CS4801 : MLE, MAP, regression

Sahely Bhadra
14/8/2017

1. Maximum Likelihood Estimation
2. Maximum-a-priori Estimation
3. Linear Regression : MLE and MAP

Slides are taken from Dr Piyush Rai

Probability Basic

Gaussian (normal)

- For a continuous univariate random variable:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}$$

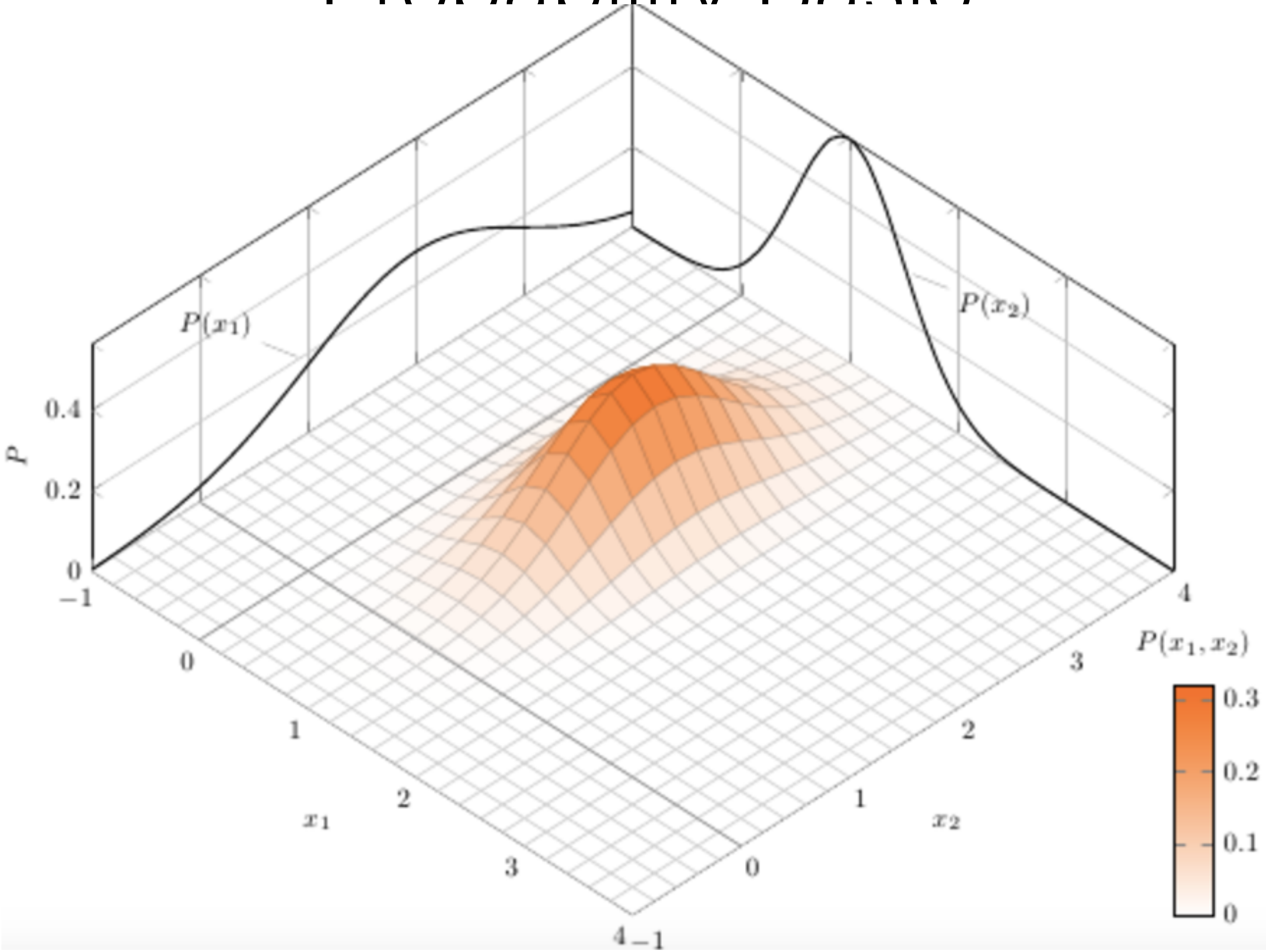
Multivariate Gaussian Distribution

- For a continuous vector random variable:

