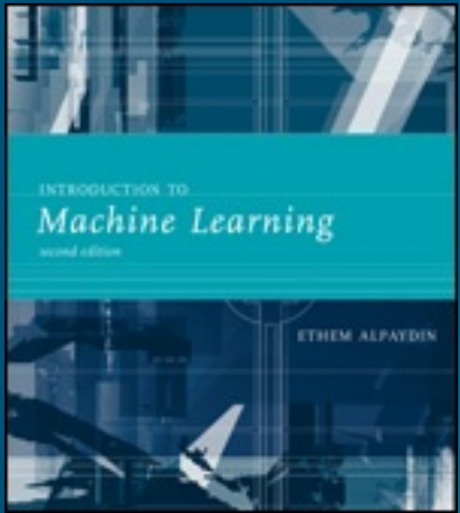
Lecture Slides for



INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN
© The MIT Press, 2010

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e

CHAPTER 14:

Bayesian Estimation

Maximum Likelihood vs. Bayes

Task: Given a dataset that comes from a normal distribution with mean μ, estimate the mean.

Maximum Likelihood Estimation: Assume μ is a unknown constant, estimate it based on data.

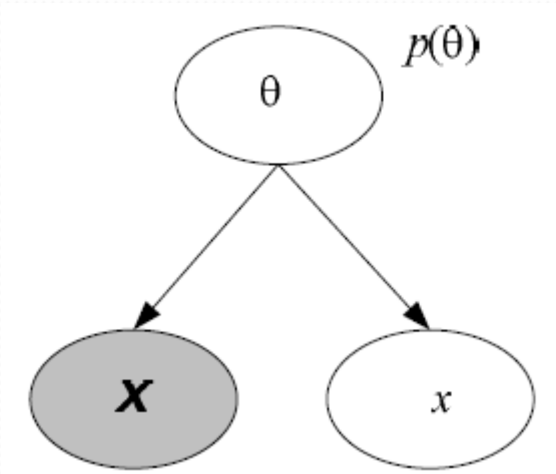
Bayes Estimation: Assume μ is a random variable with a certain prior probability distribution, using Bayes' rule, combine prior and the likelihood (based on data) to estimate the posterior distribution.

Rationale

- Bayes' Rule:

$$p(\theta \mid \mathrm{X}) = \frac{p(\theta)p(\mathrm{X} \mid \theta)}{p(\mathrm{X})}$$

- Generative model:



Arcs are in the direction of sampling:
First pick θ from p(θ)
Use θ to sample X and an instance x
X and x are independent given θ (see Bayesian networks)
Joint distr:

$$p(x,X,\theta)=p(\theta)p(X|\theta)p(x|\theta)$$

$$p(x|X)=p(x,X)/p(X)$$

$$= \int_\theta p(x,X,\theta)d\theta/p(X)$$

$$= \int_\theta p(\theta)p(X|\theta)p(x|\theta)d\theta/p(X)$$

$$= \int_\theta p(\theta|X)p(x|\theta)d\theta$$

If discrete random vars: replace integral ($\int$) with summation.

Bayesian, MAP, ML Estimator

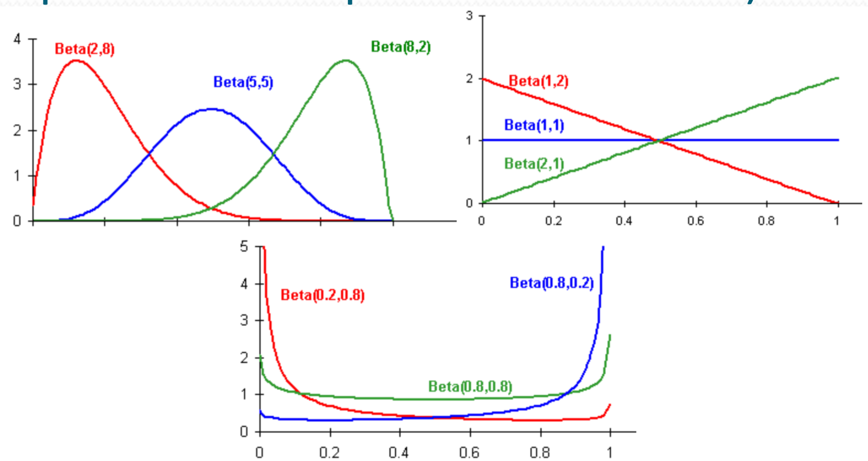
- Bayesian Estimate: Integrate to compute the posterior
  - *Problem: The integral may not be easy to compute.*
- MAP Estimate: Assuming posterior peaks around a single point (mode):
  - Θ_{MAP}=arg $\max_{\Theta}$ $p(\Theta|X)$
  - $p_{MAP}(x|X)=p(x|\Theta_{MAP})$
- Maximum Likelihood Estimate: if prior $p(\Theta)$ is uniform, then mode of posterior and mode of likelihood are at the same Θ, hence ML estimate = MAP estimate

