Parametric Techniques

Jason J. Corso

SUNY at Buffalo

Introduction

- When covering Bayesian Decision Theory, we assumed the full probabilistic structure of the problem was know.
- However, this is rarely the case in practice.
- Instead, we have some knowledge of the problem and some example data and we must estimate the probabilities.
- In the discriminants chapter, we learned how to estimate linear boundaries separating the data, assuming nothing about the specific structure of the data. Here, we resort to assuming some structure to the data and estimate the parameters of this structure.
- **Focus of this lecture** is to study a pair of techniques for estimating the parameters of the likelihood models (given a particular form of the density, such as a Gaussian).

Introduction

- When covering Bayesian Decision Theory, we assumed the full probabilistic structure of the problem was know.
- However, this is rarely the case in practice.
- Instead, we have some knowledge of the problem and some example data and we must estimate the probabilities.
- In the discriminants chapter, we learned how to estimate linear boundaries separating the data, assuming nothing about the specific structure of the data. Here, we resort to assuming some structure to the data and estimate the parameters of this structure.
- **Focus of this lecture** is to study a pair of techniques for estimating the parameters of the likelihood models (given a particular form of the density, such as a Gaussian).
- **Parametric Models** – For a particular class ω_i, we consider a set of parameters $\boldsymbol{\theta}_i$ to fully define the likelihood model.
  - For the Guassian, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

Introduction

- When covering Bayesian Decision Theory, we assumed the full probabilistic structure of the problem was know.
- However, this is rarely the case in practice.
- Instead, we have some knowledge of the problem and some example data and we must estimate the probabilities.
- In the discriminants chapter, we learned how to estimate linear boundaries separating the data, assuming nothing about the specific structure of the data. Here, we resort to assuming some structure to the data and estimate the parameters of this structure.
- **Focus of this lecture** is to study a pair of techniques for estimating the parameters of the likelihood models (given a particular form of the density, such as a Gaussian).
- **Parametric Models** – For a particular class ω_i, we consider a set of parameters $\boldsymbol{\theta}_i$ to fully define the likelihood model.
  - For the Guassian, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
- **Supervised Learning** – we are working in a supervised situation where we have an set of training data:

$$\mathcal{D} = \{(\mathbf{x}, \omega)_1, (\mathbf{x}, \omega)_2, \ldots (\mathbf{x}, \omega)_N\} \tag{1}$$

Overview of the Methods

- **Intuitive Problem:** Given a set of training data, $\mathcal{D}$, containing labels for c classes, train the likelihood models $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$ by estimating the parameters $\boldsymbol{\theta}_i$ for $i = 1, \ldots, c$.