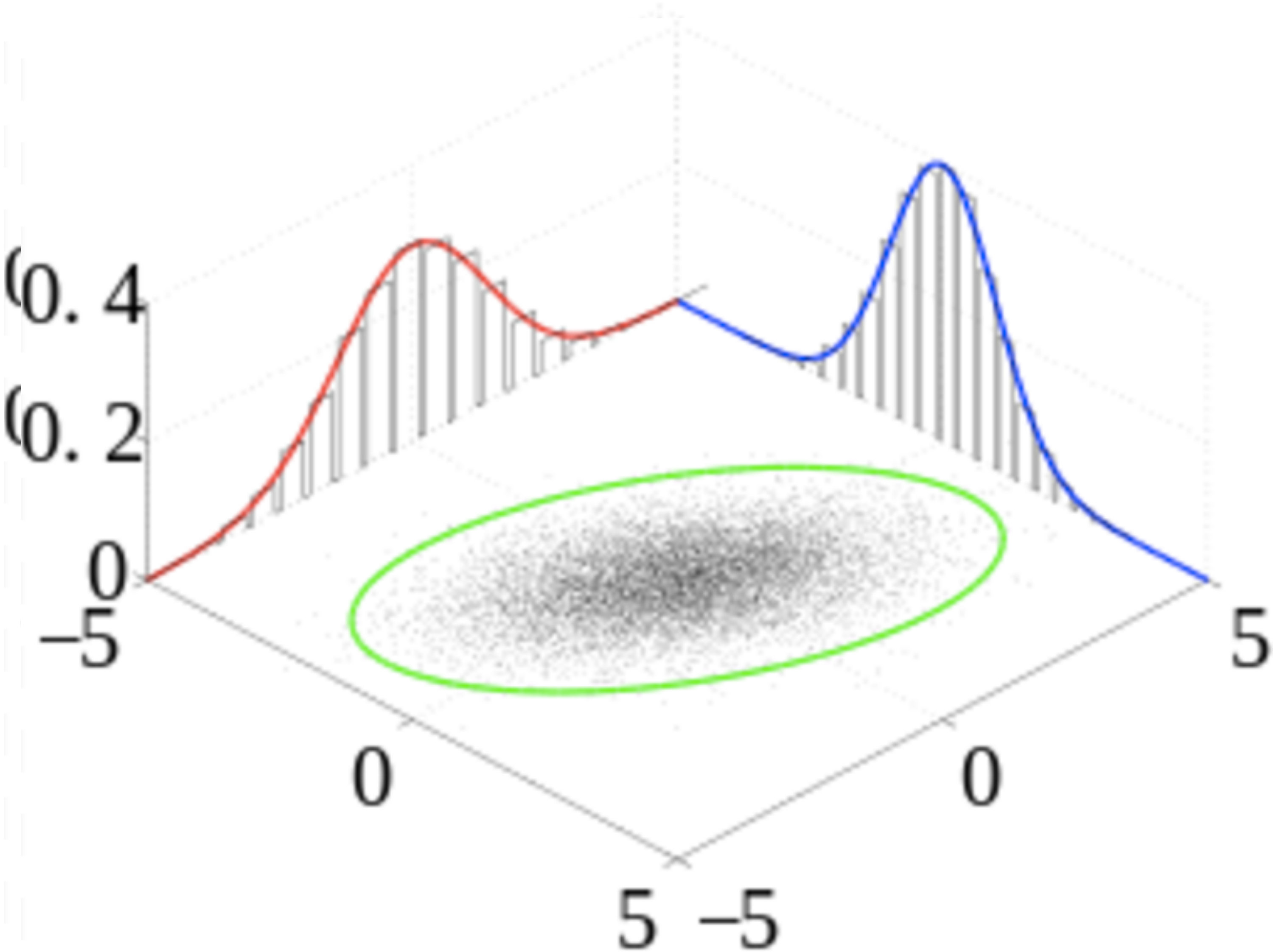
$$p(x|\mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Distribution with maximum entropy for fixed variance

