University of Amsterdam
Informatics Institute

AMLAB
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 4.4 - Supervised Learning
Bayesian Linear Regression - Sequential Bayesian Learning

Erik Bekkers

(Bishop 3.3.1)

Image credit: Kirillm | Getty Images

Slide credits: Patrick Forré and Rianne van den Berg

Example: Sequential Bayesian Learning

Data: sequences of input *x*, target *t*

Synthetic data generated by $x \sim \mathcal{U}(x \mid -1,1)$ $t = f(x, \mathbf{a}) + \varepsilon$

$f(x, \mathbf{a}) =$ $a_0 + a_1 x$ $\varepsilon \sim \mathcal{N}(0, 0.2^2)$

$a_0 = -0.3 \quad a_1 = 0.5$

Target modeling: $p(t' \mid x', \mathbf{w}, \beta) = \mathcal{N}(t' \mid y(x', \mathbf{w}), \beta^{-1})$, $\beta^{-1} =$ 0.2^2

Linear model: $y(x, \mathbf{w}) =$ $w_0 + w_1 x$

Prior: $p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$ $\alpha = 2$

e.g. $N=2$ When data arrives sequentially: posterior after N-1 datapoints is prior for arrival of N-th datapoint!

posterior after x_1 acts as a prior for the post. x_2

$$p(\underline{w} \mid x_1, x_2) = \frac{p(x_2 \mid \underline{w}) \, p(x_1 \mid \underline{w}) \cdot p(\underline{w}, \alpha)}{p(x_2) \, p(x_1)} = \frac{p(x_2 \mid \underline{w}) \, \boxed{p(\underline{w}, x_1)}}{p(x_2)}$$

