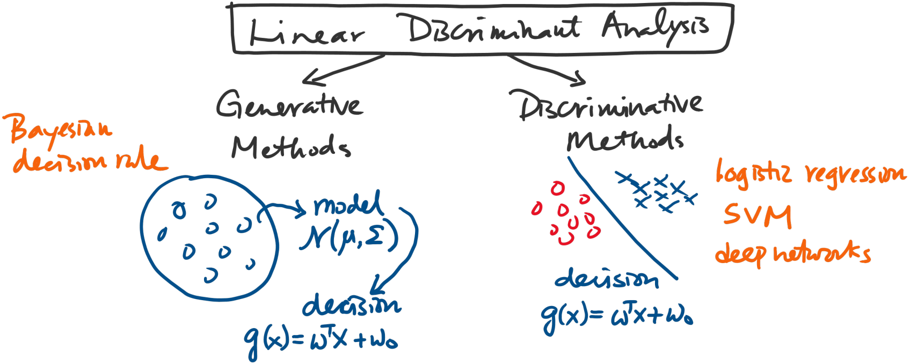
# ECE595 / STAT598: Machine Learning I
# Lecture 12 Bayesian Parameter Estimation

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

# Overview




- In linear discriminant analysis (LDA), there are generally two types of approaches
- **Generative approach**: Estimate model, then define the classifier
- **Discriminative approach**: Directly define the classifier

# Outline

**Generative Approaches**

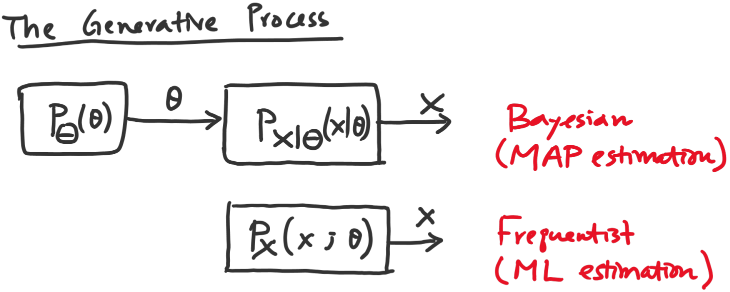
- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

**Today's Lecture**

- Basic Principles
  - Posterior
  - 1D Illustration
  - Interpretations
- Choosing Priors
  - Prior for Mean
  - Prior for Variance
  - Conjugate Prior

# MLE and MAP

There are two typical ways of estimating parameters.



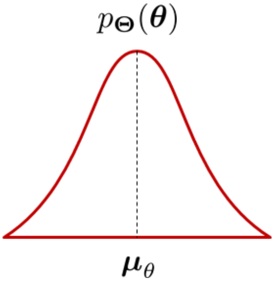

- Maximum-likelihood estimation (MLE): $\boldsymbol{\theta}$ is deterministic.
- Maximum-a-posteriori estimation (MAP): $\boldsymbol{\theta}$ is random and has a prior distribution.

# From MLE to MAP

- In MLE, the parameter $\boldsymbol{\theta}$ is **deterministic**.
- What if we assume $\boldsymbol{\theta}$ has a distribution?
- This makes $\boldsymbol{\theta}$ **probabilistic**.
- So make $\boldsymbol{\Theta}$ as a random variable, and $\boldsymbol{\theta}$ a state of $\boldsymbol{\Theta}$.
- Distribution of $\boldsymbol{\Theta}$:

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$$

- $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is the distribution of the parameter $\boldsymbol{\Theta}$.
- $\boldsymbol{\Theta}$ has its own mean and own variance.



# Maximum-a-Posteriori

By Bayes Theorem again:

$$p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}_n) = \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta})\,p_{\Theta}(\boldsymbol{\theta})}{p_{\boldsymbol{X}}(\boldsymbol{x}_n)}.$$

- To maximize the posterior distribution

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\mathcal{D}) \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \prod_{n=1}^{N} p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}_n) \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \prod_{n=1}^{N} \frac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta})\,p_{\Theta}(\boldsymbol{\theta})}{\cancel{p_{\boldsymbol{X}}(\boldsymbol{x}_n)}} \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad -\sum_{n=1}^{N} \left\{ \log p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log p_{\Theta}(\boldsymbol{\theta}) \right\}
\end{aligned}$$

# The Role of $p_{\Theta}(\theta)$

- Let's look at the MAP:

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \quad -\sum_{n=1}^{N} \left\{ \log p_{X|\Theta}(x_n|\theta) + \log p_{\Theta}(\theta) \right\}$$