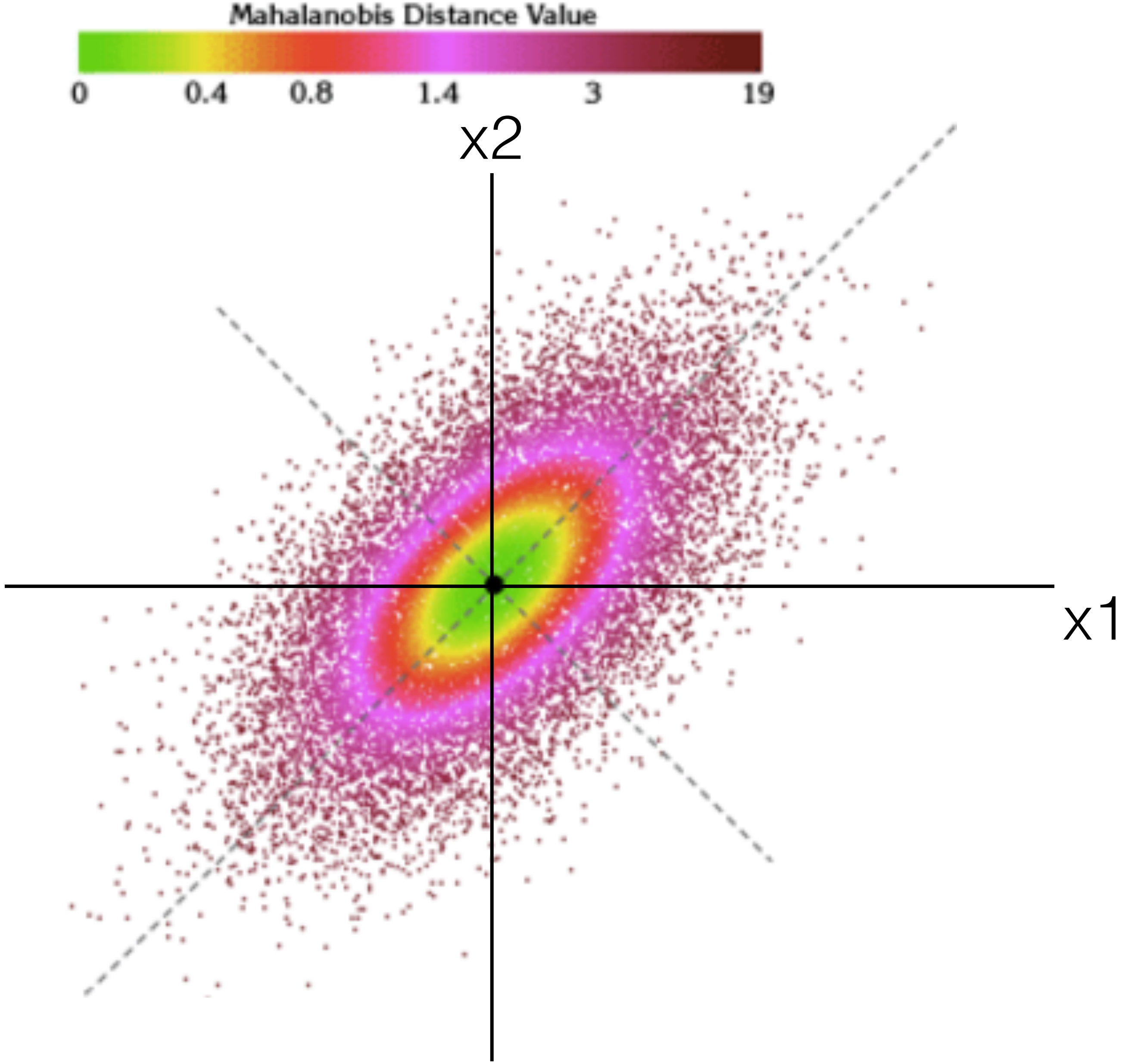
Probability Basic



Probability Basic



Probability Basic

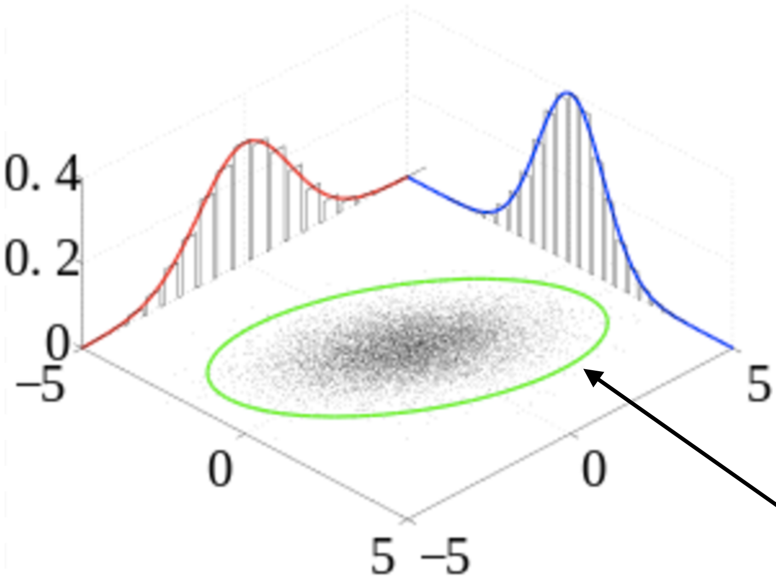


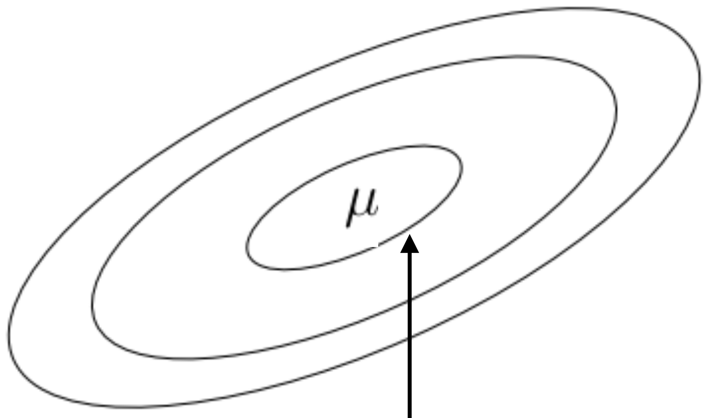
Probability Basic

Multivariate Gaussian Distribution

- For a continuous vector random variable:

$$p(x|\mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$





$$\underbrace{(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)}_{\textbf{Mahalanobis distance}} = C$$

Maximum Likelihood Estimation(MLE)

Gaussian distribution

p($\mathcal{D}/\mu, \sigma^2$)

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_1-\mu)^2}{2\sigma^2}\right) \cdots \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2.$$

log likelihood function $\ln$ p(D/μ, σ^2) $= -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2.$

$$\frac{\partial}{\partial\mu} \ln \text{p}(\mathcal{D}/\mu, \sigma^2) : \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i-\mu) = .\frac{1}{\sigma^2} n(\bar{x}-\mu)$$

$$\frac{\partial}{\partial\sigma^2} \ln \text{p}(\mathcal{D}/\mu, \sigma^2) = -\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i-\mu)^2 = \frac{n}{(\sigma^2)^2} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i-\mu)^2\right).$$

MLE $\hat{\mu}(\mathbf{x}) = \bar{x}$

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x})^2.$$

I.I.D samples of the data

