# Probability Review for Machine Learning

Chris J. Maddison[1]

University of Toronto

[1]Slides adapted from CSC 311.

# Motivation

Uncertainty comes from:

- Noisy measurements
- Variability between samples
- Finite size of data sets

Probability theory provides a consistent formalism for the quantification and manipulation of uncertainty.

# Sample Space

The data comes from a measurement of the real world, which we can think of as an experiment:

- Sample space $\Omega$ is the set of all possible outcomes of the experiment.
- Observations $\omega \in \Omega$ are points in the space also called sample outcomes, realizations, or elements.
- Events $E \subset \Omega$ are subsets of the sample space.

For example, if we flip a coin twice:

Sample space All outcomes $\Omega = \{HH, HT, TH, TT\}$

Observation $\omega = HT$ valid sample since $\omega \in \Omega$

Event Both flips same $E = \{HH, TT\}$ valid event since $E \subset \Omega$

# Probability

The probability of an event E, $P(E)$, satisfies three axioms:

1: $P(E) \geq 0$ for every $E$

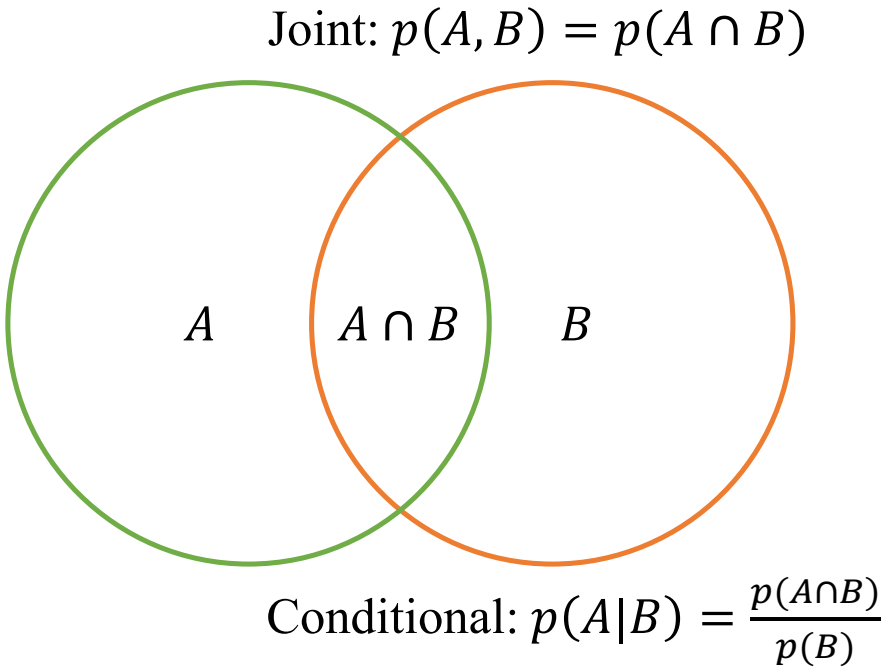
2: $P(\Omega) = 1$

3: If $E_1, E_2, \ldots$ are disjoint then

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

# Joint and Conditional Probabilities

Joint Probability of $A$ and $B$ is denoted $P(A, B)$.
Conditional Probability of $A$ given $B$ is denoted $P(A|B)$.




$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

# Conditional Example

Probability of passing the midterm is 60% and probability of passing both the final and the midterm is 45%.

What is the probability of passing the final given the student passed the midterm?

$$\begin{aligned} P(F|M) &= P(M,F)/P(M) \\ &= 0.45/0.60 \\ &= 0.75 \end{aligned}$$

# Independence

Events $A$ and $B$ are independent if $P(A, B) = P(A)P(B)$.

Suppose $P(A) = P(B) = 0.5$.

- Independent: $A$: first toss is HEAD; $B$: second toss is HEAD;

$$P(A, B) = 0.5 * 0.5 = P(A)P(B)$$

- Not Independent: $A$: first toss is HEAD; $B$: first toss is HEAD;

$$P(A, B) = 0.5 \neq P(A)P(B)$$

# Independence

Events $A$ and $B$ are conditionally independent given $C$ if

$$P(A, B|C) = P(B|C)P(A|C) \tag{1}$$

Consider two coins [2]: A regular coin and a coin which always outputs HEAD or always outputs TAIL.