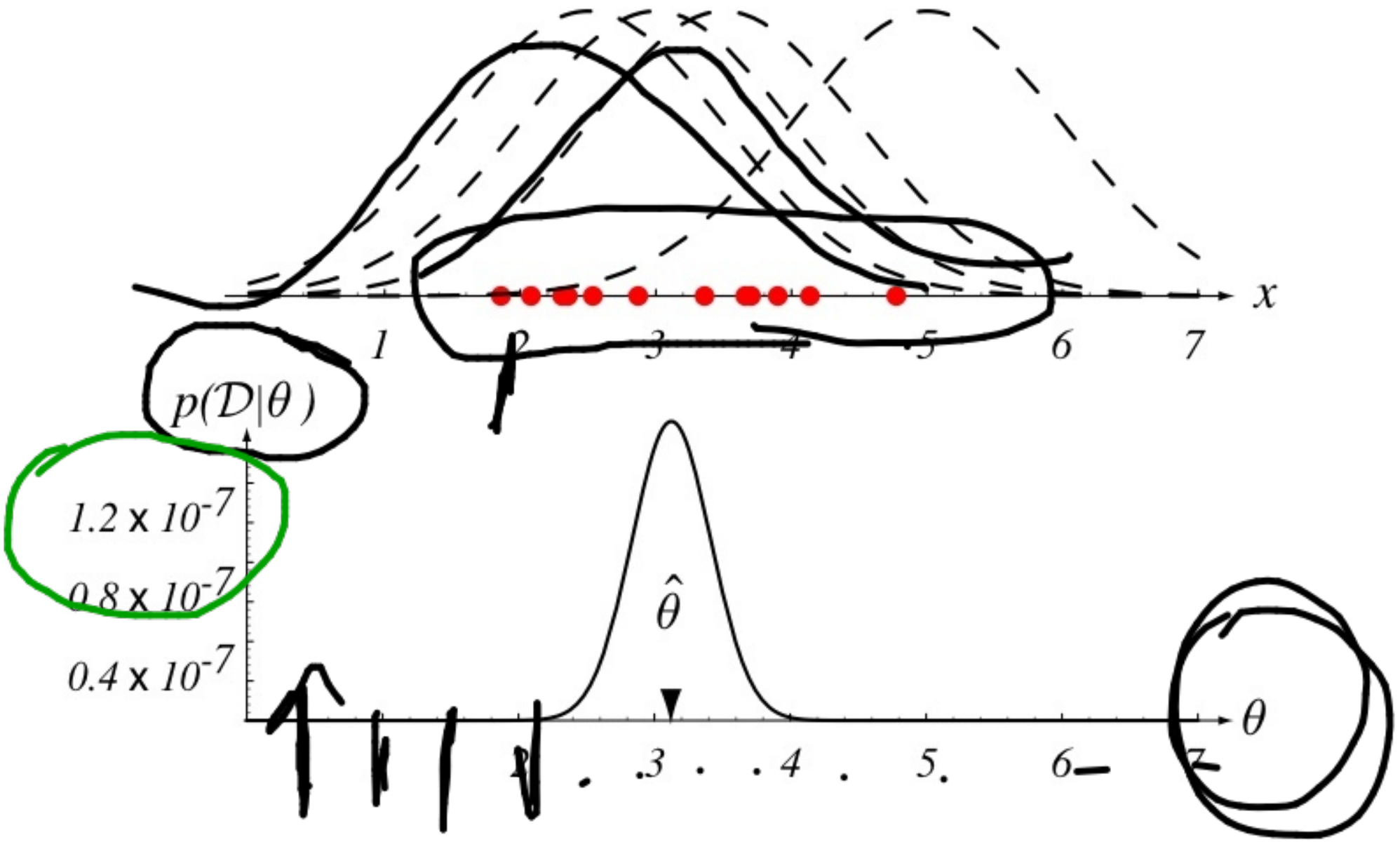
Overview of the Methods

- **Intuitive Problem:** Given a set of training data, $\mathcal{D}$, containing labels for c classes, train the likelihood models $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$ by estimating the parameters $\boldsymbol{\theta}_i$ for $i = 1, \ldots, c$.
- **Maximum Likelihood Parameter Estimation**
  - Views the parameters as quantities that are fixed by unknown.
  - The best estimate of their value is the one that maximizes the probability of obtaining the samples in $\mathcal{D}$.

Overview of the Methods

- **Intuitive Problem:** Given a set of training data, $\mathcal{D}$, containing labels for c classes, train the likelihood models $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$ by estimating the parameters $\boldsymbol{\theta}_i$ for $i = 1, \ldots, c$.
- **Maximum Likelihood Parameter Estimation**
  - Views the parameters as quantities that are fixed by unknown.
  - The best estimate of their value is the one that maximizes the probability of obtaining the samples in $\mathcal{D}$.
- **Bayesian Parameter Estimation**
  - Views the parameters as random variables having some known prior distribution.
  - The samples convert this prior into a posterior and revise our estimate of the distribution over the parameters.
  - We shall typically see that the posterior is increasingly peaked for larger $\mathcal{D}$ — *Bayesian Learning*.

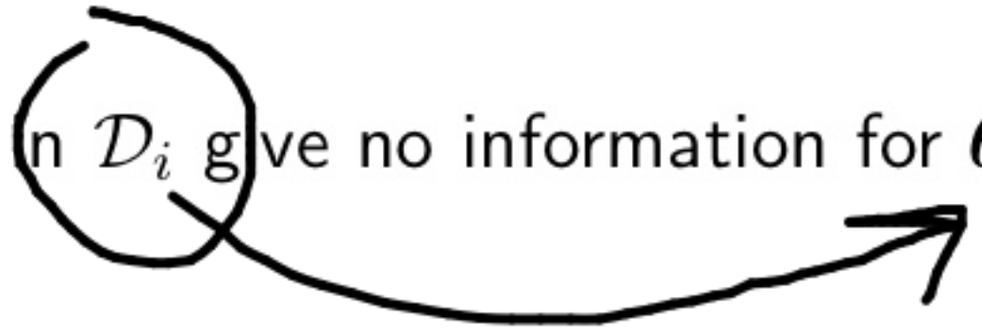
Maximum Likelihood Intuition



- Underlying model is assumed to be a Gaussian of particular variance but unknown mean.