- Special case: When

$$p_{\Theta}(\theta) = \delta(\theta - \theta_0).$$

- Then the delta function gives

$$\log p_{\Theta}(\theta) = \begin{cases} -\infty, & \text{if} \quad \theta \neq \theta_0, \\ 0, & \text{if} \quad \theta = \theta_0. \end{cases}$$

- This will give

$$\widehat{\theta} = \underset{\theta}{\text{argmin}} \quad -\sum_{n=1}^{N} \left\{ \begin{matrix} -\infty, & \text{if} \quad \theta \neq \theta_0, \\ \log p_{X|\Theta}(x_n|\theta_0), & \text{if} \quad \theta = \theta_0. \end{matrix} \right\} = \theta_0.$$

- No uncertainty. Absolutely sure $\theta = \theta_0$.

# Illustration: 1D Example

Suppose that:

$$p_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}$$

$$p_{\Theta}(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}.$$

**When $N = 1$.** The MAP problem is simply

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmax}} \quad p_{X|\Theta}(x|\theta) p_{\Theta}(\theta) \\
&= \underset{\theta}{\text{argmax}} \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} \\
&= \underset{\theta}{\text{argmax}} \quad -\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\theta_0)^2}{2\sigma_0^2}
\end{aligned}$$

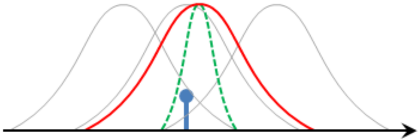
# Illustration: 1D Example

Taking derivatives:

$$
\begin{aligned}
& \frac{d}{d\theta}\left\{-\frac{(x-\theta)^2}{2\sigma^2}-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} &&= 0 \\
\Rightarrow\quad & \frac{(x-\theta)}{\sigma^2}-\frac{(\theta-\theta_0)}{\sigma_0^2} &&= 0 \\
\Rightarrow\quad & \sigma_0^2(x-\theta) &&= \sigma^2(\theta-\theta_0) \\
\Rightarrow\quad & \sigma_0^2 x+\sigma^2\theta_0 &&= (\sigma_0^2+\sigma^2)\theta
\end{aligned}
$$

Therefore, the solution is

$$\theta = \frac{\sigma_0^2 x + \sigma^2\theta_0}{\sigma_0^2+\sigma^2}.$$



# Interpreting the Result

Let us interpret the result

$$\theta = \frac{\sigma_0^2 x + \sigma^2 \theta_0}{\sigma_0^2 + \sigma^2}.$$

Does it make sense?

- If $\sigma_0 = 0$, then $\theta = \frac{\cancel{\sigma_0^2 x} + \sigma^2\theta_0}{\cancel{\sigma_0^2} + \sigma^2} = \theta_0$.
- This means: No uncertainty. Absolutely sure that $\theta = \theta_0$.
- $p_\Theta(\theta) = \delta(\theta - \theta_0)$

The other extreme

- If $\sigma_0 = \infty$, then $\theta = \frac{\sigma_0^2 x + \cancel{\sigma^2\theta_0}}{\sigma_0^2 + \cancel{\sigma^2}} = x$.
- This means: I don't trust my prior at all. Use data.
- $p_\Theta(\theta) = \frac{1}{|\Omega|}$, for all $\theta \in \Omega$.

Therefore, MAP solution gives you a trade-off between data and prior.

# When $N = 2$.

**When $N = 2$.** The MAP problem is

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmax}} \;\; p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta) p_\Theta(\theta) \\
&= \underset{\theta}{\text{argmax}} \;\; \left( \prod_{n=1}^{2} p_{X|\Theta}(x_n|\theta) \right) p_\Theta(\theta) \\
&= \underset{\theta}{\text{argmax}} \;\; \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \exp\left\{ -\frac{(x_1-\theta)^2 + (x_2-\theta)^2}{2\sigma^2} \right\} \\
&\qquad\qquad \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{(\theta-\theta_0)^2}{2\sigma_0^2} \right\} \\
&= \underset{\theta}{\text{argmin}} \;\; -\log( \;\; \cdot \;\; ) \\
&= \underset{\theta}{\text{argmin}} \;\; \left\{ \frac{(x_1-\theta)^2}{2\sigma^2} + \frac{(x_2-\theta)^2}{2\sigma^2} + \frac{(\theta-\theta_0)^2}{2\sigma_0^2} \right\}
\end{aligned}$$

# When $N = 2$.

Taking derivatives and setting to zero

$$\frac{d}{d\theta}\left\{\frac{(x_1-\theta)^2}{2\sigma^2}+\frac{(x_2-\theta)^2}{2\sigma^2}+\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\}$$
$$=-\frac{x_1-\theta}{\sigma^2}-\frac{x_2-\theta}{\sigma^2}+\frac{\theta-\theta_0}{\sigma_0^2}=0.$$

Equating to zero yields

$$\theta=\frac{(x_1+x_2)\sigma_0^2+\theta_0\sigma^2}{2\sigma_0^2+\sigma^2}.$$

- If $\sigma_0 = 0$ (certain prior), then $\theta = \theta_0$.
- If $\sigma_0 = \infty$ (useless prior), then $\theta = \frac{x_1+x_2}{2}$.