Assume data generated via a probabilistic model

$$\mathbf{d} \sim P(\mathbf{d} \mid \theta)$$

$P(\mathbf{d} \mid \theta)$: Probability distribution underlying the data

- θ: fixed but unknown distribution parameter

Given: N independent and identically distributed (i.i.d.) samples of the data

$$\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_N\} \qquad \mathbf{d_i = \{ x_i, y_i \}}$$

Independent and Identically Distributed:

- Given θ, each sample $\mathbf{d}_\mathbf{i}$ is independent of all other samples
- All samples $\mathbf{d}_\mathbf{i}$ drawn from the same distribution

Goal: Estimate parameter θ that best models/describes the data

Several ways to define the "best"

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE): Choose the parameter θ that maximizes the probability of the data, *given* that parameter

Probability of the data, given the parameters is called the Likelihood, a function of θ and defined as:

$$\mathcal{L}(\theta) = P(\mathcal{D} \mid \theta) = P(\mathbf{d}_1, \ldots, \mathbf{d}_N \mid \theta) = \prod_{i=1}^{N} P(\mathbf{d}_i \mid \theta)$$

MLE typically maximizes the Log-likelihood instead of the likelihood

Log-likelihood:

$$\log \mathcal{L}(\theta) = \log P(\mathcal{D} \mid \theta) = \log \prod_{i=1}^{N} P(\mathbf{d}_i \mid \theta) = \sum_{i=1}^{N} \log P(\mathbf{d}_i \mid \theta)$$