Now consider the following events.

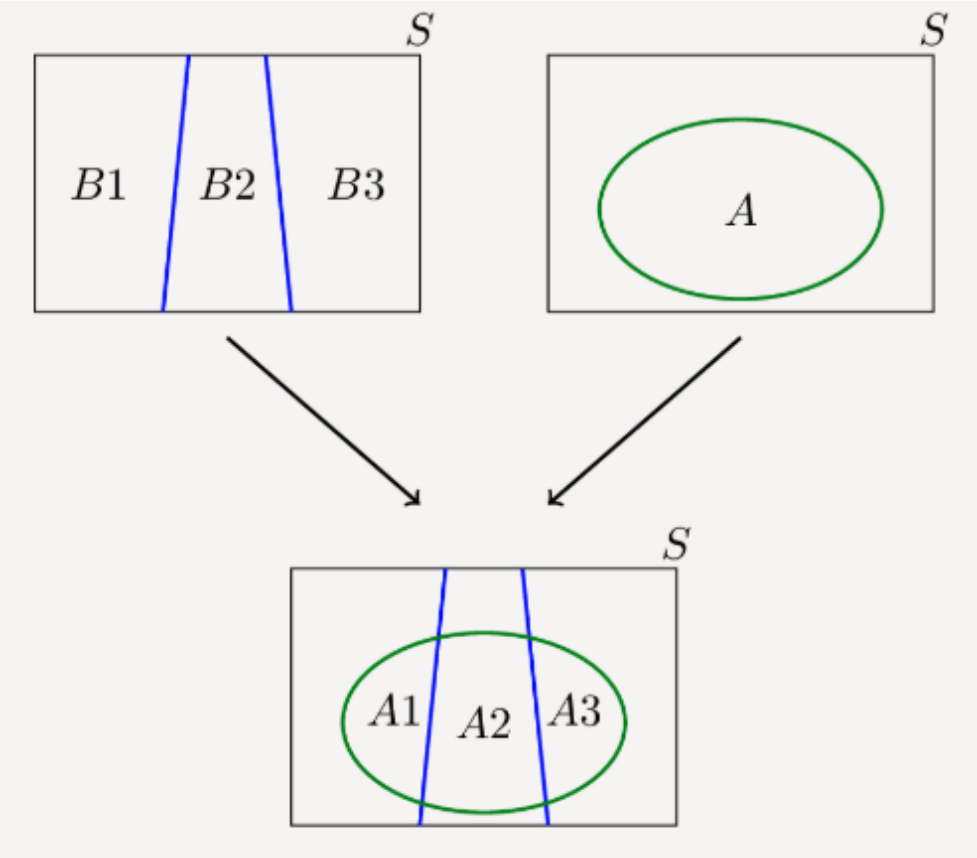
- $A$=The first toss is HEAD.
- $B$=The second toss is HEAD.
- $C$=The regular coin is used.
- $D$=The other coin is used.

Then $A$ and $B$ are conditionally independent given $C$, but $A$ and $B$ are NOT conditionally independent given $D$.

[2] `www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php`

# Marginalization and Law of Total Probability[3]

$$P(X) = \sum_Y P(X, Y) = \sum_Y P(X|Y)P(Y)$$




[3] www.probabilitycourse.com/chapter1/1_4_2_total_probability.php

# Bayes' Rule

Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}$$

# Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?
This depends on the prior probability of the disease:

- $P(T = 1|D = 1) = 0.95$ (likelihood)
- $P(T = 1|D = 0) = 0.10$ (likelihood)
- $P(D = 1) = 0.1$ (prior)

So $P(D = 1|T = 1) =$?

# Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

$P(T = 1|D = 1) = 0.95$ (true positive)

$P(T = 1|D = 0) = 0.10$ (false positive)

$P(D = 1) = 0.1$ (prior)

So $P(D = 1|T = 1) =?$
Use Bayes' Rule:

$$P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)} = \frac{0.95 * 0.1}{P(T = 1)} = 0.51$$

$$P(T = 1) = P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0)$$
$$= 0.95 * 0.1 + 0.1 * 0.90 = 0.185$$

# Random Variable

How do we connect sample spaces and events to data?