# When $N$ is Arbitrary

**General** $N$.

$$
\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\text{argmax}} \quad \left[\prod_{n=1}^{N} p_{X|\Theta}(x_n|\theta)\right] p_{\Theta}(\theta) \\
&= \underset{\theta}{\text{argmax}} \quad \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N} \frac{(x_n-\theta)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} \\
&= \underset{\theta}{\text{argmin}} \quad \left\{\sum_{n=1}^{N} \frac{(x_n-\theta)^2}{2\sigma^2} + \frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right\} \\
&= \frac{\sigma_0^2 \sum_{n=1}^{N} x_n + \theta_0\sigma^2}{N\sigma_0^2 + \sigma^2}.
\end{aligned}
$$

What does it mean?

# Interpreting the MAP solution: $N \to \infty$

Let's do some algebra:

$$\widehat{\theta} = \frac{(\sum_{n=1}^N x_n)\sigma_0^2 + \theta_0\sigma^2}{N\sigma_0^2 + \sigma^2} = \frac{(\sum_{n=1}^N x_n)\sigma_0^2 + \theta_0\sigma^2}{N(\sigma_0^2 + \frac{\sigma^2}{N})}$$

$$= \frac{\left(\frac{1}{N}\sum_{n=1}^N x_n\right)\sigma_0^2 + \frac{\sigma^2}{N}\theta_0}{\sigma_0^2 + \frac{\sigma^2}{N}}.$$

- Fix $\sigma_0$ and $\sigma$
- As $N \to \infty$,

$$\widehat{\theta} = \frac{\left(\frac{1}{N}\sum_{n=1}^N x_n\right)\sigma_0^2 + \cancel{\frac{\sigma^2}{N}\theta_0}}{\sigma_0^2 + \cancel{\frac{\sigma^2}{N}}} = \frac{1}{N}\sum_{n=1}^N x_n$$

- This is the maximum-likelihood estimate.
- When I have a lot of samples, the prior does not really matter.

# Interpreting the MAP solution: $N \to 0$

$$\widehat{\theta} = \frac{\left(\frac{1}{N}\sum_{n=1}^N x_n\right)\sigma_0^2 + \frac{\sigma^2}{N}\theta_0}{\sigma_0^2 + \frac{\sigma^2}{N}}$$

- Fix $\sigma_0$ and $\sigma$
- As $N \to 0$,

$$\widehat{\theta} = \frac{\cancel{\left(\frac{1}{N}\sum_{n=1}^N x_n\right)\sigma_0^2} + \frac{\sigma^2}{N}\theta_0}{\cancel{\sigma_0^2} + \frac{\sigma^2}{N}} = \theta_0.$$

- This is just the prior.
- When I have very few sample, I should rely on the prior.
- If the prior is good, then I can do well even if I have very few samples.
- Maximum-likelihood does not have the same luxury!

# What Happens to the Posterior?

Consider

$$p(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}.$$

We can show that the posterior is

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left\{-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right\},$$

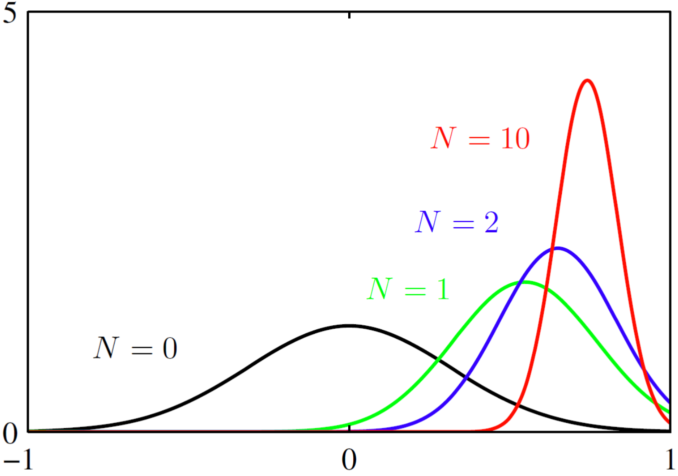
where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}$$

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

# What Happens to the Posterior?

- $x_n$ are generated from $\mu = 0.8$ and $\sigma^2 = 0.1$.
- When $N$ increases, the posterior shifts towards the true distribution.