Example: Sequential Bayesian Learning

- Data generated by $t = a_0 + a_1 x + \varepsilon$

  $a_0 = -0.3 \quad a_1 = 0.5$

sample w

- Prior

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

- Sample 1 datapoint

  x_1, t_1

- Likelihood

  $p(t_1|x_1, \mathbf{w}, \beta) = $

  $\mathcal{N}(t_1 | w_0 + w_1 x_1, \beta^{-1})$

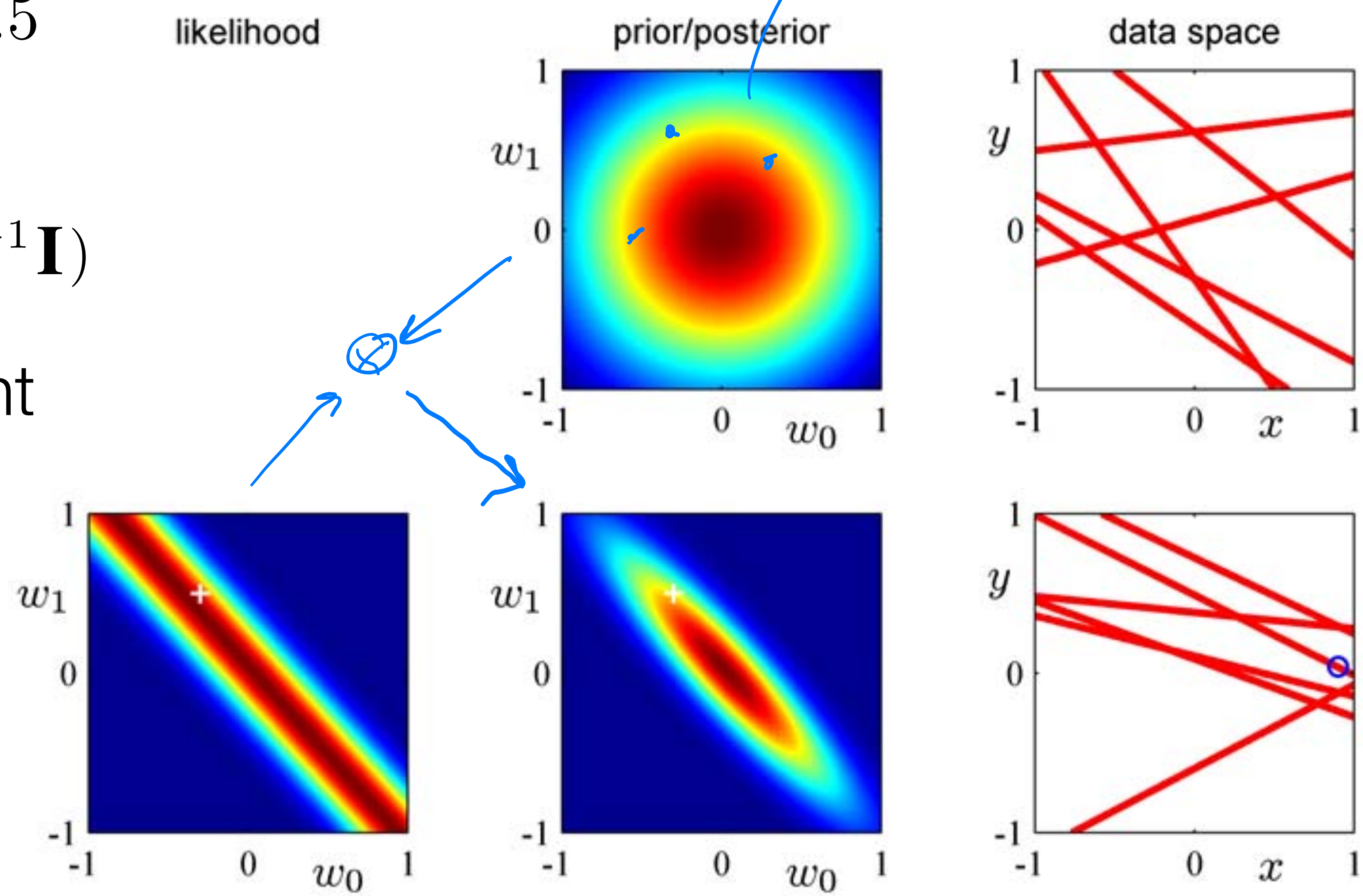

Figure: Sequential Bayesian learning (Bishop 3.7)

- Posterior

  $p(\mathbf{w}|x_1, t_1, \alpha, \beta) \propto$ $p(t_1 | x_1, \underline{w}, \beta)\, p(\underline{w}, \alpha)$

Example: Sequential Bayesian Learning

- Sample second datapoint:

  x_2, t_2

- Posterior ➡ prior :

- Likelihood

  $$p(t_2|x_2, \mathbf{w}, \beta)$$

- Posterior

  $$p(\mathbf{w}|(x_1, t_1), (x_2, t_2), \alpha, \beta) \propto p(t_2|x_2, \underline{w}, \beta) \cdot p(\underline{w}|(x_1, t_1), \alpha, \beta)$$