A random variable is a mapping which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$

For example, let's flip a coin 10 times. $X(\omega)$ counts the number of Heads we observe in our sequence. If $\omega = HHTHTHHTHT$ then $X(\omega) = 6$.

# Discrete and Continuous Random Variables

## Discrete Random Variables

- Takes countably many values, e.g., number of heads
- Distribution defined by probability mass function (PMF) $p_X$.
- The probability that $X(\omega) \in B$ is given by:

$$P(X(\omega) \in B) = \sum_{x \in B} p_X(x). \tag{2}$$

## Continuous Random Variables

- Takes uncountably many values, e.g., time to complete task
- Distribution defined by probability density function (PDF) $f_X$.
- The probability that $X(\omega) \in B$ is given by:

$$P(X(\omega) \in B) = \int_B f_X(x)\, dx \tag{3}$$

# Joint Distributions

Two random variables $X(\omega), Y(\omega)$ have a joint distribution.

- If both are discrete, then they have a joint PMF $p_{X,Y}(x, y)$.
- If both are continuous, then they have a joint PDF $f_{X,Y}(x, y)$.

We can marginalize the joint to get the marginal distributions of $X$ or $Y$:

- $\sum_y p_{X,Y}(x, y) = p_X(x)$.
- $\int f_{X,Y}(x, y)\, dx = f_Y(y)$.

$p_X$ and $f_Y$ are often just called marginals.

# Conditional Distributions

The conditional distribution of $X$ given $Y$ can be described using conditional PMF or PDF:

- If both are discrete, then the conditional PMF of $X$ given $Y$ is

$$p_{X \mid Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

- If both are continuous, then the conditional PMF of $X$ given $Y$ is

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

$X$ is independent of $Y$ if $p_{X \mid Y}(x \mid y) = p_X(x)$ or $f_{X \mid Y}(x \mid y) = f_X(x)$.

# I.I.D.

Random variables are said to be independent and identically distributed (i.i.d.) if they are sampled from the same probability distribution and are mutually independent.

This is a common assumption for observations. For example, coin flips are assumed to be iid.

# A Note on Notation

The machine learning discipline has a variety of different notation norms for probability, which range from very formal to very informal. We will lean informal, because it's efficient and important to learn the norms.

We use lower case for random variables, use $p$ for all PMFs / PDFs, and sometimes even omit the ranges of integration.

$$P(X(\omega) \in B) = \int_B f_X(x)\,dx$$

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

$$f_{X \mid Y}(x \mid y)$$

$\xrightarrow{\text{becomes}}$

$$P(x \in B) = \int_B p(x)\,dx$$

$$p(x,y) = p(x)p(y)$$

$$p(x \mid y)$$

This seems like a bit of a mess, but with practice it's always clear what is intended from context.

# Probability Distribution Statistics

Mean: First Moment, $\mu$

$$\mathbb{E}[x] = \sum_{i=1}^{\infty} x_i p(x_i) \qquad \text{(univariate discrete r.v.)}$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x p(x) dx \qquad \text{(univariate continuous r.v.)}$$

