

CS489/698

Lecture 5: Jan 17, 2018

Linear Regression by Maximum
Likelihood, Maximum A Posteriori and
Bayesian Learning

[B] Sections 3.1 – 3.3, [M] Chapt. 7

Noisy Linear Regression

- Assume y is obtained from $\mathbf{x}$ by a deterministic function f that has been perturbed (i.e., noisy measurement)

$$y = f(\bar{\mathbf{x}}) + \epsilon$$

$\downarrow \qquad \searrow$

$$\mathbf{w}^T \bar{\mathbf{x}} \qquad N(0, \sigma^2)$$

- Gaussian noise:

$$\begin{aligned} \Pr(\mathbf{y}|\bar{\mathbf{X}}, \mathbf{w}, \sigma) &= N(\mathbf{y}|\mathbf{w}^T \bar{\mathbf{X}}, \sigma^2) \\ &= \prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2}} \end{aligned}$$

Maximum Likelihood

- Possible objective: find best $\mathbf{w}^*$ by maximizing the likelihood of the data

$$\begin{aligned}\mathbf{w}^* &= \operatorname{argmax}_{\mathbf{w}} \Pr(\mathbf{y}|\bar{\mathbf{X}}, \mathbf{w}, \sigma) \\ &= \operatorname{argmax}_{\mathbf{w}} \prod_n e^{-\frac{(y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2}} \\ &= \operatorname{argmax}_{\mathbf{w}} \sum_n -\frac{(y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2} \\ &= \operatorname{argmin}_{\mathbf{w}} \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2\end{aligned}$$

- We arrive at the original least square problem!

Maximum A Posteriori

- Alternative objective: find $\mathbf{w}^*$ with highest posterior probability
- Consider Gaussian prior: $\Pr(\mathbf{w}) = N(\mathbf{0}, \Sigma)$
- Posterior:

$$\Pr(\mathbf{w}|\mathbf{X}, \mathbf{y}) \propto \Pr(\mathbf{w}) \Pr(\mathbf{y}|\mathbf{X}, \mathbf{w})$$

$$= k e^{-\frac{\mathbf{w}^T \Sigma^{-1} \mathbf{w}}{2}} e^{-\frac{\sum_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2}{2\sigma^2}}$$

Maximum A Posteriori

- Optimization:

$$\begin{aligned}\mathbf{w}^* &= \operatorname{argmax}_{\mathbf{w}} \Pr(\mathbf{w}|\bar{\mathbf{X}}, \mathbf{y}) \\ &= \operatorname{argmax}_{\mathbf{w}} -\sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2 - \mathbf{w}^T \Sigma^{-1} \mathbf{w} \\ &= \operatorname{argmin}_{\mathbf{w}} \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2 + \mathbf{w}^T \Sigma^{-1} \mathbf{w}\end{aligned}$$

- Let $\Sigma^{-1} = \lambda \mathbf{I}$ then

$$= \operatorname{argmin}_{\mathbf{w}} \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2 + \lambda \|\mathbf{w}\|_2^2$$

- We arrive at the original **regularized** least square problem!

Expected Squared Loss

- Even though we use a statistical framework, it is interesting to evaluate the expected squared loss

$$\begin{aligned} E[L] &= \int_{\mathbf{x}, y} \Pr(\mathbf{x}, y) (y - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x} dy \\ &= \int_{\mathbf{x}, y} \Pr(\mathbf{x}, y) (y - f(\mathbf{x}) + f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x} dy \\ &= \int_{\mathbf{x}, y} \Pr(\mathbf{x}, y) \left[(y - f(\mathbf{x}))^2 + \underbrace{2(y - f(\mathbf{x}))(f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})}_{\text{Expectation with respect to } y \text{ is } 0} + (f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 \right] d\mathbf{x} dy \end{aligned}$$