Estimating the Parameters of a Distribution: Discrete case

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx$$

$$\Gamma(n) = (n-1)!$$

- x^t_i=1 if in instance *t* is in state *i*, probability of state *i* is q_i
- Dirichlet prior, α_i are hyperparameters $$\text{Dirichlet}(\mathbf{q} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{i=1}^{K} q_i^{\alpha_i - 1}$$
- Sample likelihood $$p(X \mid \mathbf{q}) = \prod_{t=1}^{N} \prod_{i=1}^{K} q_i^{x_i^t}$$
- Posterior $$p(\mathbf{q} \mid X) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + N_1)\cdots\Gamma(\alpha_K + N_K)} \prod_{i=1}^{K} q_i^{\alpha_i + N_i - 1}$$ $$= \text{Dirichlet}(\mathbf{q} \mid \boldsymbol{\alpha} + \mathbf{n})$$
- Dirichlet is a conjugate prior (shape of the posterior and prior are the same)
- With *K*=2, Dirichlet distr reduces to
- Beta distribution



$$f(x) = \frac{(x)^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad \mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$$

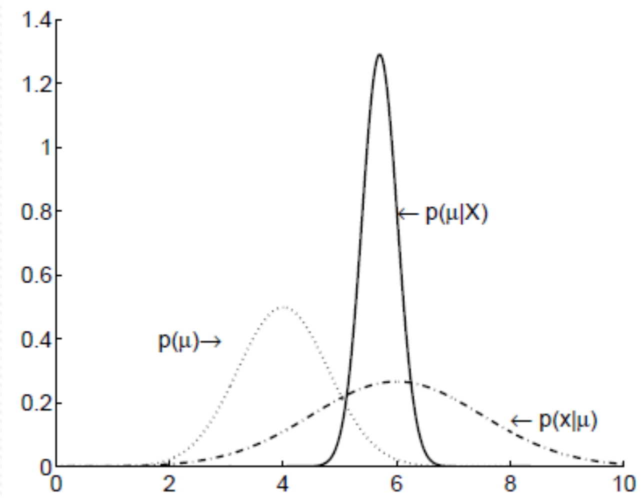
where $B(\alpha,\beta)$ is a Beta function

Estimating the Parameters of a Distribution: Continuous case

- $p(x^t) \sim N(\mu, \sigma^2)$
- Gaussian prior for mean μ, $p(\mu) \sim N(\mu_0, \sigma_0^2)$
- Posterior: $p(\mu|X) \propto p(\mu)p(X|\mu)$
- Posterior is also Gaussian $p(\mu|X) \sim N(\mu_N, \sigma_N^2)$ where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m$$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$



- To estimate the precision (λ=1/variance)
- Use Gamma prior, posterior is also Gamma

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx$$

$$\Gamma(n) = (n-1)!$$

$$g(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

prior: $p(\lambda)$=Gamma(a_0,b_0) posterior:$p(\lambda|X) \propto p(X|\lambda)p(\lambda) \sim$ Gamma(a_N,b_N), $a_N = a_0 + N/2$