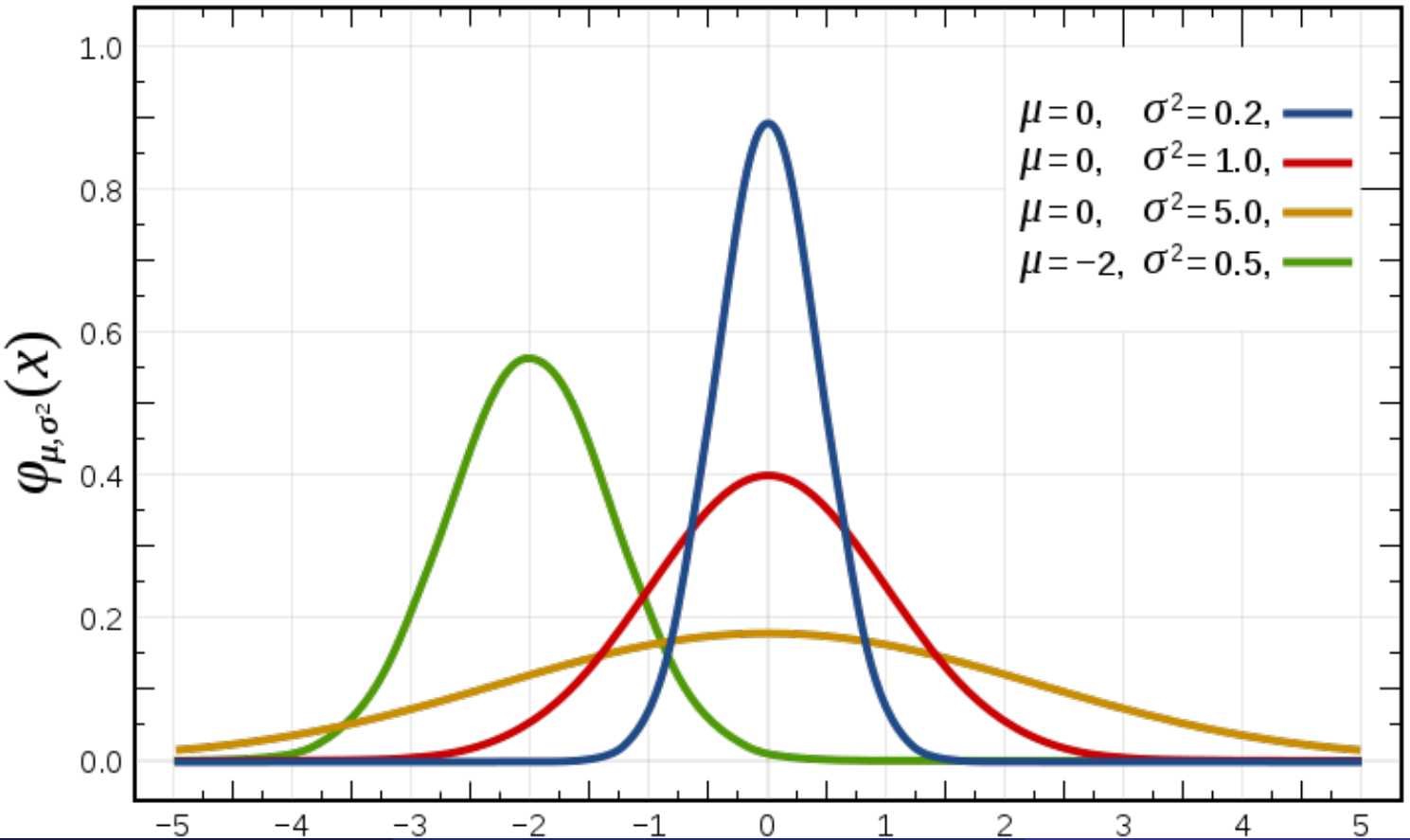
Variance: Second (central) Moment, $\sigma^2$

$$\begin{aligned} \text{Var}[x] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \\ &= \mathbb{E}\left[(x-\mu)^2\right] \\ &= \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 \end{aligned}$$

# Univariate Gaussian Distribution

Also known as the Normal Distribution, $\mathcal{N}(\mu, \sigma^2)$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$



# Mixed Discrete and Continuous

It is very common in machine learning to have a joint distribution over a mixture of discrete and continuous random variables. In this case, we can define the joint distribution in terms of marginals and conditionals.

Suppose $x$ is discrete and $y$ is continuous. Then we can define the joint distribution by

$$p(x, y) = p(x)p(y|x)$$

More specifically, suppose $x \in \{1, \ldots, n\}$ with probabilities given by $p_x$. Suppose also we have $n$ real numbers $\mu_i$, then we can define a joint distribution between $x$ and a continuous $y$ via