Expectation with respect to y is 0

$$E[L] = \underbrace{\int_{\mathbf{x}, y} \Pr(\mathbf{x}, y) (y - f(\mathbf{x}))^2 d\mathbf{x} dy}_{\text{noise (constant)}} + \underbrace{\int_{\mathbf{x}} \Pr(\mathbf{x}) (f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x}}_{\text{error (depends on } \mathbf{w} \text{)}}$$

Expected Squared Loss

- Let's focus on the error part, which depends on $\mathbf{w}$

$$E_{\mathbf{x}}[(f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2] = \int_{\mathbf{x}} \text{Pr}(\mathbf{x}) (f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x}$$

- But the choice of $\mathbf{w}$ depends on the dataset S
- Instead consider expectation with respect to S

$$E_S[(f(\mathbf{x}) - \mathbf{w}_S^T \bar{\mathbf{x}})^2]$$

where $\mathbf{w}_S$ is the weight vector obtained based on S

Bias-Variance Decomposition

- Decompose squared loss

$$\begin{aligned} E_S \left[(f(\mathbf{x}) - \mathbf{w}_S^T \bar{\mathbf{x}})^2 \right] &= E_S \left[f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] + E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}} \right]^2 \\ &= E_S \left[(f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}])^2 \right. \\ &\quad + 2(f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}]) (E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}}) \\ &\quad \left. + (E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}})^2 \right] \underbrace{\hspace{10em}}_{\text{Expectation is 0}} \\ &= \underbrace{(f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}])^2}_{\text{bias}^2} + \underbrace{E_S \left[(E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}})^2 \right]}_{\text{variance}} \end{aligned}$$

Bias-Variance Decomposition

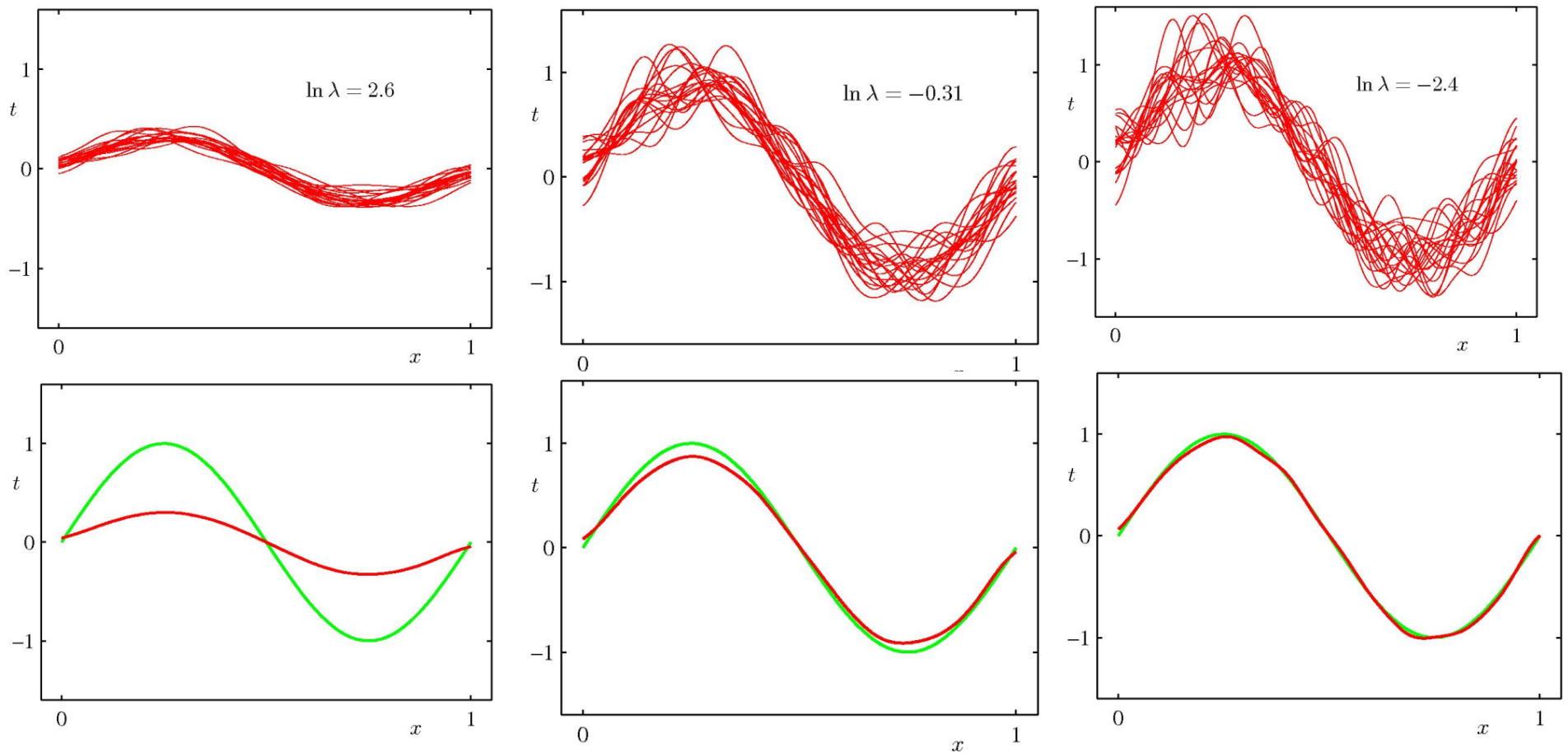
- Hence:

