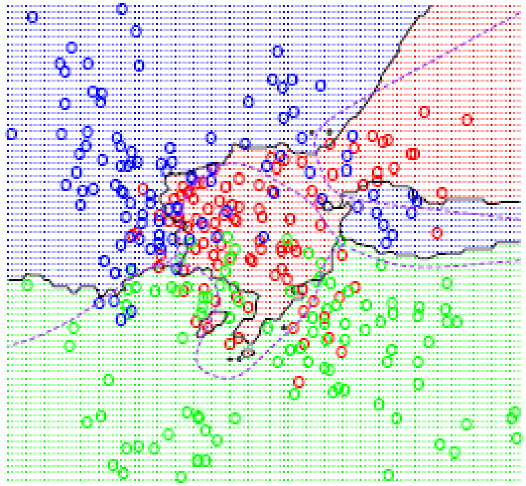
Machine Learning

Lecture 22: The Bayesian approach to machine learning

The lectures are mainly offered on white board accompanied by some slides.

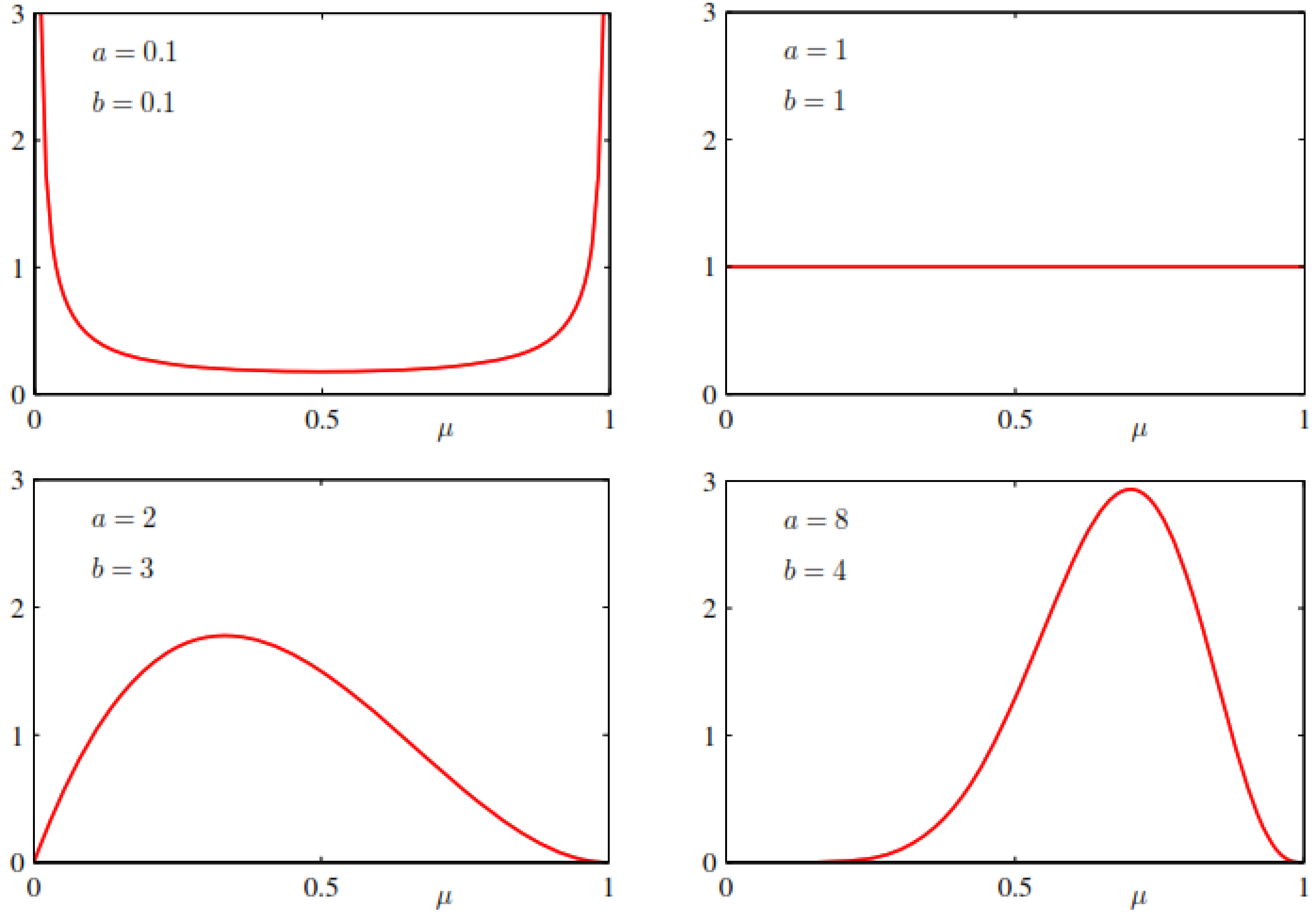


Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

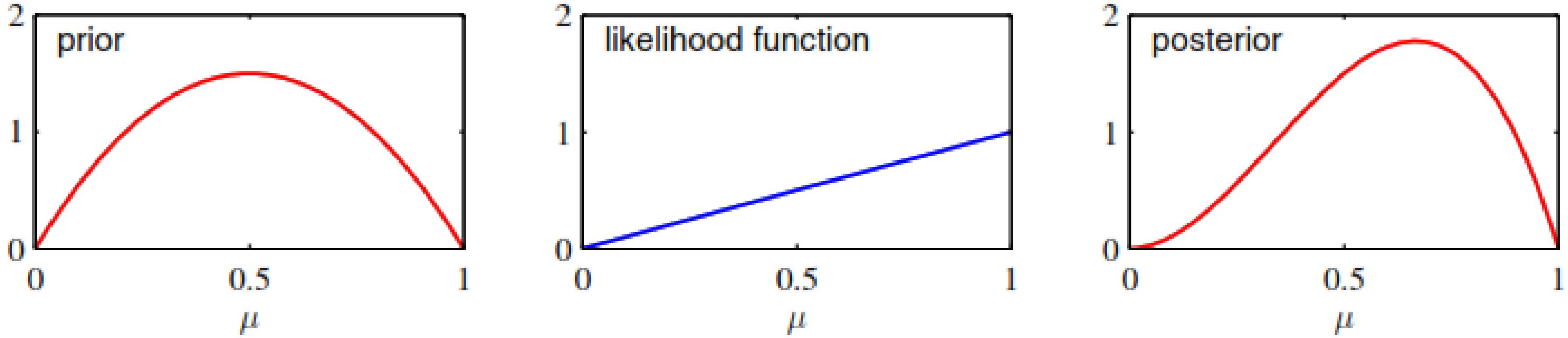
Azar 30, 1398

University of TEHRAN

Prior distributions

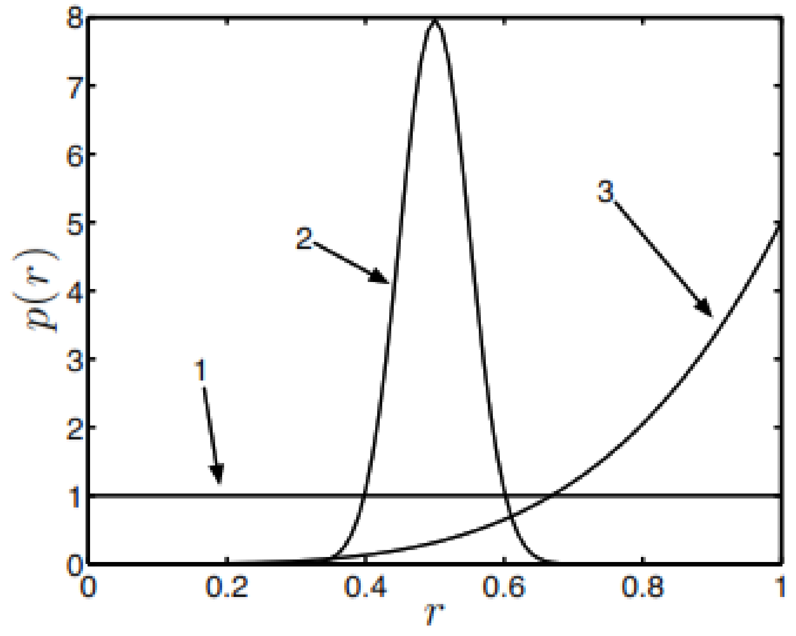


Prior, likelihood, posterior



A coin game [from FCML]