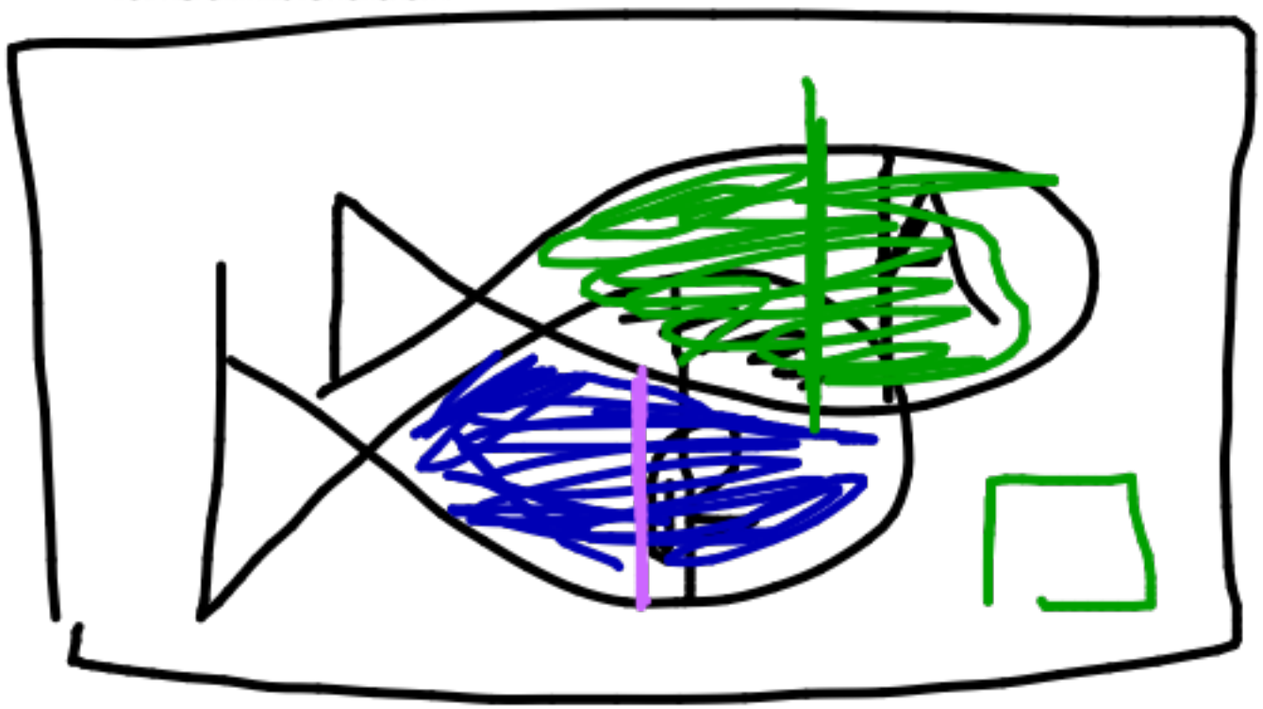
Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.



Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.
- Assume the samples in $\mathcal{D}_j$ have been drawn independently according to the (unknown but) fixed density $p(\mathbf{x}|\omega_j)$.
  - We say these samples are **i.i.d.** — independent and identically distributed.



Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.
- Assume the samples in $\mathcal{D}_j$ have been drawn independently according to the (unknown but) fixed density $p(\mathbf{x}|\omega_j)$.
  - We say these samples are **i.i.d.** — independent and identically distributed.
- Assume $p(\mathbf{x}|\omega_j)$ has some fixed parametric form and is fully described by $\boldsymbol{\theta}_j$; hence we write $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$.

Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.
- Assume the samples in $\mathcal{D}_j$ have been drawn independently according to the (unknown but) fixed density $p(\mathbf{x}|\omega_j)$.
  - We say these samples are **i.i.d.** — independent and identically distributed.
- Assume $p(\mathbf{x}|\omega_j)$ has some fixed parametric form and is fully described by $\boldsymbol{\theta}_j$; hence we write $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$.
- We thus have c separate problems of the form:

Definition

Use a set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of training samples drawn independently from the density $p(\mathbf{x}|\boldsymbol{\theta})$ to estimate the unknown parameter vector $\boldsymbol{\theta}$.