$b_N = b_0 + s^2 N/2$

Estimating the Parameters of a Function: Regression

- $r=\boldsymbol{w}^T\boldsymbol{x}+\varepsilon$ where $p(\varepsilon)\sim N(0,1/\beta)$, and $p(r^t|x^t,\boldsymbol{w},\beta) \sim N(\boldsymbol{w}^T\boldsymbol{x}^t, 1/\beta)$
- Log likelihood

$$L(\mathbf{r}\,|\,\mathbf{X},\mathbf{w},\beta)=\log\prod_t p(r^t\,|\,\mathbf{x}^t,\mathbf{w},\beta)$$

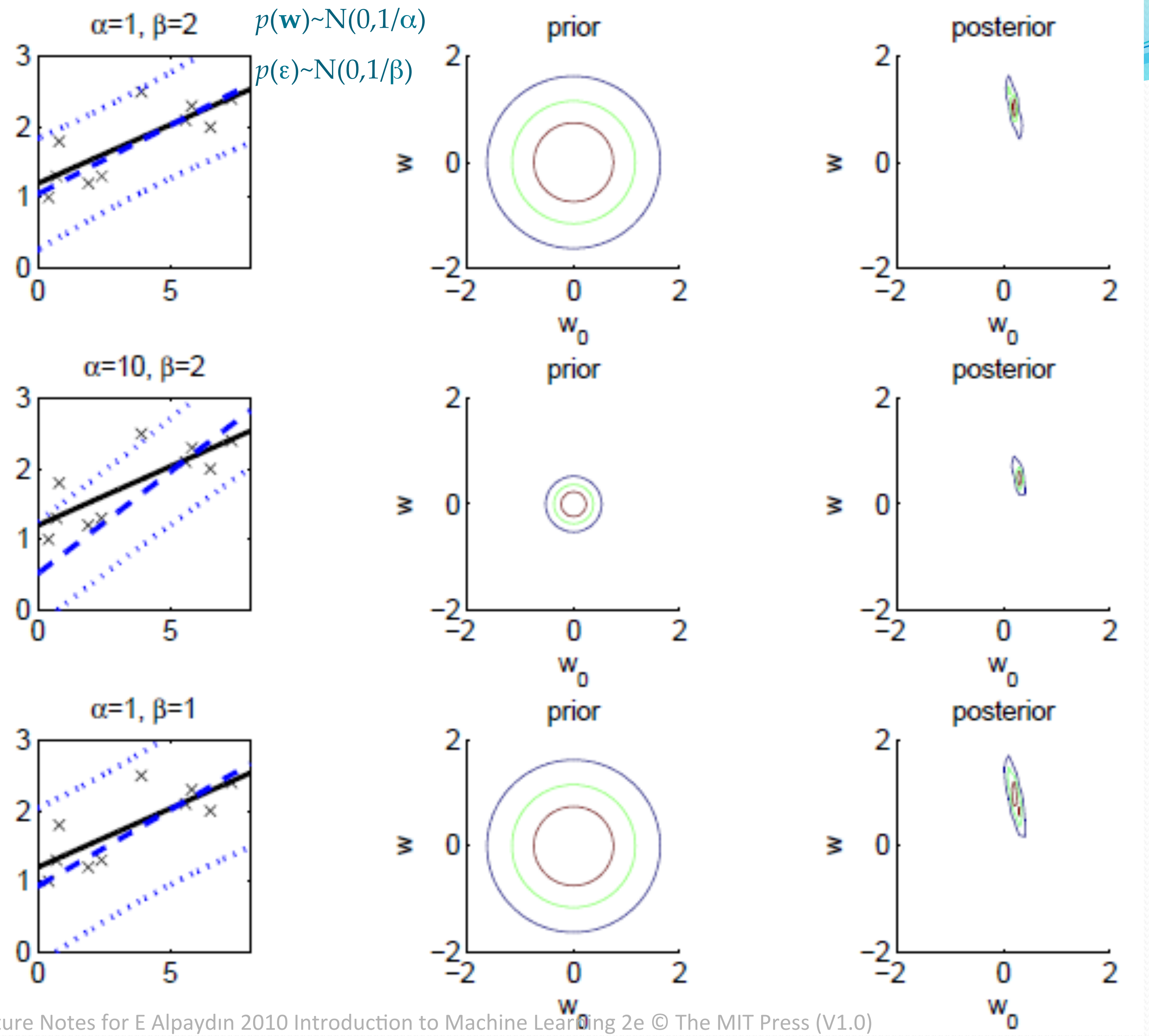
$$=-N\log\left(\sqrt{2\pi}\right)+N\log\beta-\frac{\beta}{2}\sum_t\left(r^t-\mathbf{w}^T\mathbf{x}^t\right)$$