$$E[loss] = (bias)^2 + variance + noise$$

- Picture:

Bias-Variance Decomposition

- Example



Bayesian Linear Regression

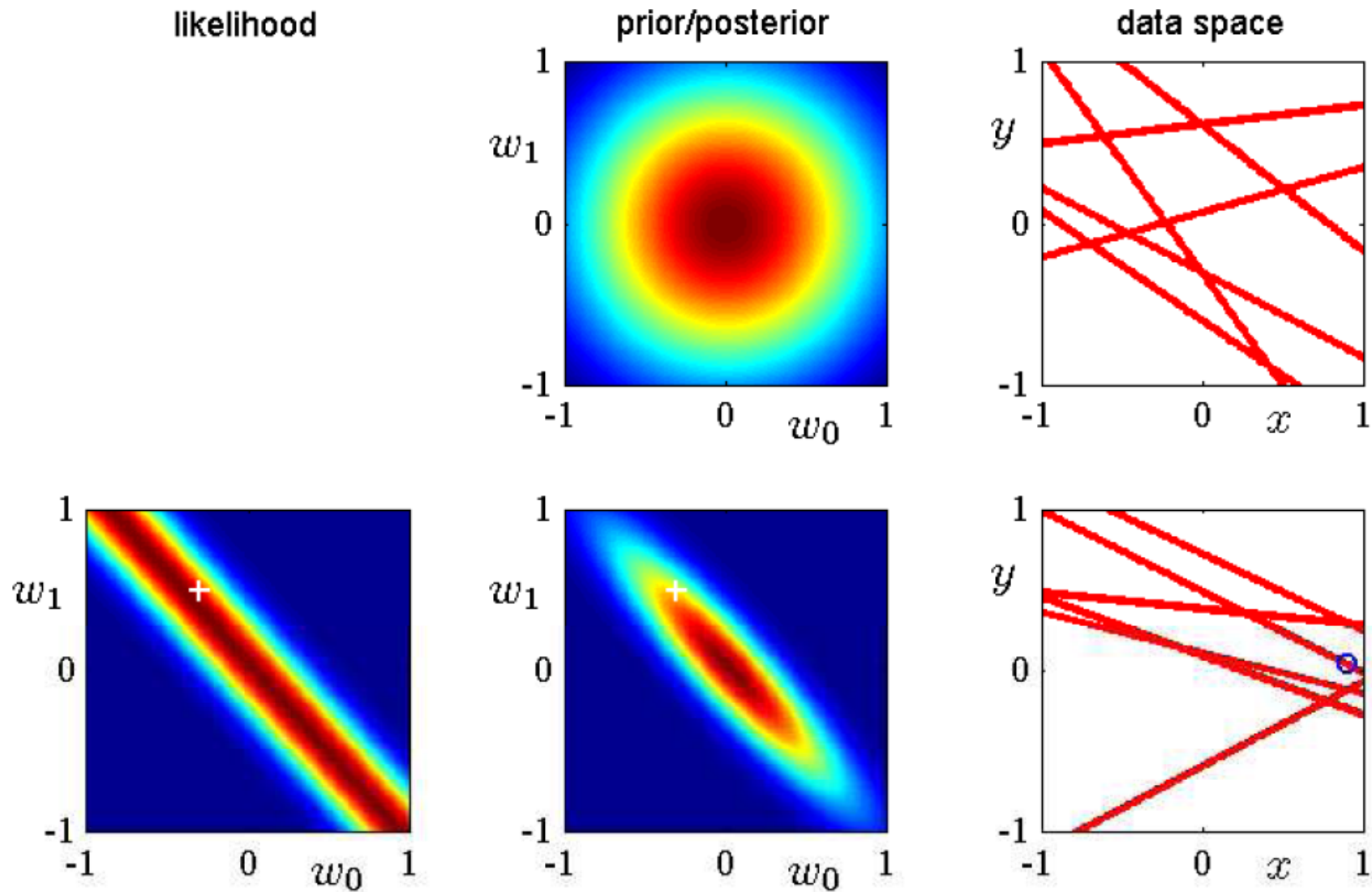
- We don't know if $\mathbf{w}^*$ is the true underlying $\mathbf{w}$
- Instead of making predictions according to $\mathbf{w}^*$, compute the weighted average prediction according to $\Pr(\mathbf{w}|\bar{\mathbf{X}}, \mathbf{y})$