$$p(x, y) = p_x \mathcal{N}(y|\mu_x, 1)$$

# Multivariate Gaussian Distribution

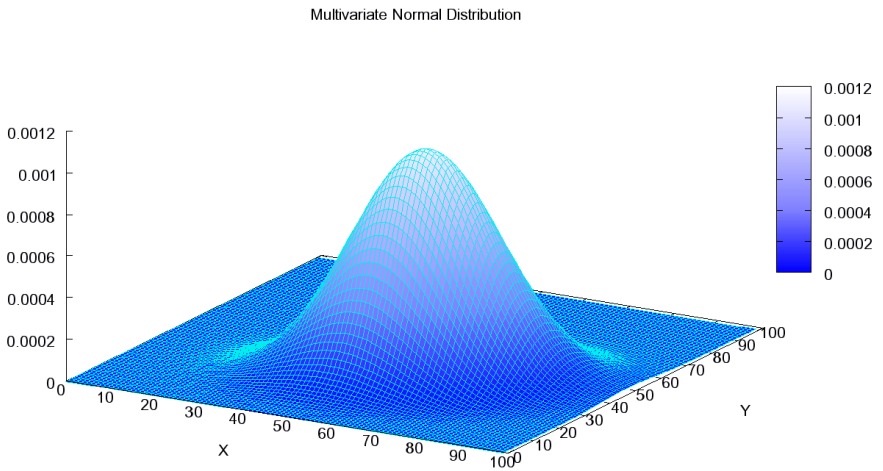
$\mathbf{x}$ is a D-dimensional vector

$\mu$ is a D-dimensional mean vector

$\Sigma$ is a $D \times D$ covariance matrix with determinant $|\Sigma|$.

When $\Sigma$ is invertible, the Gaussian has a density:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$



# Covariance Matrix

Mean $\boldsymbol{\mu}$ is the D-dimensional vector whose $i$ entry is the expected value

$$\begin{aligned}\boldsymbol{\mu}_i &= \mathbb{E}[\boldsymbol{x}_i] \\ &= \int \boldsymbol{x}_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)\, d\boldsymbol{x}\end{aligned}$$

Covariance matrix $\Sigma$ is a matrix whose $(i, j)$ entry is the covariance

$$\begin{aligned}\Sigma_{ij} &= Cov(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \mathbb{E}\left[(\boldsymbol{x}_i - \mu_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_j)\right] \\ &= \mathbb{E}\left[(\boldsymbol{x}_i \boldsymbol{x}_j)\right] - \boldsymbol{\mu}_i \boldsymbol{\mu}_j\end{aligned}$$

so notice that the diagonal entries are the variance of each elements.

The covariance matrix has the property that it is symmetric and positive-semidefinite.

# Inferring Parameters

We have data $\boldsymbol{x}$ and we assume it is sampled from some distribution $p(\boldsymbol{x}|\theta)$ with parameters $\theta$, which are unknown to us.

How do we figure out the $\theta$ that 'best' fit that distribution?

One classical approach is maximum likelihood estimation (MLE),

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log p(\boldsymbol{x}|\theta)$$

We will discuss this in more detail in the later part of the course. For now, just assume that this makes sense to do.

# MLE for Univariate Gaussian Distribution

For example, we are trying to infer the parameters for a Univariate Gaussian Distribution, mean ($\mu$) and variance ($\sigma^2$).

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

The likelihood that our observations $\boldsymbol{x} = (x_1,\ldots,x_N)$ were generated by a univariate Gaussian with parameters $\mu$ and $\sigma^2$ is

$$\text{Likelihood} = p(x_1 \ldots x_N|\mu,\sigma^2) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\}$$

In this case $\theta = (\mu,\sigma^2)$.

# MLE for Univariate Gaussian Distribution

For MLE we want to maximize this likelihood, which is difficult because it is represented by a product of terms

$$\text{Likelihood} = p(x_1 \dots x_N | \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\}$$

So we take the log of the likelihood so the product becomes a sum

$$\begin{aligned} \text{Log Likelihood} &= \log p(x_1 \dots x_N | \mu, \sigma^2) \\ &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\} \end{aligned}$$

Since log is monotonically increasing $\max L(\theta) = \max \log L(\theta)$

# MLE for Univariate Gaussian Distribution

The log Likelihood simplifies to

$$\begin{aligned}\mathcal{L}(\mu, \sigma) &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\} \\ &= -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\end{aligned}$$

Which we want to maximize. How?

# MLE for Univariate Gaussian Distribution

To maximize we take the derivatives, set equal to 0, and solve:

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Derivative w.r.t. $\mu$, set equal to 0, and solve for $\hat{\mu}$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \mu} = 0 \implies \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Therefore the $\hat{\mu}$ that maximizes the likelihood is the average of the data points.

Derivative w.r.t. $\sigma^2$, set equal to 0, and solve for $\hat{\sigma}^2$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$