# Outline

**Generative Approaches**

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

**Today's Lecture**

- Basic Principles
    - Posterior
    - 1D Illustration
    - Interpretations
- Choosing Priors
    - Prior for Mean
    - Prior for Variance
    - Conjugate Prior

# Prior for $\mu$

So far we have considered

$$p(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}.$$

- Unknown $\mu$, and known $\sigma^2$.
- The likelihood is Gaussian (by problem setup).
- The prior for $\mu$ is Gaussian (by our choice).
- Good, because posterior remains a Gaussian.
- What happens if $\sigma^2$ is unknown but $\mu$ is known?

# Prior for $\sigma^2$

- Let us define the precision: $\lambda = \frac{1}{\sigma^2}$.
- The likelihood is

$$
\begin{aligned}
p(\mathcal{D}|\lambda) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{\lambda^{N/2}}{(\sqrt{2\pi})^N}\right) \exp\left\{-\sum_{n=1}^{N} \frac{\lambda}{2}(x_n-\mu)^2\right\} \\
&= \frac{1}{(2\pi)^{N/2}}\, \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.
\end{aligned}
$$

- We want to choose $p(\lambda)$ in a similar form:

$$
p(\lambda) = A\lambda^B \exp\{-C\lambda\}
$$

so that the posterior $p(\lambda|\mathcal{D})$ is easy to compute.

# Prior for $\sigma^2$

- We want to choose $p(\lambda)$ in a similar form:

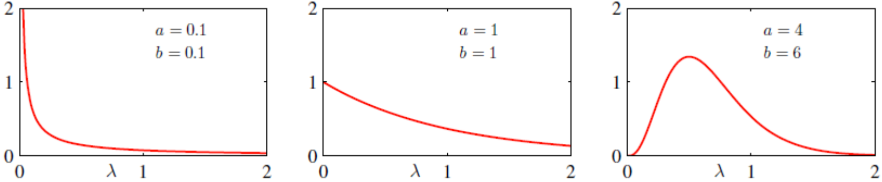
$$p(\lambda) = A\lambda^B \exp\{-C\lambda\}$$

- The candidate is ...

$$p(\lambda) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

- This distribution is called the Gamma distribution $\mathrm{Gam}(\lambda|a, b)$.
- We can show that

$$\mathbb{E}[\lambda] = \frac{a}{b}, \qquad \mathrm{Var}[\lambda] = \frac{a}{b^2}.$$



# Prior for $\sigma^2$

- If we consider this pair of likelihood and prior

$$p(\mathcal{D}|\lambda) = \frac{1}{(2\pi)^{N/2}} \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

$$p(\lambda) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0-1} \exp(-b_0\lambda),$$

- then the posterior is

$$p(\lambda|\mathcal{D}) \propto \lambda^{(a_0+N/2)-1} \exp\left\{-\left(b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)\lambda\right\}$$

- Just another Gamma distribution.
- You can now do estimation on this Gamma by finding $\lambda$ which maximizes the posterior. Details: See Appendix.

# Prior for Both $\mu$ and $\sigma^2$

- Again, let $\lambda = \frac{1}{\sigma^2}$.
- The likelihood is

$$p(\mathcal{D}|\mu,\lambda) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}$$

- Candidate for the prior is

$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^{\beta} \exp\left\{c\lambda\mu - d\lambda\right\}$$

$$= \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu-c/\beta)^2\right\}}_{\mathcal{N}(\mu|\mu_0,\sigma_0^2)}\underbrace{\lambda^{\beta/2}\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\}}_{\mathrm{Gam}(\lambda|a,b)}$$

- The prior distribution is called the Normal-Gamma distribution.

# Priors for High-dimension Gaussians

- Let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$.
- The likelihood is

$$p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}_m|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

- Prior for $\boldsymbol{\mu}$: Gaussian.

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}).$$

- Prior for $\boldsymbol{\Sigma}$: Wishart.

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu).$$

- Prior for both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$: Normal-Wishart.