$$\begin{aligned}\Pr(\mathbf{w}|\bar{\mathbf{X}}, \mathbf{y}) &= k e^{-\frac{\mathbf{w}^T \Sigma^{-1} \mathbf{w}}{2}} e^{-\frac{\sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2}} \\ &= k e^{-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A}(\mathbf{w} - \bar{\mathbf{w}})} = N(\bar{\mathbf{w}}, \mathbf{A}^{-1})\end{aligned}$$

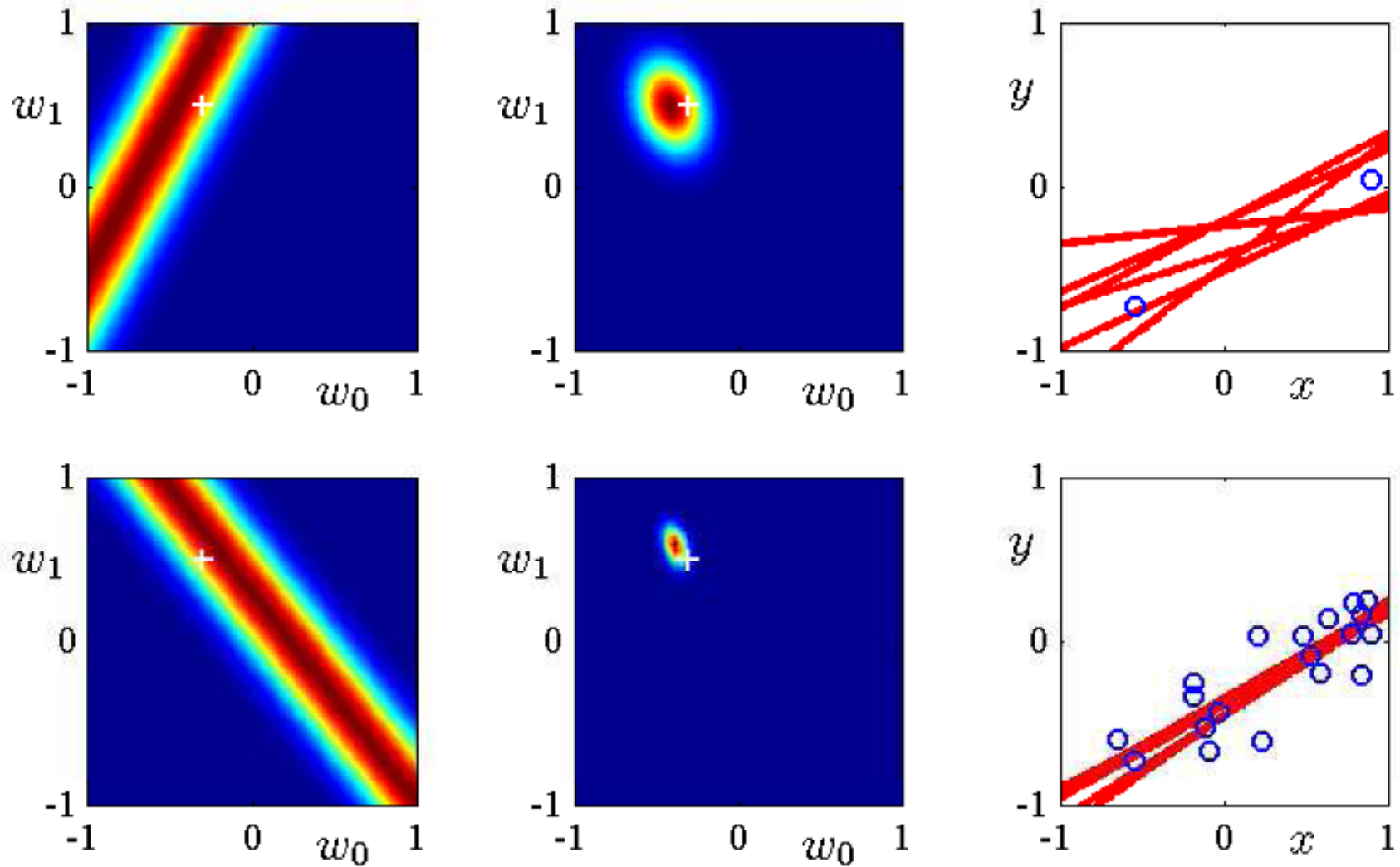
where $\bar{\mathbf{w}} = \sigma^{-2} \mathbf{A}^{-1} \bar{\mathbf{X}} \mathbf{y}$

$$\mathbf{A} = \sigma^{-2} \bar{\mathbf{X}} \bar{\mathbf{X}}^T + \Sigma^{-1}$$

Bayesian Learning



Bayesian Learning



Bayesian Prediction

- Let $\mathbf{x}_*$ be the input for which we want a prediction and y_* be the corresponding prediction

$$\begin{aligned}\Pr(y_* | \bar{\mathbf{x}}_*, \bar{\mathbf{X}}, \mathbf{y}) &= \int_{\mathbf{w}} \Pr(y_* | \bar{\mathbf{x}}_*, \mathbf{w}) \Pr(\mathbf{w} | \bar{\mathbf{X}}, \mathbf{y}) d\mathbf{w} \\ &= k \int_{\mathbf{w}} e^{-\frac{(y_* - \bar{\mathbf{x}}_*^T \mathbf{w})^2}{2\sigma^2}} e^{-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A}(\mathbf{w} - \bar{\mathbf{w}})} d\mathbf{w} \\ &= N(\bar{\mathbf{x}}_*^T \mathbf{A}^{-1} \bar{\mathbf{X}} \mathbf{y}, \bar{\mathbf{x}}_*^T \mathbf{A}^{-1} \bar{\mathbf{x}}_*)\end{aligned}$$