- Definition and notation:
  - r: the probability of head
  - The likelihood: $P(y) = \binom{N}{y} r^y (1-r)^{N-y}$
  - The prior: $r \sim \beta\text{eta}(\alpha, \beta)$

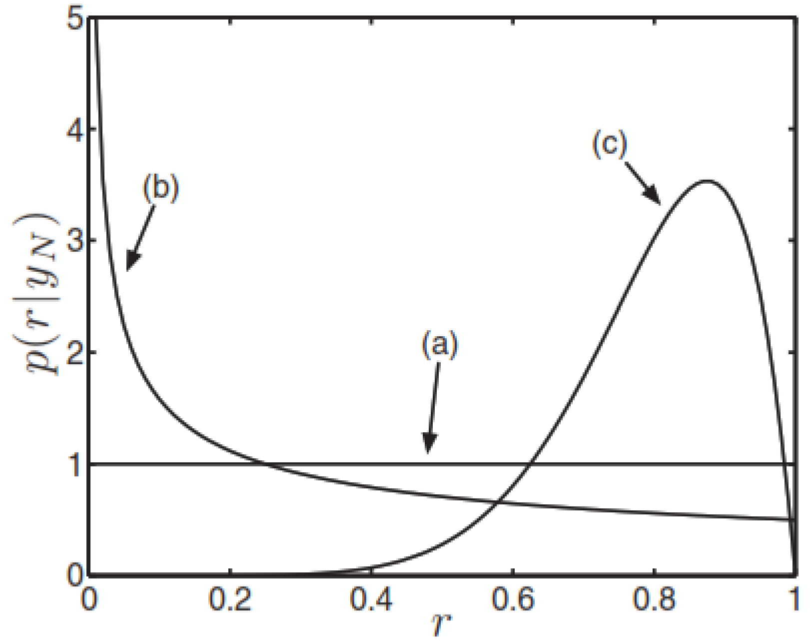


1. Know nothing: $\alpha = 1,\ \beta = 1.$
2. Fair coin: $\alpha = 50,\ \beta = 50.$
3. Biased: $\alpha = 5,\ \beta = 1.$

University of TEHRAN

A coin game [from FCML]

- (a) posterior distribution is uniform.
  - Combining the likelihood and the prior together has left all values of r equally likely.
- (b) suggests that r is most likely to be low but could be high.
  - Might be the result of a uniform prior and observing more tails than heads.
- (c) suggests the coin is biased to observing heads more often.

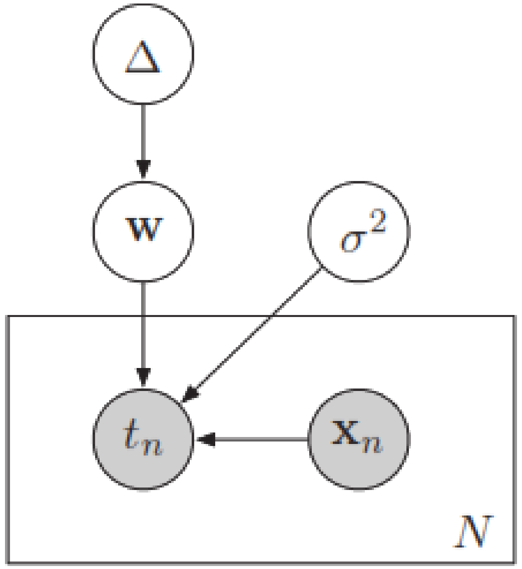


Bayesian linear regression-derivations

- From FCML Sec. 3.7
- Model

$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \cdots + w_K x_n^K + \epsilon_n, \qquad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

In vector form: $t_n = \mathbf{w}^{\mathsf{T}} \mathbf{x}_n + \epsilon_n$



- Likelihood

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

- The prior

$$p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$

Bayesian linear regression-derivations-2

- The posterior

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

$$
p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)\propto\exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X}\mathbf{w}+\frac{1}{\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}+\mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}-2\boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}.
$$

From the form of the function we know the posterior is Gaussian.

