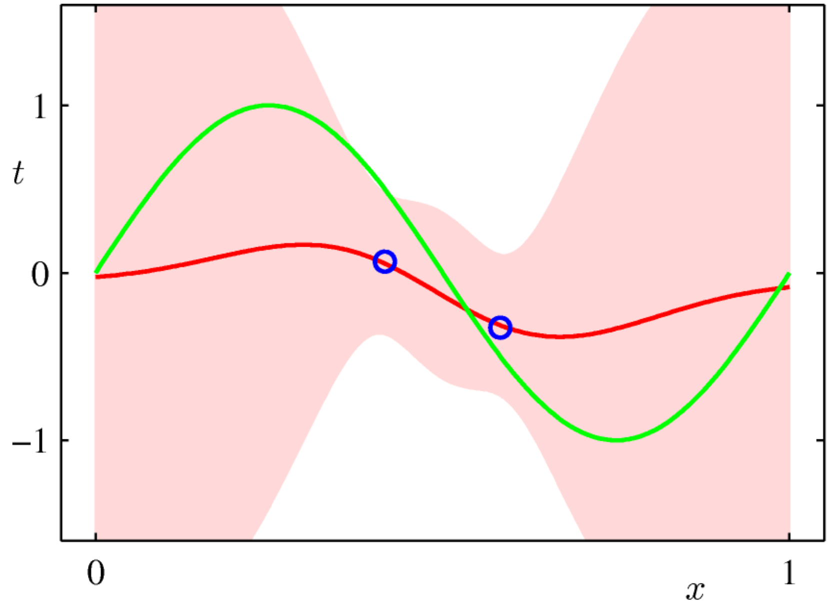
Predictive Distribution$_{(2)}$

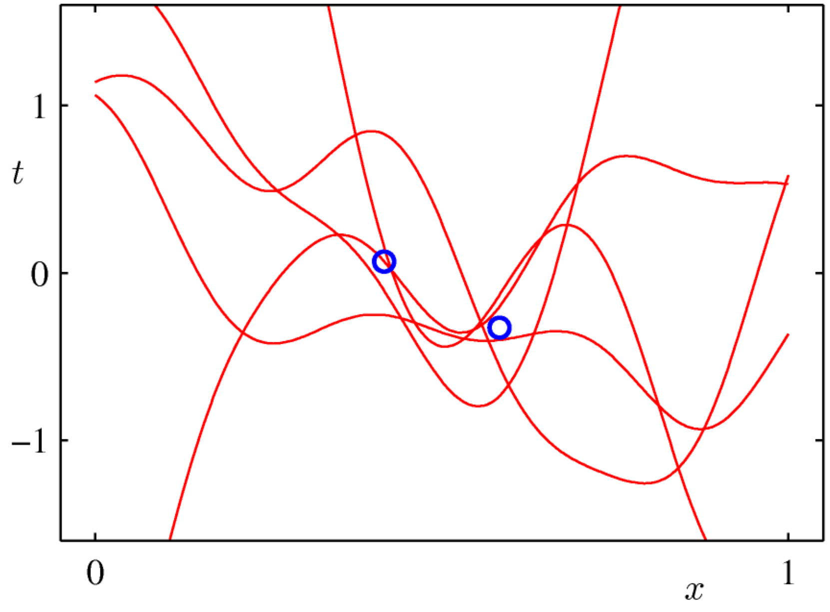
- Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point