- ML solution $\mathbf{w}_{ML}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}$
- Gaussian conjugate prior: $p(\mathbf{w})\sim N(0,1/\alpha)$
- Posterior: $p(\mathbf{w}|\mathbf{X})\sim N(\boldsymbol{\mu}_N,\Sigma_N)$ where

$$\boldsymbol{\mu}_N=\beta\boldsymbol{\Sigma}_N\mathbf{X}^T\mathbf{r}$$

$$\boldsymbol{\Sigma}_N=(\alpha\mathbf{I}+\beta\mathbf{X}^T\mathbf{X})^{-1}$$

- Generating output for input x': Integrate over the full posterior:
- $r' = \int w^T x' p(w|X)dw$ Integrate over all possible w's


α=1, β=2
p(w)~N(0,1/α)
p(ε)~N(0,1/β)
prior
posterior
α=10, β=2
prior
posterior
α=1, β=1
prior
posterior
w
w0

Basis/Kernel Functions

- For new ***x***', the estimate r' is calculated as

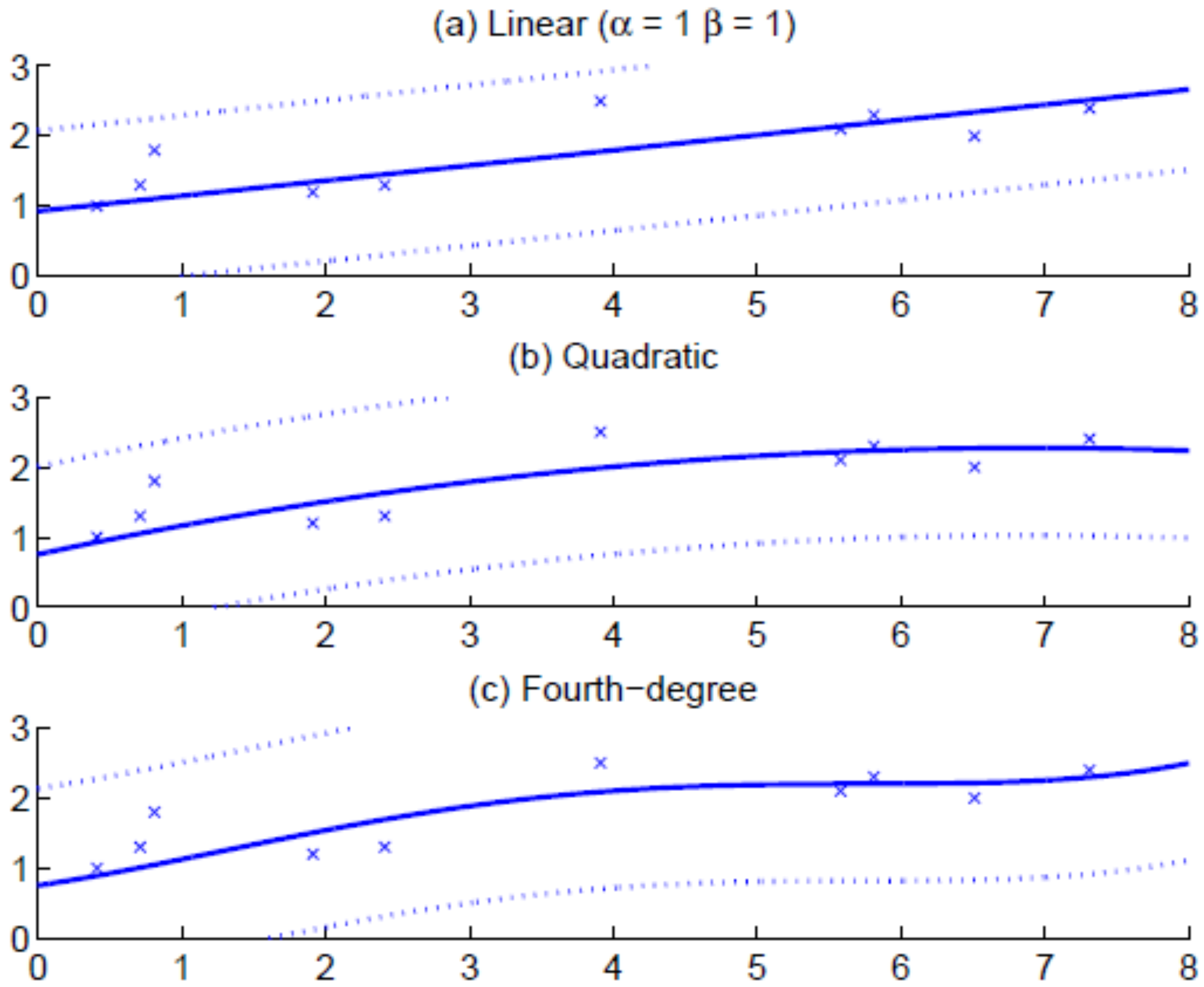
$$r' = (\mathbf{x}')^T w$$
$$= \beta(\mathbf{x}')^T \mathbf{\Sigma}_N \mathbf{X}^T \mathbf{r}$$
$$= \sum_t \beta(\mathbf{x}')^T \mathbf{\Sigma}_N \mathbf{x}^t r^t$$