University of TEHRAN

Bayesian linear regression-derivations-3

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) \propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^\mathsf{T}\mathbf{X}\mathbf{w} + \frac{1}{\sigma^2}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} + \mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w} - 2\boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}.$$

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

Solving for Σ_w

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

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

Solving for μ_w

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

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

Bayesian linear regression-derivations-4

- Therefore the posterior is:

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

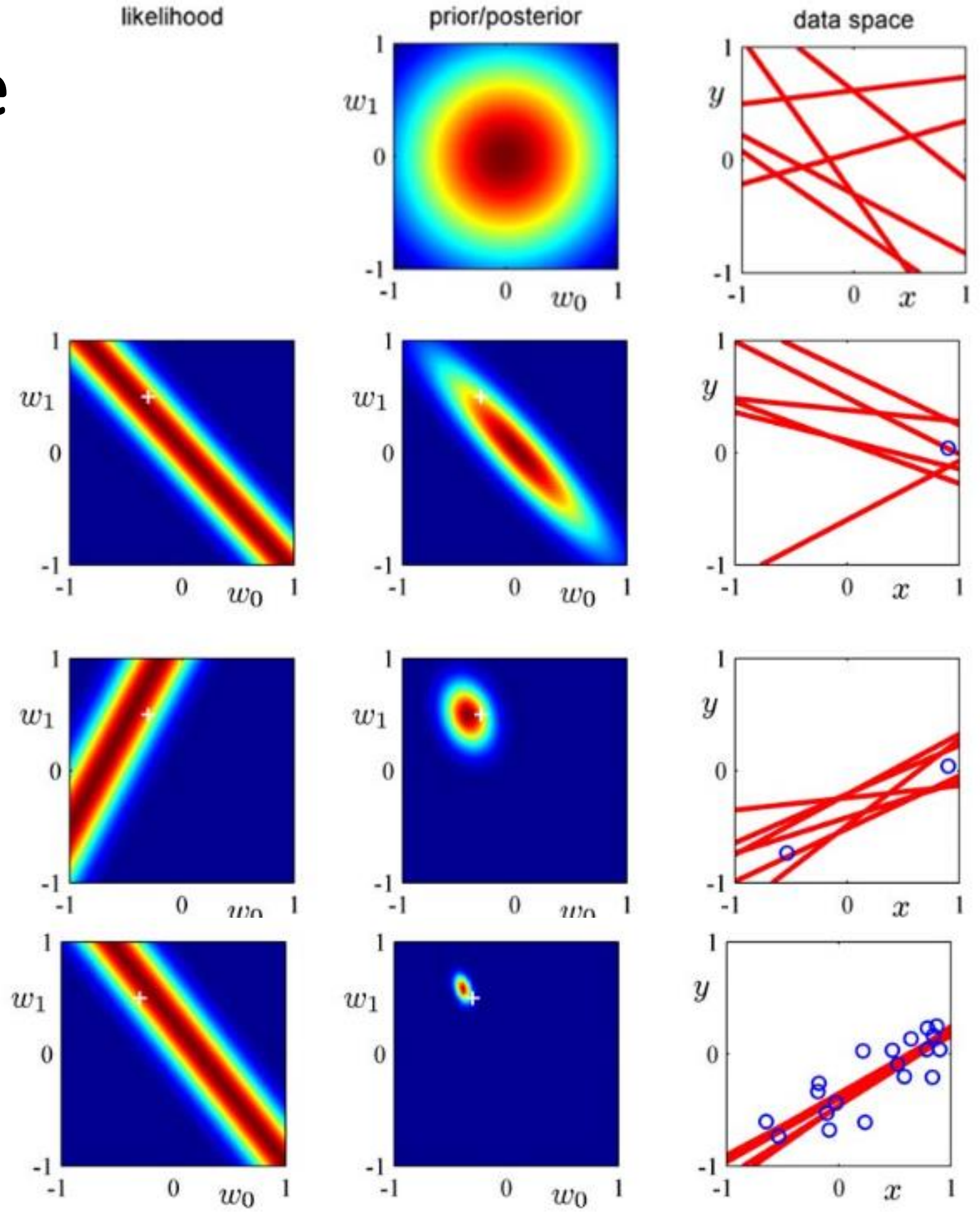
$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{w}} &= \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \\
\boldsymbol{\mu}_{\mathbf{w}} &= \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)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