Maximum Likelihood parameter estimation

$$\boxed{\hat{\theta}_{MLE} = \arg\max_{\theta} \log \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \log P(\mathbf{d}_i \mid \theta)}$$

Maximum-a-posteriori Estimation

Maximum-a-Posteriori Estimation (MAP): Choose θ that maximizes the posterior probability of θ (i.e., probability in the light of the observed data)

Posterior probability of θ is given by the Bayes Rule

$$P(\theta \mid \mathcal{D}) = \frac{P(\theta)P(\mathcal{D} \mid \theta)}{P(\mathcal{D})}$$

$P(\theta)$: Prior probability of θ (without having seen any data)

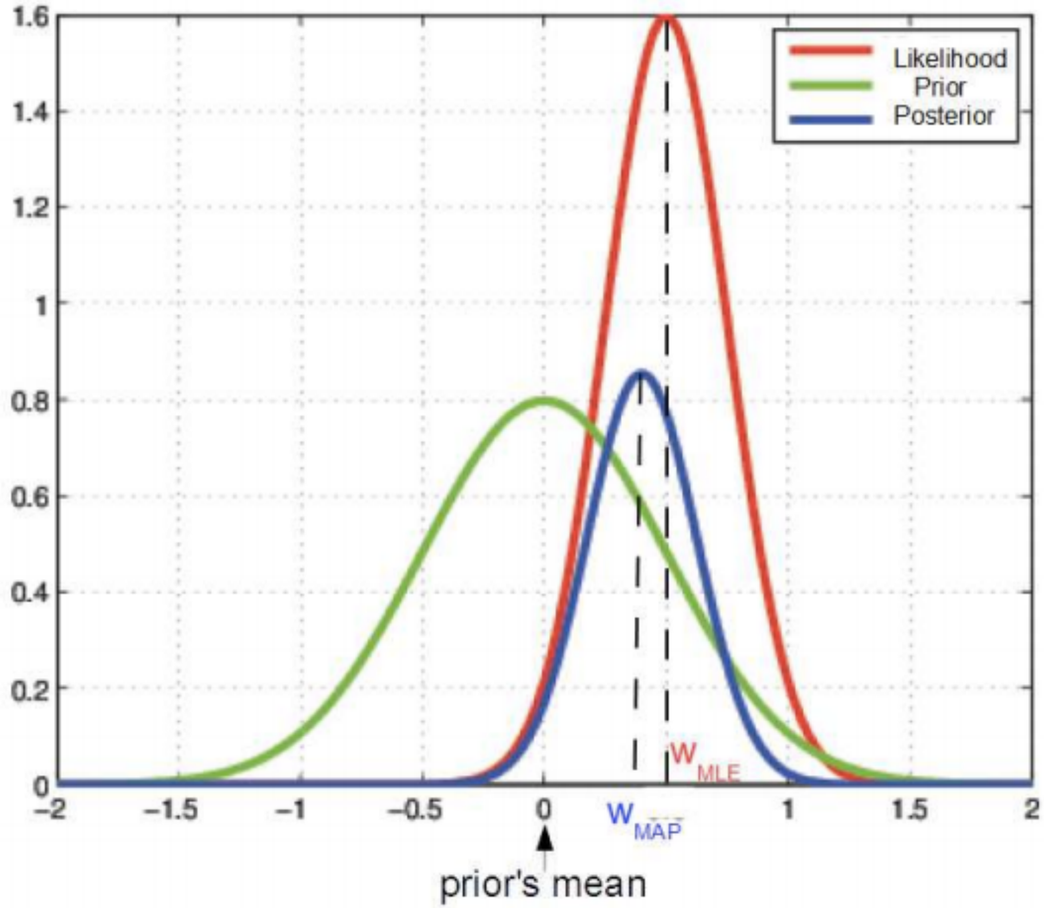
$P(\mathcal{D} \mid \theta)$: Likelihood

$P(\mathcal{D})$: Probability of the data (independent of θ)

$$P(\mathcal{D}) = \int P(\theta)P(\mathcal{D} \mid \theta)d\theta \quad \text{(sum over all } \theta\text{'s)}$$

The Bayes Rule lets us update our belief about θ in the light of observed data

While doing MAP, we usually maximize the log of the posterior probability



1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−2
−1.5
−1
−0.5
0
1
1.5
2
Likelihood
Prior
Posterior
w_MLE
w_MAP
prior's mean