Dual representation

- Linear kernel $\quad r' = \sum_t \beta(\mathbf{x}')^T \mathbf{\Sigma}_N \mathbf{x}^t r^t \sum_t \beta K(\mathbf{x}', \mathbf{x}^t) r^t$

- For any other ϕ(**x**), we can write $K(\mathbf{x}',\mathbf{x})=\phi(\mathbf{x}')^T\phi(\mathbf{x})$

Kernel Functions



Bayesian Classification

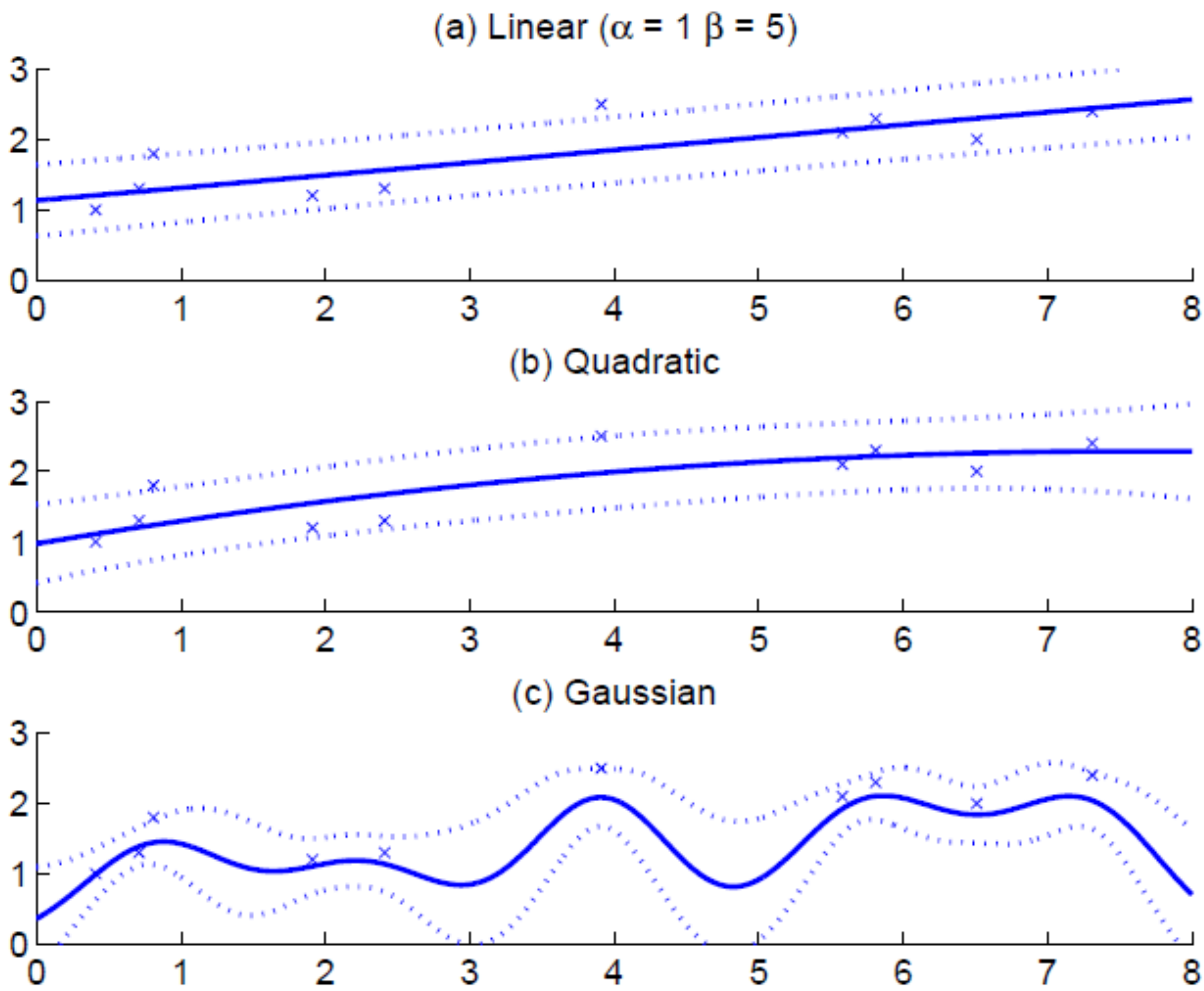
- Assume weights have a zero mean Gaussian prior
- Write down the posterior for weights (given X and r)
- Posterior is not Gaussian and can not be computed exactly.
- Use Laplace Approximation to the posterior
- Find the mode of the posterior
- Fit a Gaussian centered at this mode
- Variance: Taylor expression involving the second derivatives matrix (Hessian)

Gaussian Processes

- For the linear model, instead of a single output y for input x, obtain an output distribution based on the distribution p(w) of weights
- p(w) is a Gaussian,y is a linear combination of Gaussians, y is Gaussian
- We want to compute the joint distr of y values calculated at N points