Bayesian linear regression-example

- Input: a single variable *x*
- Output: a single target
- Linear model of the form

$$y(x, w) = w_0 + w_1 x$$

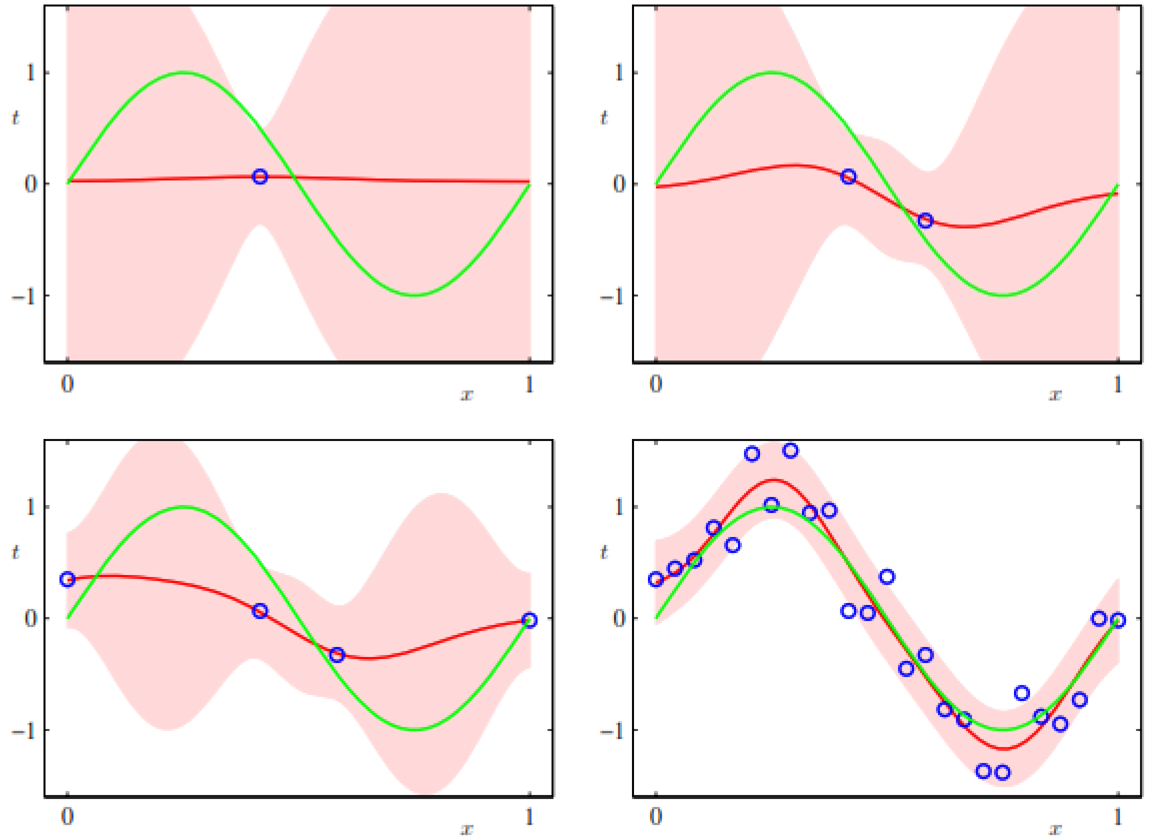
- We generate synthetic data from function $f(x) = -0.3 - 0.5x$ with addition of some noise to the target values



Bishop, pages 154-155, fig 3.7

Bayesian linear regression: predictive distribution

- Underlying function: $\sin(2\pi x)$ (green)
- Mean of Gaussian predictive distribution (red curve)
- **Question:** analyze in terms of bias/variance?