(Log-)Likelihood

- Because we assume i.i.d. we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \ . \tag{2}$$

(Log-)Likelihood

- Because we assume i.i.d. we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \quad . \tag{2}$$

- The log-likelihood is typically easier to work with both analytically and numerically.

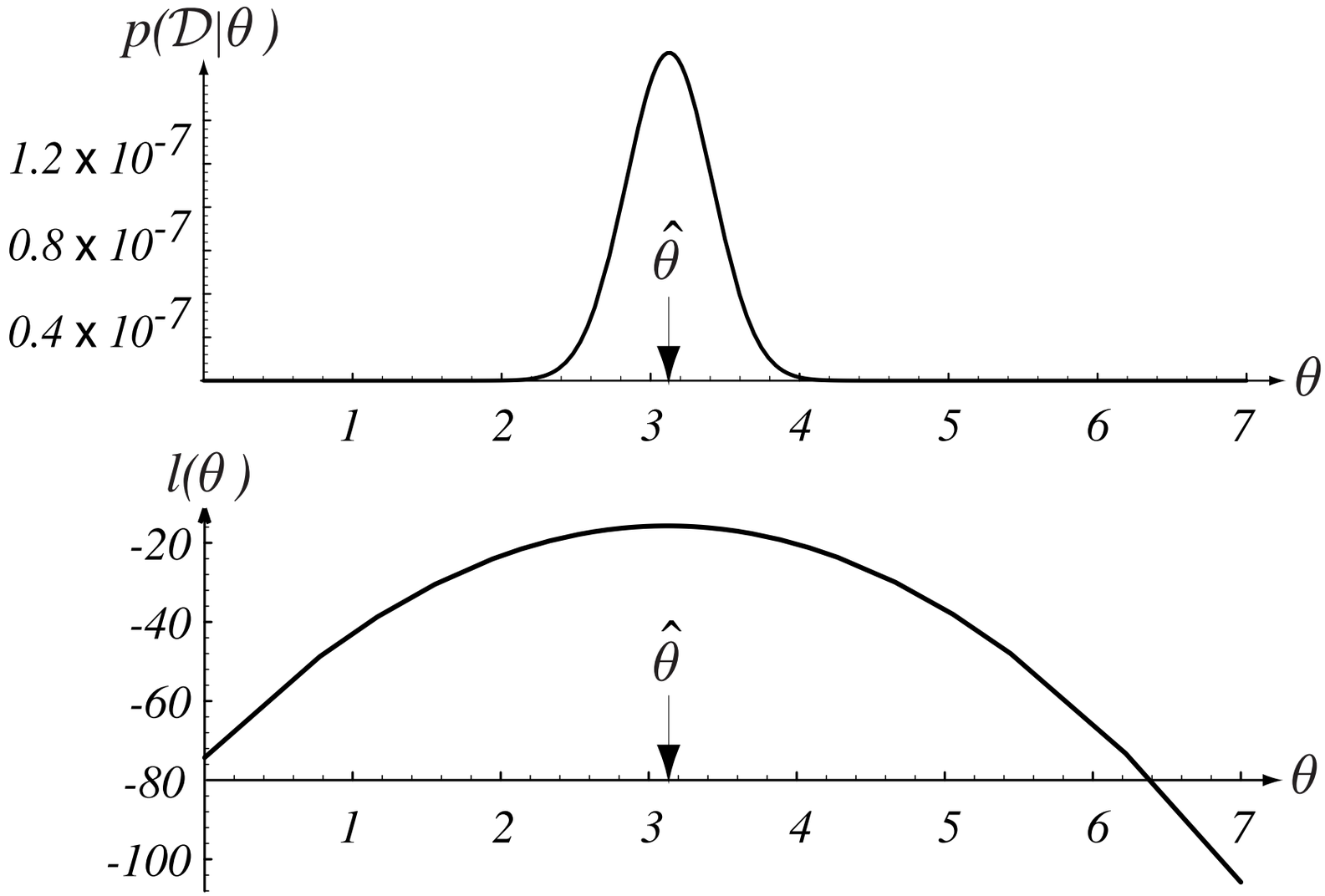
$$l_{\mathcal{D}}(\boldsymbol{\theta}) \equiv l(\boldsymbol{\theta}) \doteq \ln p(\mathcal{D}|\boldsymbol{\theta}) \tag{3}$$

$$= \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{4}$$

Maximum (Log-)Likelihood

- The **maximum likelihood estimate** of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$ or equivalently maximizes $l_{\mathcal{D}}(\boldsymbol{\theta})$.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l_{\mathcal{D}}(\boldsymbol{\theta}) \tag{5}$$