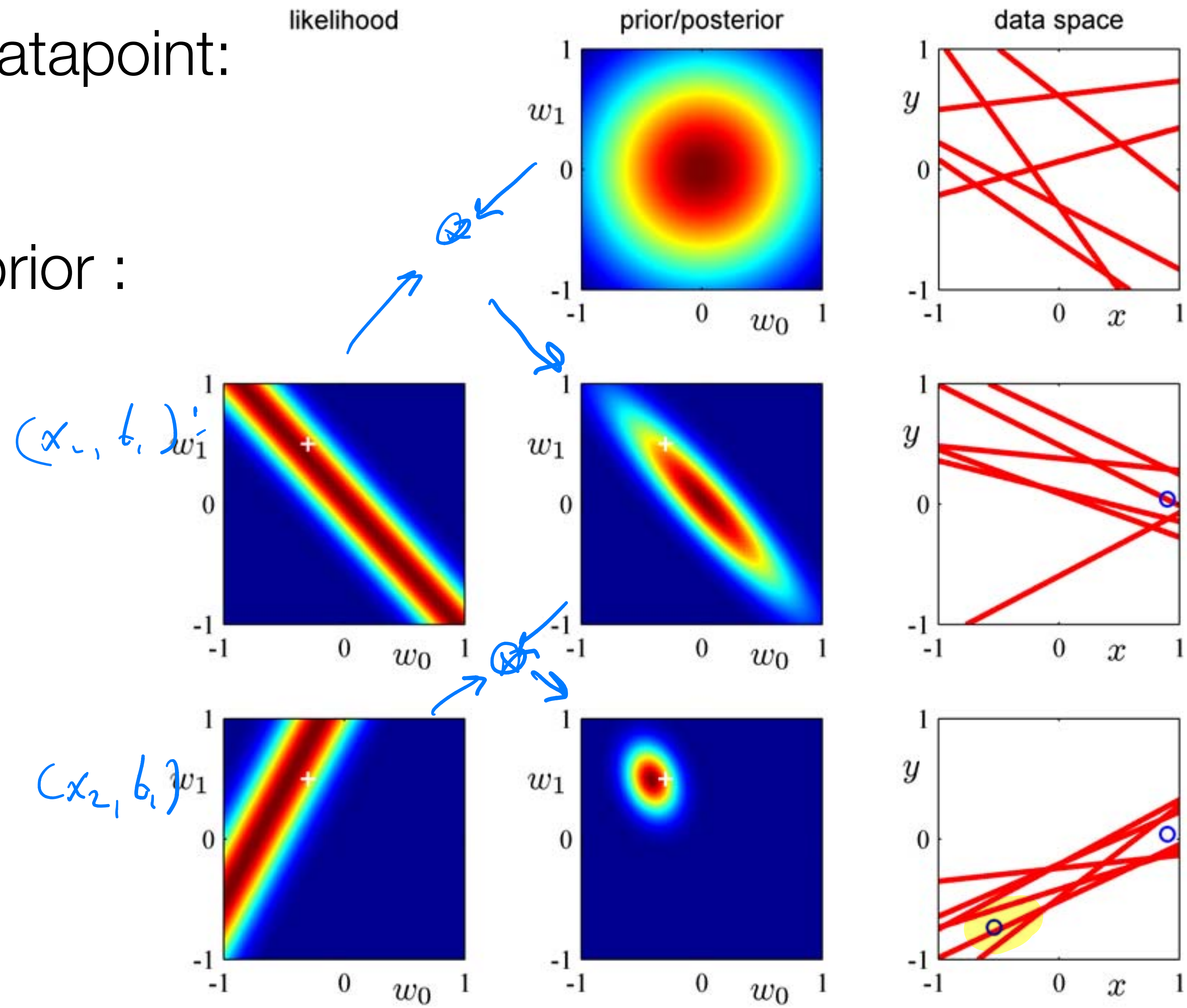

Figure: Sequential Bayesian learning (Bishop 3.7)

Example: Sequential Bayesian Learning

- After 19 datapoints

$(x_1, t_1) \; --- \; (x_{19}, t_{19})$

- Prior

$$p(\mathbf{w}|\{(x_n, t_n)\}_{n=1}^{19}, \alpha, \beta)$$

- Likelihood

$$p(t_{20}|x_{20}, \mathbf{w}, \beta)$$

- Posterior

$$p(\mathbf{w}|\{(x_n, t_n)\}_{n=1}^{20}, \alpha, \beta) \propto$$

- Much sharper posterior!