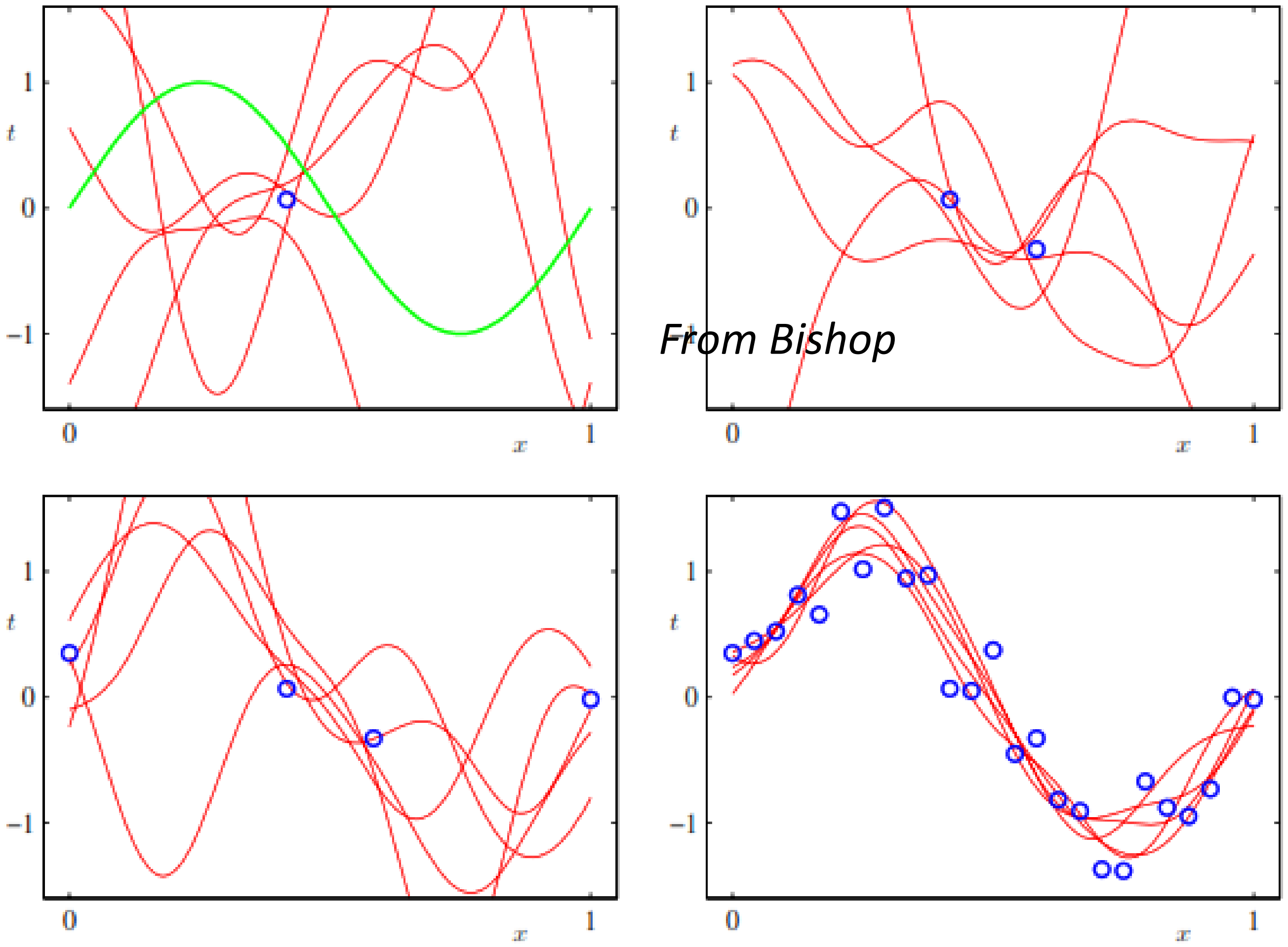
Bishop, pages fig 3.8

From Bishop

University of TEHRAN

Bayesian linear regression: predictive distribution

- Plots of the function $y(x, w)$ using samples from the posterior distributions over w.