Predictive Distribution$_{(3)}$

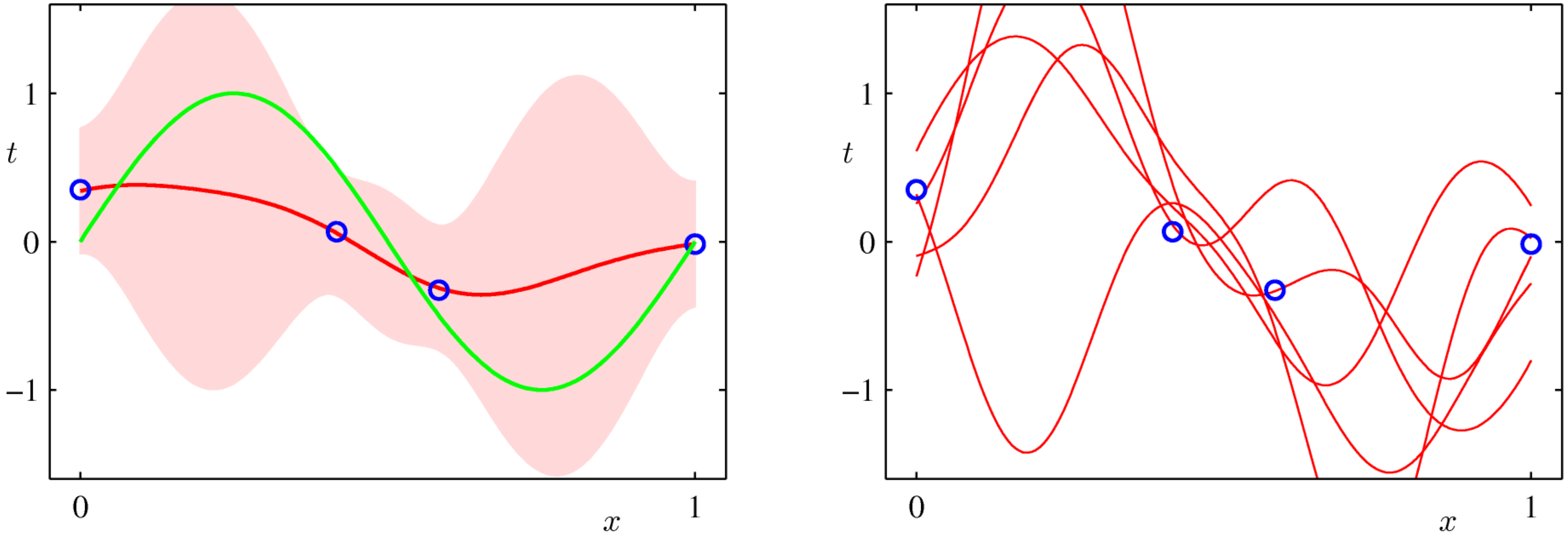
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data point