Necessary Conditions for MLE

- For p parameters, $\boldsymbol{\theta} \doteq [\theta_1 \quad \theta_2 \quad \dots \quad \theta_p]^{\mathsf{T}}$.
- Let $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ be the gradient operator, then $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \doteq \left[\frac{\partial}{\partial\theta_1} \quad \dots \quad \frac{\partial}{\partial\theta_p}\right]^{\mathsf{T}}$.
- The set of **necessary conditions** for the maximum likelihood estimate of $\boldsymbol{\theta}$ are obtained from the following system of p equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = 0 \tag{6}$$

Necessary Conditions for MLE

- For p parameters, $\boldsymbol{\theta} \doteq \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_p \end{bmatrix}^{\mathsf{T}}$.
- Let $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ be the gradient operator, then $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \doteq \begin{bmatrix} \frac{\partial}{\partial \theta_1} & \dots & \frac{\partial}{\partial \theta_p} \end{bmatrix}^{\mathsf{T}}$.
- The set of **necessary conditions** for the maximum likelihood estimate of $\boldsymbol{\theta}$ are obtained from the following system of p equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = 0 \tag{6}$$

- A solution $\hat{\boldsymbol{\theta}}$ to (6) can be a true global maximum, a local maximum or minimum or an inflection point of $l(\boldsymbol{\theta})$.

Necessary Conditions for MLE

- For p parameters, $\boldsymbol{\theta} \doteq \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_p \end{bmatrix}^{\mathsf{T}}$.
- Let $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ be the gradient operator, then $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \doteq \begin{bmatrix} \frac{\partial}{\partial \theta_1} & \dots & \frac{\partial}{\partial \theta_p} \end{bmatrix}^{\mathsf{T}}$.
- The set of **necessary conditions** for the maximum likelihood estimate of $\boldsymbol{\theta}$ are obtained from the following system of p equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = 0 \tag{6}$$

- A solution $\hat{\boldsymbol{\theta}}$ to (6) can be a true global maximum, a local maximum or minimum or an inflection point of $l(\boldsymbol{\theta})$.
- Keep in mind that $\hat{\boldsymbol{\theta}}$ is only an estimate. Only in the limit of an infinitely large number of training samples can we expect it to be the true parameters of the underlying density.

Gaussian Case with Known Σ and Unknown μ

- For a single sample point $\mathbf{x}_k$:

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) \quad (7)$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) \quad (8)$$

$$p(x_k|\mu) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}\,\star\right)$$

Gaussian Case with Known Σ and Unknown μ

- For a single sample point $\mathbf{x}_k$:

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{7}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{8}$$

- We see that the ML-estimate must satisfy

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0 \tag{9}$$

Gaussian Case with Known Σ and Unknown μ

- For a single sample point $\mathbf{x}_k$:

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{7}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{8}$$

- We see that the ML-estimate must satisfy

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0 \tag{9}$$

- And we get the sample mean!

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \tag{10}$$

Univariate Gaussian Case with Unknown μ and σ^2

The Log-Likelihood

- Let $\boldsymbol{\theta} = (\mu, \sigma^2)$. The log-likelihood of x_k is

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}(x_k - \mu)^2 \tag{11}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2}(x_k - \mu) \\ -\frac{1}{2\sigma^2} + \frac{(x_k-\mu)^2}{2\sigma^2} \end{bmatrix} \tag{12}$$

Univariate Gaussian Case with Unknown μ and σ^2

Necessary Conditions

- The following conditions are defined:

$$\sum_{k=1}^{n} \frac{1}{\hat{\sigma}^2}(x_k - \hat{\mu}) = 0 \tag{13}$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\sigma}^2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\mu})^2}{\hat{\sigma}^2} = 0 \tag{14}$$

Univariate Gaussian Case with Unknown μ and σ^2

ML-Estimates

- After some manipulation we have the following:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{15}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \tag{16}$$

- These are encouraging results – even in the case of unknown μ and σ^2 the ML-estimate of μ corresponds to the sample mean.