From Bishop

Bishop, pages fig 3.9

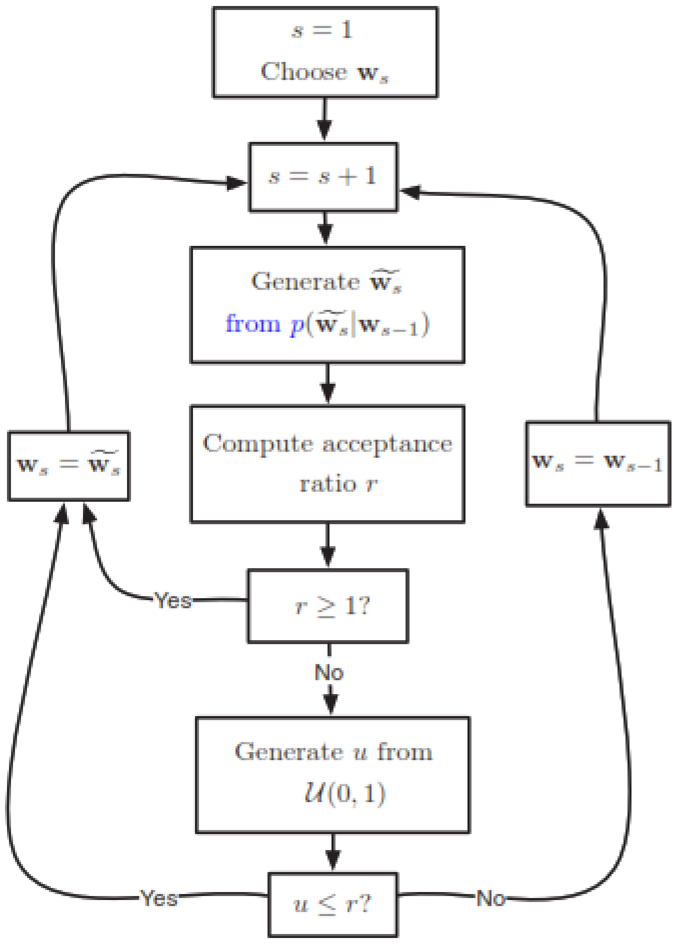
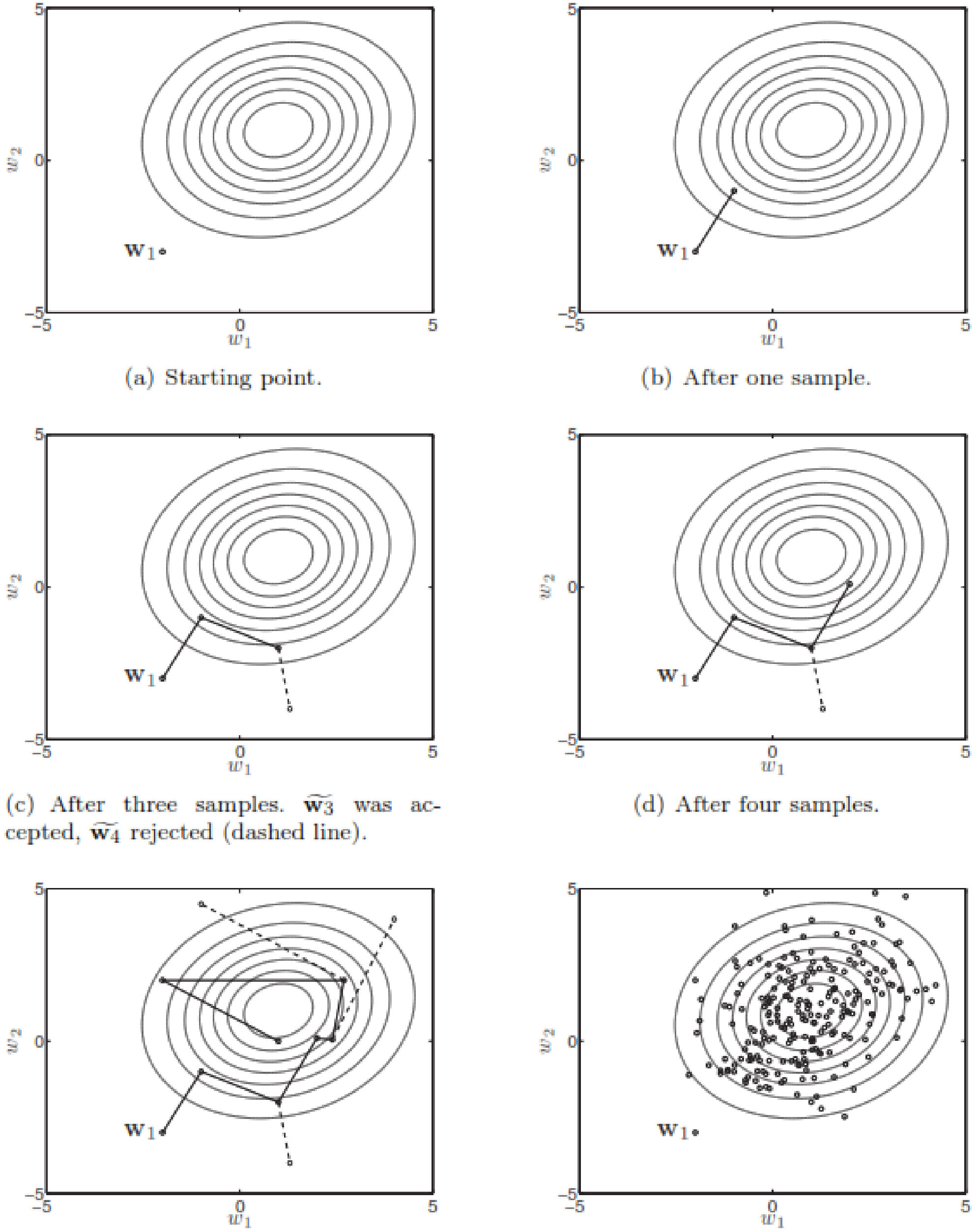