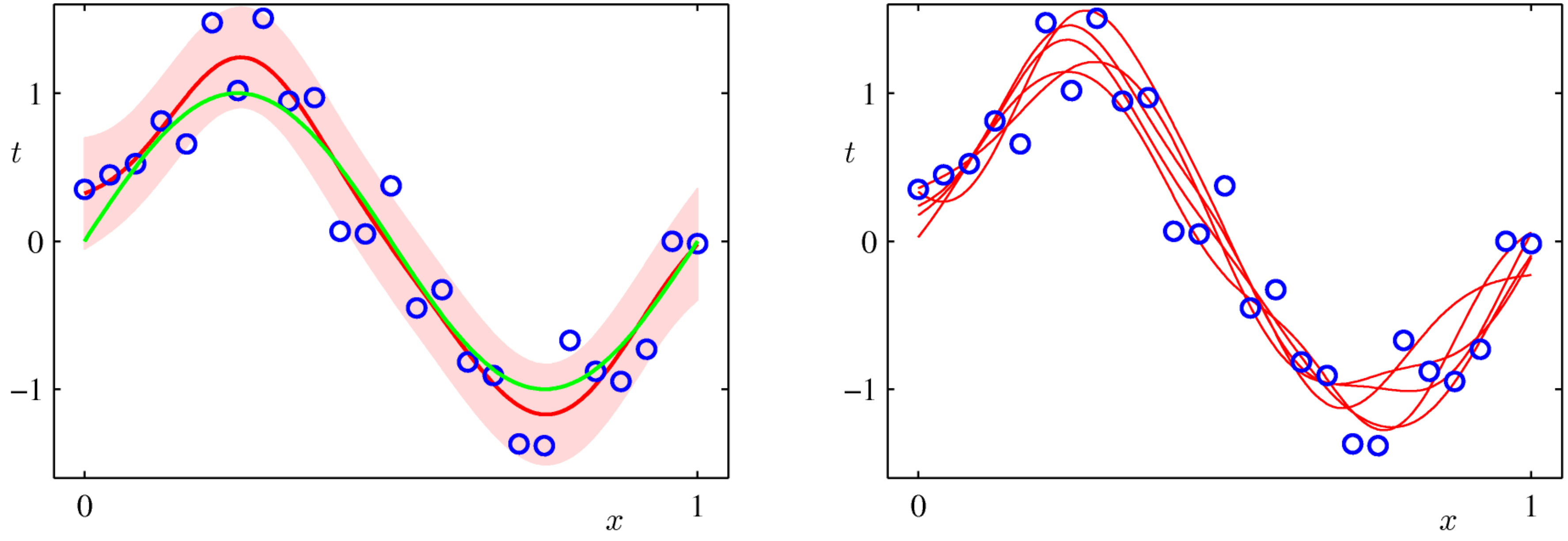
Predictive Distribution$_{(4)}$

- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data point



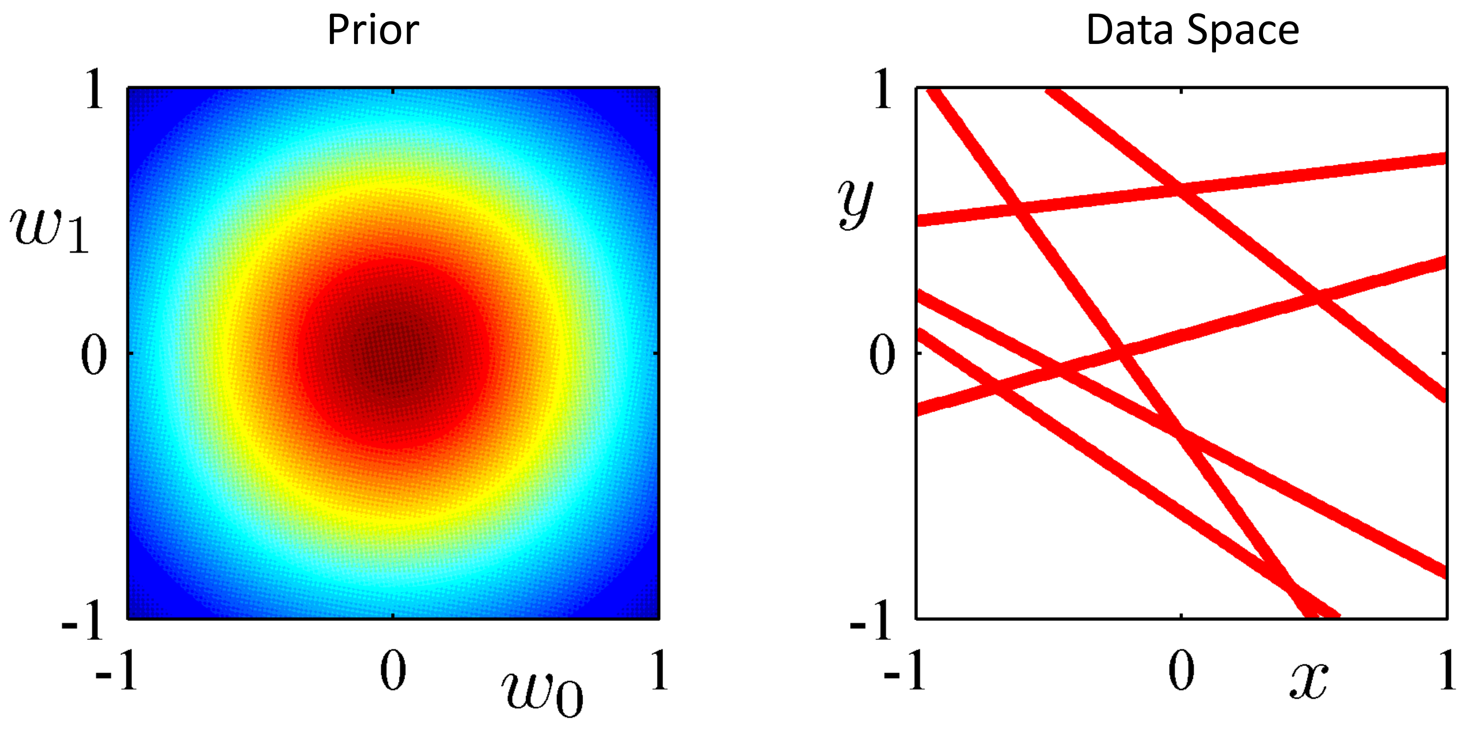
Predictive Distribution$_{(4)}$

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data point