Bias

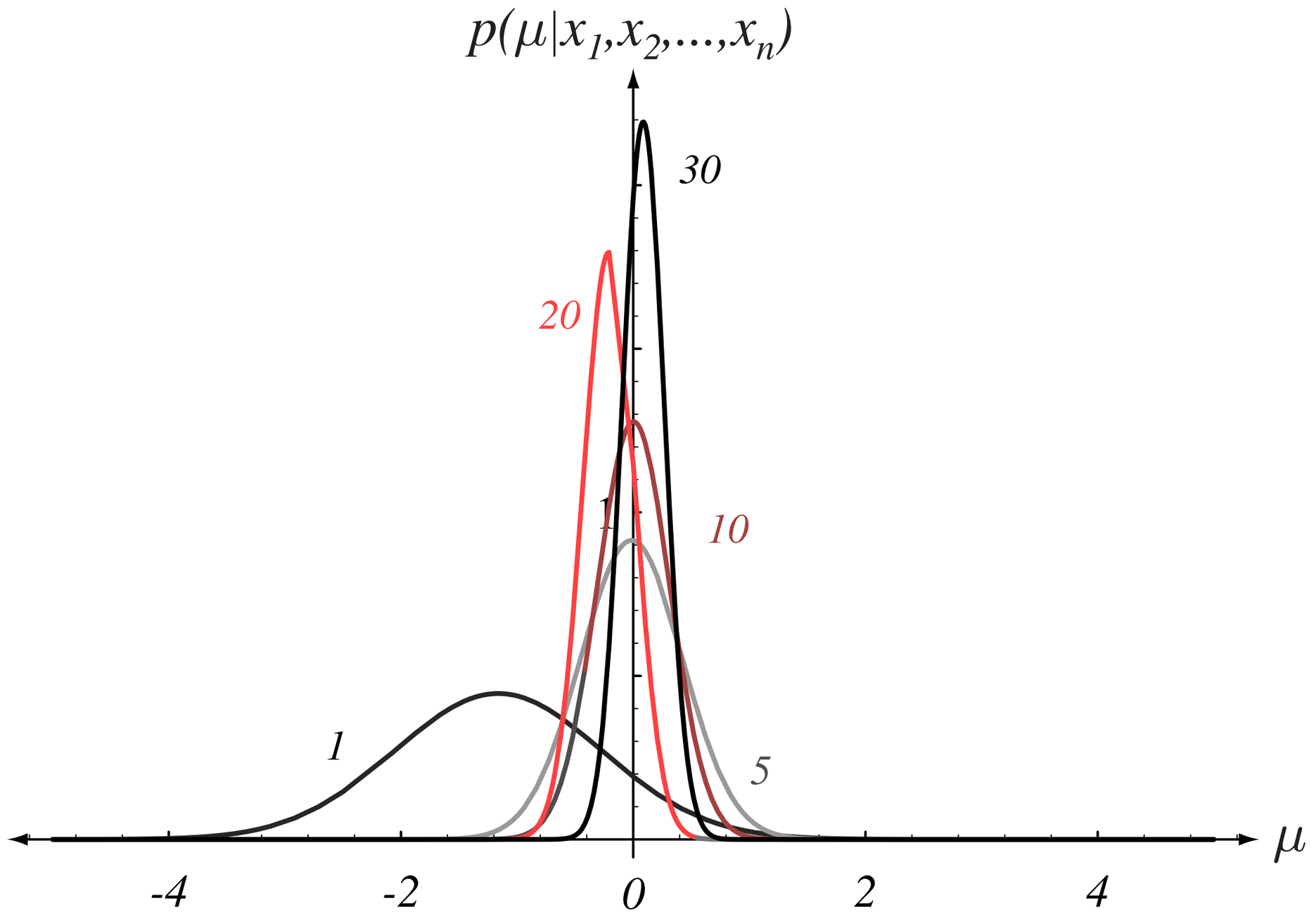
- The maximum likelihood estimate for the variance σ^2 is **biased**.
- The expected value over datasets of size n of the sample variance is not equal to the true variance

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \tag{17}$$

- In other words, the ML-estimate of the variance systematically underestimates the variance of the distribution.
- As $n \to \infty$ the problem of bias is reduced or removed, but bias remains a problem of the ML-estimator.
- An unbiased ML-estimator of the variance is

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \tag{18}$$

Bayesian Parameter Estimation Intuition



General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.
- Our initial knowledge about the parameters is summarized in a prior distribution $p(\boldsymbol{\theta})$.