- Assume Gaussian prior on inputs $p(\mathbf{w}) \sim N(0,1/\alpha)$
- $\mathbf{y}=\mathbf{X}\mathbf{w}$, where $E[\mathbf{y}]=0$ and $\text{Cov}(\mathbf{y})=\mathbf{K}$ with Gram Matrix $\mathbf{K}$, $\mathbf{K}_{ij}= (\mathbf{x}^i)^T\mathbf{x}^i$
- $\mathbf{K}$ is the covariance function, here linear
- With basis function $\phi(\mathbf{x})$, $\mathbf{K}_{ij}= (\phi(\mathbf{x}^i))^T\phi(\mathbf{x}^i)$
- $r \sim N_N(\mathbf{0},C_N)$ where $C_N= (1/\beta)\mathbf{I}+\mathbf{K}$
- With new $\mathbf{x}'$ added as $\mathbf{x}_{N+1}$, $r_{N+1} \sim N_{N+1}(0,C_{N+1})$

$$\mathbf{C}_{N+1} = \begin{bmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k} & c \end{bmatrix}$$

where $\mathbf{k} = [K(\mathbf{x}',\mathbf{x}^t)_t]^T$ and $c=K(\mathbf{x}',\mathbf{x}')+1/\beta$.
$p(r'|\mathbf{x}',\mathbf{X},\mathbf{r}) \sim N(\mathbf{k}^T\mathbf{C}_{N-1}\mathbf{r}, c-\mathbf{k}^T\mathbf{C}_{N-1}\mathbf{k})$


(a) Linear (α = 1 β = 5)
(b) Quadratic
(c) Gaussian
0
1
2
3
4
5
6
7
8