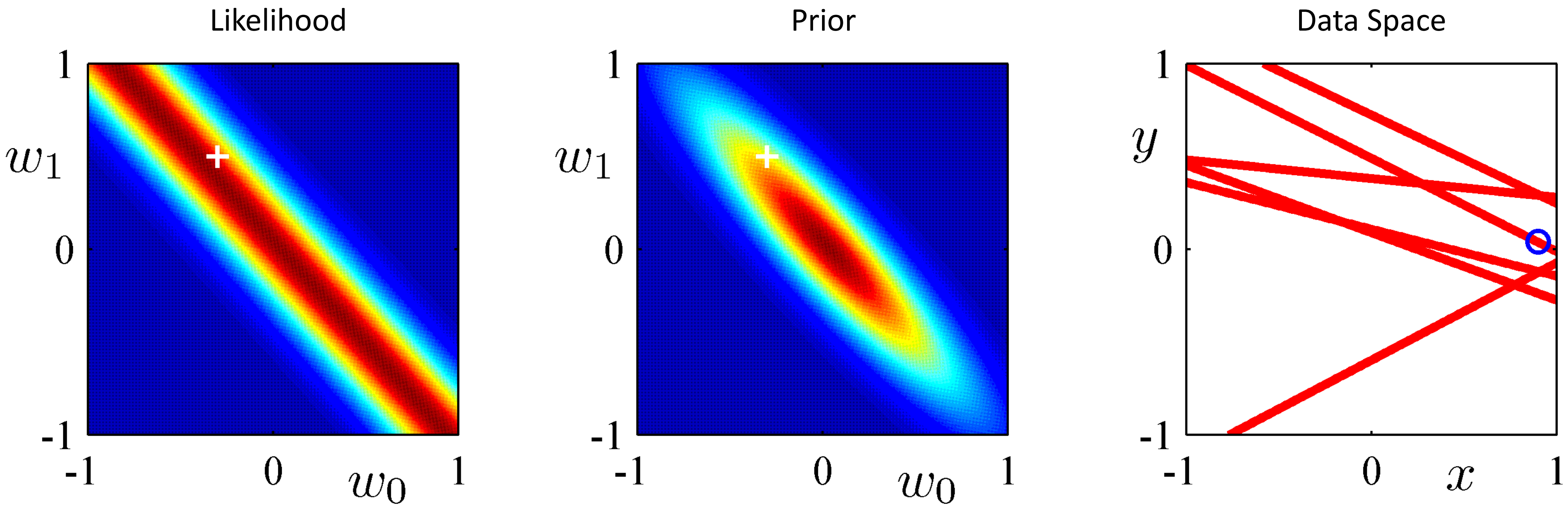
Bayesian Linear Regression$_{(3)}$

- 0 data points observed



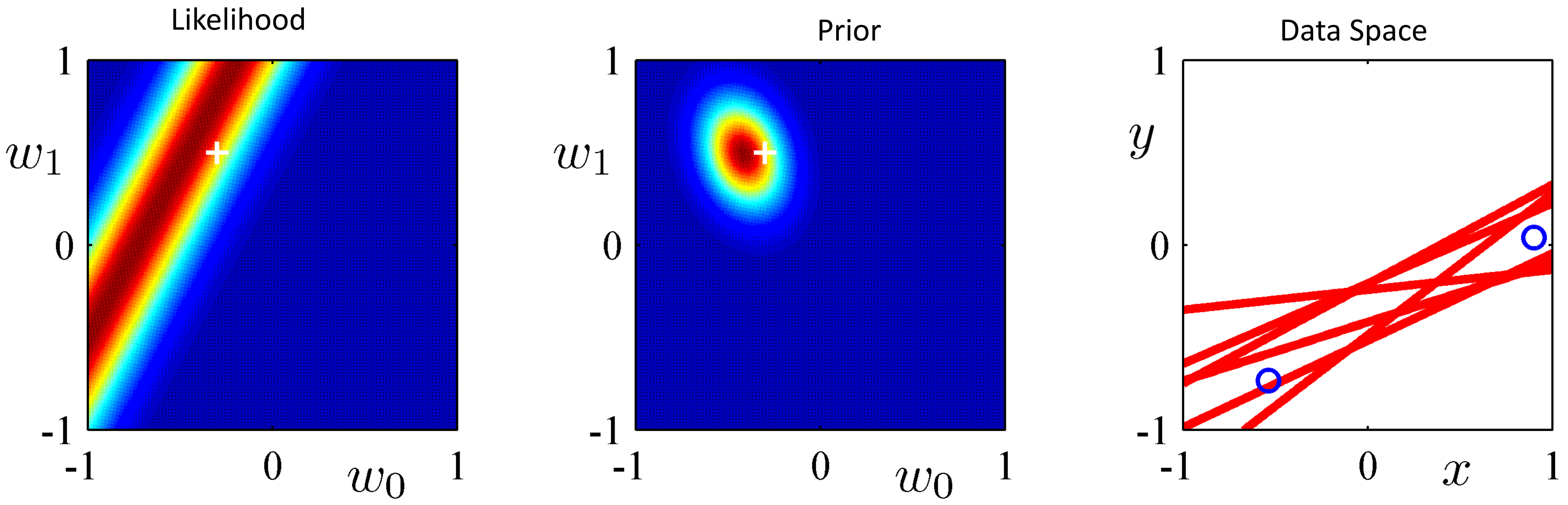
Bayesian Linear Regression$_{(4)}$

- 1 data points observed



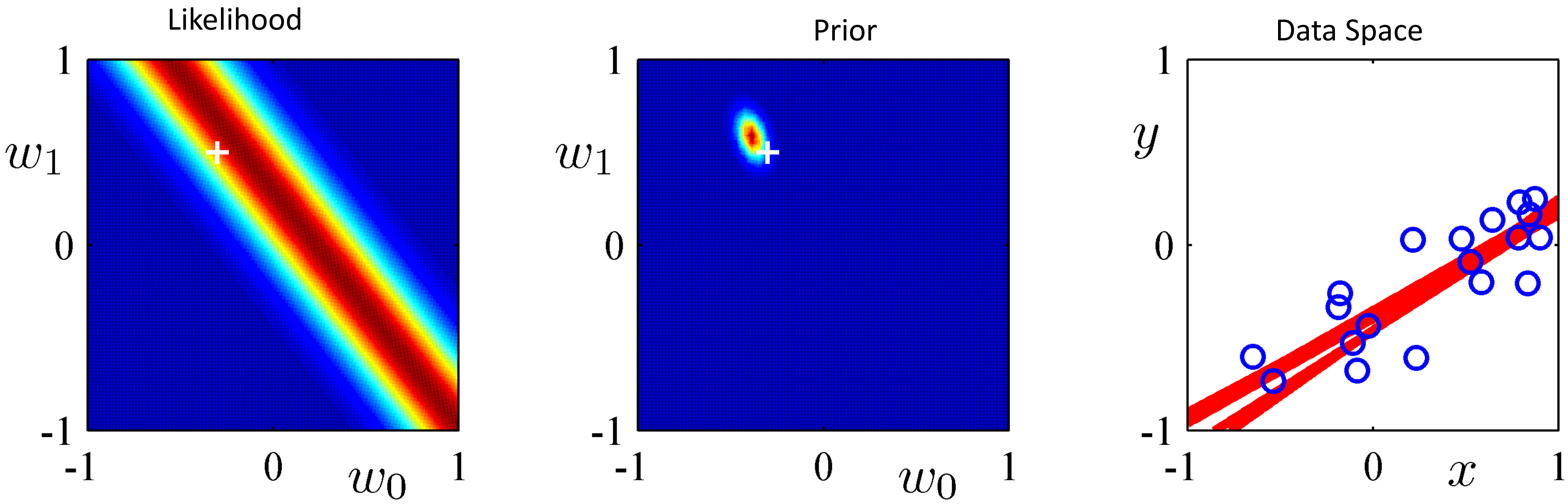