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.
- Our initial knowledge about the parameters is summarized in a prior distribution $p(\boldsymbol{\theta})$.
- The rest of our knowledge about $\boldsymbol{\theta}$ is contained in a set $\mathcal{D}$ of n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn according to fixed $p(\mathbf{x})$.

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.
- Our initial knowledge about the parameters is summarized in a prior distribution $p(\boldsymbol{\theta})$.
- The rest of our knowledge about $\boldsymbol{\theta}$ is contained in a set $\mathcal{D}$ of n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn according to fixed $p(\mathbf{x})$.

Goal

Our ultimate goal is to estimate $p(\mathbf{x}|\mathcal{D})$, which is as close as we can come to estimating the unknown $p(\mathbf{x})$.

Linking Likelihood and the Parameter Distribution

- How do we relate the prior distribution on the parameters to the samples?

Linking Likelihood and the Parameter Distribution

- How do we relate the prior distribution on the parameters to the samples?
- **Missing Data!** The samples will convert our prior $p(\boldsymbol{\theta})$ to a posterior $p(\boldsymbol{\theta}|\mathcal{D})$, by integrating the joint density over $\boldsymbol{\theta}$:

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{19}$$

$$= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{20}$$

Linking Likelihood and the Parameter Distribution

- How do we relate the prior distribution on the parameters to the samples?
- **Missing Data!** The samples will convert our prior $p(\boldsymbol{\theta})$ to a posterior $p(\boldsymbol{\theta}|\mathcal{D})$, by integrating the joint density over $\boldsymbol{\theta}$:

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{19}$$

$$= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{20}$$

- And, because the distribution of $\mathbf{x}$ is known given the parameters $\boldsymbol{\theta}$, we simplify to

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{21}$$

Linking Likelihood and the Parameter Distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

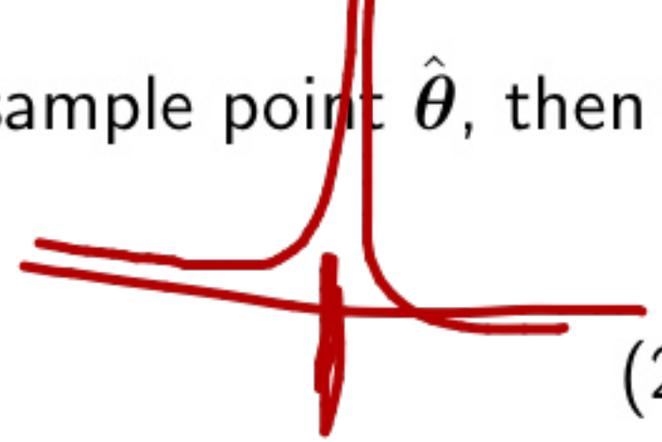
- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.

Linking Likelihood and the Parameter Distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply for sample point $\hat{\boldsymbol{\theta}}$, then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \ . \tag{22}$$



Linking Likelihood and the Parameter Distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply for sample point $\hat{\boldsymbol{\theta}}$, then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \quad . \tag{22}$$

- And, we will see that during Bayesian parameter estimation, the distribution over the parameters will get increasingly "peaky" as the number of samples increases.

Linking Likelihood and the Parameter Distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply for sample point $\hat{\boldsymbol{\theta}}$, then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \quad . \tag{22}$$

- And, we will see that during Bayesian parameter estimation, the distribution over the parameters will get increasingly "peaky" as the number of samples increases.
- What if the integral is not readily analytically computed?

The Posterior Density on the Parameters

- The primary task in Bayesian Parameter Estimation is the computation of the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$.
- By Bayes formula

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{23}$$

- Z is a normalizing constant:

$$Z = \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{24}$$

The Posterior Density on the Parameters

- The primary task in Bayesian Parameter Estimation is the computation of the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$.
- By Bayes formula

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{23}$$

- Z is a normalizing constant:

$$Z = \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{24}$$

- And, by the independence assumption on $\mathcal{D}$:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{25}$$

- Let's see an example now.