FIGURE 4.10 The Metropolis–Hastings algorithm.

From FCML

University of TEHRAN



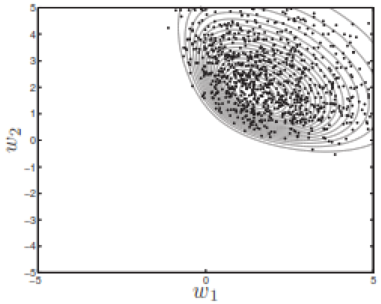
(a) Starting point. (b) After one sample.

(c) After three samples. $\widetilde{\mathbf{w}_3}$ was accepted, $\widetilde{\mathbf{w}_4}$ rejected (dashed line). (d) After four samples.

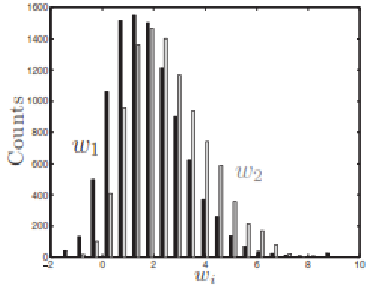
(e) After ten samples. (f) The first 300 samples.

FIGURE 4.11 Example of the Metropolis–Hastings agorithm in operation.

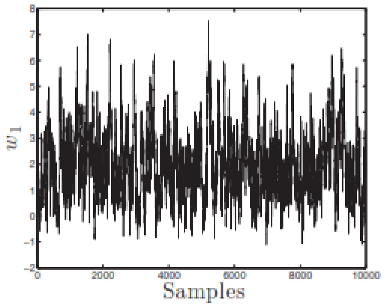
From FCML



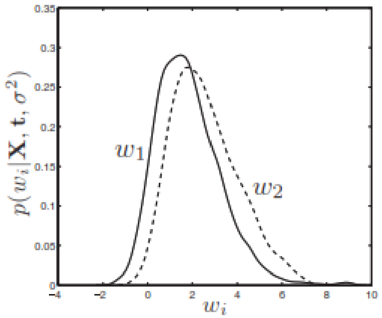
(a) One thousand of the MH samples along with the posterior contours.



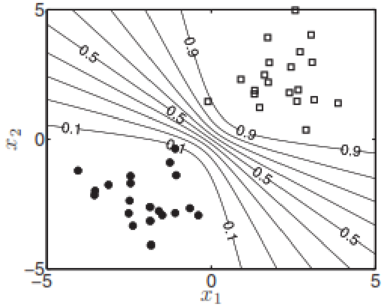
(b) Histograms of the samples for both w_1 (black) and w_2 (grey).



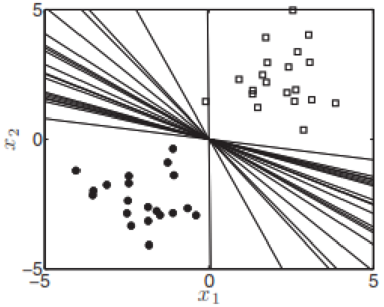
(c) All of the w_1 samples plotted against iteration, s.



(d) Continuous densities fitted to the w_1 and w_2 samples.



(e) Predictive probability contours. The contours show the probability of classifying an object at any location as a square. The probability of classifying an object as a circle at any point is 1 minus this value.



(f) Decision boundaries created from 20 randomly selected MH samples.

FIGURE 4.12 Results of applying the MH sampling algorithm to the binary response model.

References

- Pattern Recognition and Machine Learning by Christopher Bishop [PRML]
- A First Course in Machine Learning by Simon Rogers and Mark Girolami [FCML]