Bayesian Linear Regression$_{(5)}$

- 2 data points observed



Bayesian Linear Regression$_{(6)}$

- 20 data points observed



Equivalent Kernel (1)

- The predictive mean can be written

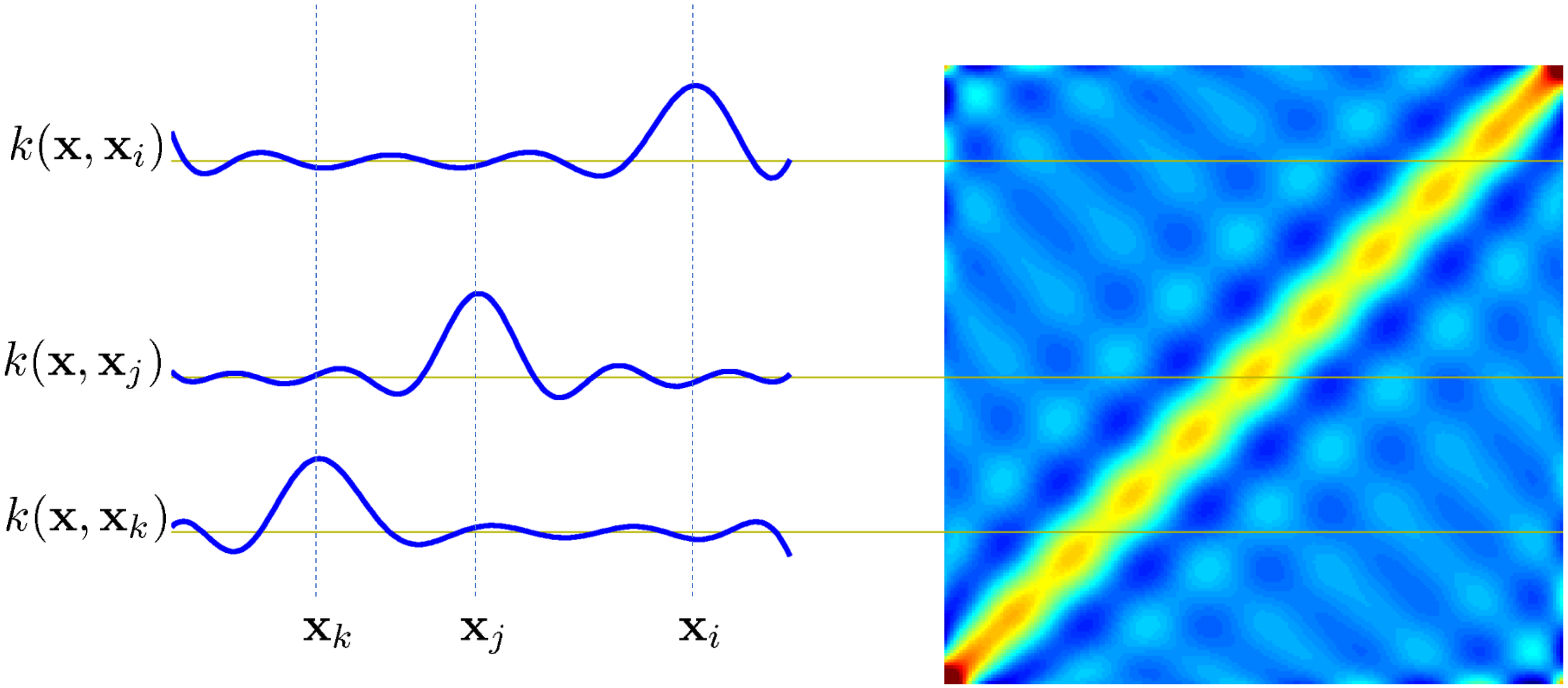
$$
\begin{aligned}
y(\mathbf{x}, \mathbf{m}_N) &= \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\
&= \sum_{n=1}^{N} \underbrace{\beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_n)} t_n \\
&= \sum_{n=1}^{N} \underbrace{k(\mathbf{x}, \mathbf{x}_n)} t_n .
\end{aligned}
$$