Maximum-a-posteriori Estimation

Maximum-a-Posteriori parameter estimation

$$\begin{aligned}
\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) &= \arg\max_{\theta} \frac{P(\theta)P(\mathcal{D} \mid \theta)}{P(\mathcal{D})} \\
&= \arg\max_{\theta} P(\theta)P(\mathcal{D} \mid \theta) \\
&= \arg\max_{\theta} \log P(\theta)P(\mathcal{D} \mid \theta) \\
&= \arg\max_{\theta} \{\log P(\theta) + \log P(\mathcal{D} \mid \theta)\}
\end{aligned}$$

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \{\log P(\theta) + \sum_{i=1}^{N} \log P(\mathbf{d}_i \mid \theta)\}}$$

Same as MLE except the extra log-prior-distribution term!

MAP allows incorporating our prior knowledge about θ in its estimation

Linear Regression

Each response generated by a linear model plus some Gaussian noise

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon$$

Noise ϵ is drawn from a Gaussian distribution:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Each response y then becomes a draw from the following Gaussian:

$$y \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$$

Probability of each response variable

$$P(y \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(y \mid \mathbf{w}^\top \mathbf{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y - \mathbf{w}^\top \mathbf{x})^2}{2\sigma^2}\right]$$

Given data $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, we want to estimate the weight vector $\mathbf{w}$

Linear Regression : MLE

Log-likelihood:

$$
\begin{aligned}
\log \mathcal{L}(\mathbf{w}) = \log P(\mathcal{D} \mid \mathbf{w}) = \log P(\mathbf{Y} \mid \mathbf{X}, \mathbf{w}) &= \log \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w}) \\
&= \sum_{i=1}^{N} \log P(y_i \mid \mathbf{x}_i, \mathbf{w}) \\
&= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} \left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2}\right\}
\end{aligned}
$$

Maximum Likelihood Solution: $\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \log P(\mathcal{D} \mid \mathbf{w})$

$$
\begin{aligned}
&= \arg\max_{\mathbf{w}} -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \\
&= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2
\end{aligned}
$$

For $\sigma = 1$ (or some constant) for each input, it's equivalent to the least-squares objective for linear regression

Linear Regression : MAP

Let's assume a Gaussian prior distribution over the weight vector $\mathbf{w}$

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid 0, \lambda^{-1}\mathbf{I}) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}\right)$$

Log posterior probability:

$$\log P(\mathbf{w} \mid \mathcal{D}) = \log \frac{P(\mathbf{w})P(\mathcal{D} \mid \mathbf{w})}{P(\mathcal{D})} = \log P(\mathbf{w}) + \log P(\mathcal{D} \mid \mathbf{w}) - \log P(\mathcal{D})$$

Maximum-a-Posteriori Solution: $\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \log P(\mathbf{w} \mid \mathcal{D})$

$$\begin{aligned}
&= \arg\max_{\mathbf{w}} \{\log P(\mathbf{w}) + \log P(\mathcal{D} \mid \mathbf{w}) - \log P(\mathcal{D})\} \\
&= \arg\max_{\mathbf{w}} \{\log P(\mathbf{w}) + \log P(\mathcal{D} \mid \mathbf{w})\} \\
&= \arg\max_{\mathbf{w}} \left\{-\frac{D}{2}\log(2\pi) - \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w} + \sum_{i=1}^{N} \left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2}\right\}\right\} \\
&= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w} \quad \text{(ignoring constants and changing max to min)}
\end{aligned}$$

For $\sigma = 1$ (or some constant) for each input, it's equivalent to the regularized least-squares objective

MLE vs MAP

MLE solution:

$$\boxed{\hat{\mathbf{w}}_{MLE} = \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2}$$

MAP solution:

$$\boxed{\hat{\mathbf{w}}_{MAP} = \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w}}$$

Take-home messages:

- MLE estimation of a parameter leads to unregularized solutions
- MAP estimation of a parameter leads to regularized solutions
- The prior distribution acts as a regularizer in MAP estimation

Note: For MAP, different prior distributions lead to different regularizers

- Gaussian prior on $\mathbf{w}$ regularizes the ℓ_2 norm of $\mathbf{w}$
- Laplace prior $\exp(-C||\mathbf{w}||_1)$ on $\mathbf{w}$ regularizes the ℓ_1 norm of $\mathbf{w}$

Next Classes

- 16/8
  - Classification