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu).$$

# Conjugate Prior

- You have a likelihood $p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\boldsymbol{\theta})$
- You want to choose a prior $p_{\Theta}(\boldsymbol{\theta})$ so that ...
- the posterior $p_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$ takes the same form as the prior
- Such prior is called the **conjugate prior**
- Conjugate *with respect to* the likelihood
- Finding the conjugate prior may not be easy!
- Good news: Any likelihood belong to the **exponential family** will have a conjugate prior also in the exponential family.
- Exponential family: Gaussian, Exponential, Poisson, Bernoulli, etc
- For more discussions, see Bishop Chapter 2.4

# Reading List

**Bayesian Parameter Estimation**

- Duda-Hart-Stork, Pattern Classification, Chapter 3.3 - 3.5
- Bishop, Pattern Recognition and Machine Learning, Chapter 2.4
- M. Jordan (Berkeley), https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter9.pdf
- CMU Note, http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/slides/MLE_MAP_Part1.pdf
- A. Kak (Purdue), https://engineering.purdue.edu/kak/Tutorials/Trinity.pdf

# Appendix

# Prior for $\sigma^2$: Solution

- The posterior is

$$p(\lambda|\mathcal{D}) \propto \lambda^{(a_0+N/2)-1} \exp\left\{-\left(b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)\lambda\right\}$$
$$\propto \lambda^{a_N-1}\exp\left\{-b_N\lambda\right\}.$$

- The maximum-a-posteriori estimate of $\lambda$ is

$$\widehat{\lambda} = \underset{\lambda}{\text{argmax}} \;\; p(\lambda|\mathcal{D})$$
$$= \underset{\lambda}{\text{argmax}} \;\; \lambda^{a_N-1}\exp\left\{-b_N\lambda\right\}$$
$$= \underset{\lambda}{\text{argmax}} \;\; (a_N-1)\log\lambda - b_N\lambda.$$

- Taking derivative and setting to zero:

$$\frac{d}{d\lambda}\Big((a_N-1)\log\lambda - b_N\lambda\Big) = \frac{a_N-1}{\lambda} - b_N = 0.$$

# Prior for $\sigma^2$: Solution

- Therefore,

$$\lambda = \frac{a_N - 1}{b_N}.$$

- where the parameters are

$$a_N = a_0 + \frac{N}{2},$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\text{ML}}.$$

- Hence, the MAP estimate is

$$\lambda = \frac{a_0 + \frac{N}{2}}{b_0 + \frac{N}{2}\sigma^2_{\text{ML}}}.$$

- As $N \to \infty$, $\lambda \to \frac{1}{\sigma^2_{\text{ML}}}$.
- As $N \to 0$, $\lambda \to \frac{a_0}{b_0}$.

# Prior for Both $\mu$ and $\sigma^2$: Detailed Derivation

- Again, let $\lambda = \frac{1}{\sigma^2}$.
- The likelihood is

$$
\begin{aligned}
p(\mathcal{D}|\mu, \lambda) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \\
&= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n^2 - 2\mu x_n + \mu^2)\right\} \\
&= \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N} x_n^2 + \lambda\mu\sum_{n=1}^{N} x_n\right\} \left[\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^N \\
&= \left(\frac{1}{2\pi}\right)^{N/2} \left[\lambda^{1/2} \exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N} x_n - \frac{\lambda}{2}\sum_{n=1}^{N} x_n^2\right\}
\end{aligned}
$$

# Prior for Both $\mu$ and $\sigma^2$: Detailed Derivation

- The likelihood is

$$p(\mathcal{D}|\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^{N} \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}$$

- Candidate for the prior is

$$\begin{aligned} p(\mu,\lambda) &\propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^{\beta} \exp\left\{c\lambda\mu - d\lambda\right\} \\ &= \left[\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^{\beta}\left[\lambda^{\beta/2}\exp\left\{c\lambda\mu - d\lambda\right\}\right] \\ &= \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}}_{\mathcal{N}(\mu|\mu_0,\sigma_0^2)}\underbrace{\lambda^{\beta/2}\exp\left\{-\left(d - \frac{c^2}{2\beta}\right)\lambda\right\}}_{\text{Gam}(\lambda|a,b)} \end{aligned}$$

- 

$$\mu_0 = c/\beta, \quad \sigma_0^2 = (\beta\lambda)^{-1}, \quad a = 1 + \beta/2, \quad b = d - c^2/2\beta$$

# Prior for Both $\mu$ and $\sigma^2$: Detailed Derivation

- The prior distribution is

$$p(\mu, \lambda) = \mathcal{N}\left(\mu | \mu_0, (\beta\lambda)^{-1}\right) \mathrm{Gam}(\lambda | a, b)$$

- This is called the Normal-Gamma distribution
- Here is a 2D plot of $p(\mu, \lambda)$ when $\mu_0 = 0$, $\beta = 2$, $a = 5$, $b = 6$.