> *Equivalent kernel* or *smoother matrix*.

- This is a weighted sum of the training data target values, t_n.

Equivalent Kernel (2)

- Weight of t_n depends on distance between x and x_n; nearby x_n carry more weight.



The Evidence Approximation (1)

- The fully Bayesian predictive distribution is given by

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})\,\mathrm{d}\mathbf{w}\,\mathrm{d}\alpha\,\mathrm{d}\beta$$

- but this integral is intractable. Approximate with

$$p(t|\mathbf{t}) \simeq p\left(t|\mathbf{t},\widehat{\alpha},\widehat{\beta}\right) = \int p\left(t|\mathbf{w},\widehat{\beta}\right)p\left(\mathbf{w}|\mathbf{t},\widehat{\alpha},\widehat{\beta}\right)\,\mathrm{d}\mathbf{w}$$

•where $\left(\widehat{\alpha},\widehat{\beta}\right)$ is the mode of $p(\alpha,\beta|\mathbf{t})$, which is assumed to be sharply peaked; a.k.a. *empirical Bayes, type II* or *generalized maximum likelihood,* or *evidence approximation*.

The Evidence Approximation (2)

- From Bayes' theorem we have

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta)$$

- and if we assume to $p(\alpha, \beta)$ be flat we see that

$$\begin{aligned} p(\alpha, \beta|\mathbf{t}) &\propto p(\mathbf{t}|\alpha, \beta) \\ &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \end{aligned}$$

- General results for Gaussian integrals give

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) + \frac{1}{2}\ln|\mathbf{S}_N| - \frac{N}{2}\ln(2\pi).$$

Maximizing the Evidence Function (1)

- To maximise $p(t|\alpha, \beta)$ w.r.t. α and β, we define the eigenvector equation

$$\left(\beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right) \mathbf{u}_i = \lambda_i \mathbf{u}_i .$$

- Thus

$$\mathbf{A} = \mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

- has eigenvalues $\lambda_i + \alpha$.

- We can now differentiate $p(t|\alpha, \beta)$ w.r.t. α and β and set the results to zero, to get

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}} \mathbf{m}_N}$$

$$\frac{1}{\beta} = \frac{1}{N-\gamma} \sum_{n=1}^{N} \left\{ t_n - \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\}^2$$

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

N.B. γ depends on both α and β

Any questions?