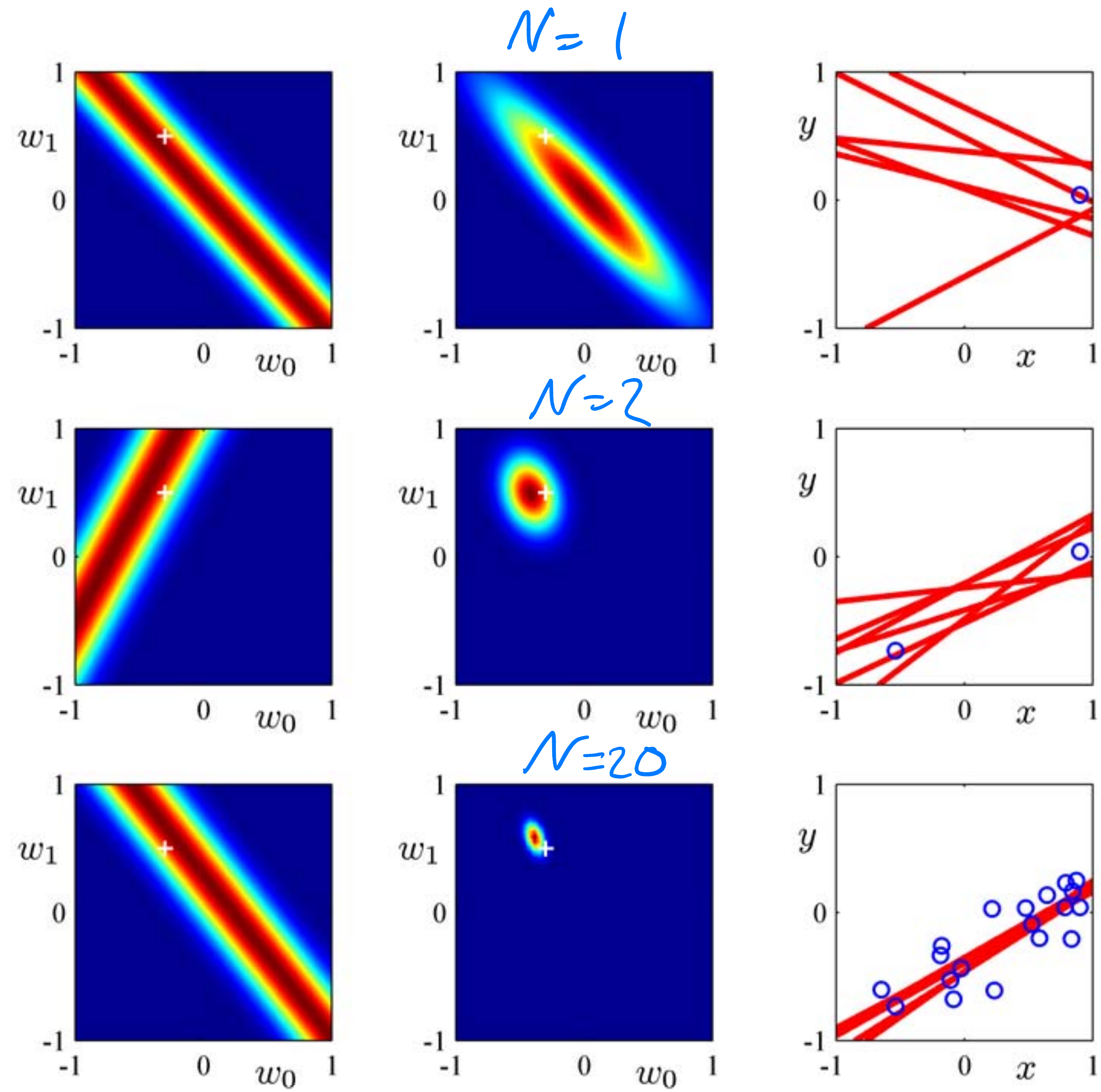

Figure: Sequential Bayesian learning (Bishop 3.7)

Infinite Data in Bayesian Linear Regression

- Poster distribution after observing N data points:

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$$

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{S}_N^{-1} = \alpha \mathbb{1} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$$

- After an infinite amount of data :

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \dots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \dots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

$$\lim_{N\to\infty} S_N = 0$$ (zero matrix)

$$\lim_{N\to\infty} \left[\mathbf{\Phi}^T \mathbf{\Phi}\right]_{ij} = \lim_{N\to\infty} \propto N$$

$$\lim_{N\to\infty} \mathbf{m}_N = \lim_{N\to\infty} \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t} = \lim_{N\to\infty} \beta(\cancel{\alpha I} + \beta \Phi^T \Phi)^{-1} \Phi^T t$$

$$= \lim_{N\to\infty} \underbrace{(\Phi^T \Phi)^{-1} \Phi^T t}_{w_{ML}}$$

Bayesian, MAP, ML all agree at $N \to \infty$