

# Bayesian Learning: Gaussian Models

ELG 5218 - Uncertainty Evaluation in Engineering Measurements and Machine Learning

Miodrag Bolic

University of Ottawa

January 21, 2026

# Roadmap (40 slides, 3 hours)

- ➊ **Gaussian models fundamentals:** Why Gaussians? Central Limit Theorem, conjugacy.
- ➋ **Known variance case:** Precision additivity, conjugate priors, weighted averaging.
- ➌ **Online learning (sequential updates):** Recursive Bayesian filtering, streaming data.
- ➍ **Unknown variance:** Normal-Gamma conjugate family, Student- $t$  marginal, robustness.
- ➎ **Monte Carlo methods:** Posterior sampling, functions of parameters, practical inference.
- ➏ **Multivariate case:** Matrix algebra, sensor fusion, conditional independence.
- ➐ **Discussion:** Advanced topics, extensions, exercises.

# The Gaussian family: central to Bayesian inference

## Four reasons Gaussians dominate:

- **Universality (CLT):** Sums of iid random variables  $\rightarrow$  Gaussian (approximately).
- **Analytical tractability:** Closed-form posterior, predictive, marginals, conditionals.
- **Conjugacy:** Normal prior  $\times$  Normal likelihood  $\Rightarrow$  Normal posterior.
- **Optimization:** Gaussian assumption  $\rightarrow$  convex inference, efficient algorithms.

## Univariate density:

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

## Precision notation (cleaner for Bayesian updates):

$$\lambda = \frac{1}{\sigma^2} \quad \Rightarrow \quad p(x \mid \mu, \lambda) \propto \sqrt{\lambda} \exp\left(-\frac{\lambda(x - \mu)^2}{2}\right).$$

# Multivariate Gaussian: covariance and precision

For  $\mathbf{x} \in \mathbb{R}^d$ :

**Covariance form:**

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

**Precision form** (more convenient for updates):

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \quad \Rightarrow \quad p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{1/2} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

**Key insight: Precision combines additively**

In Bayesian updates, posterior precision = prior precision + data precision.  
This additive structure is the **engine** driving conjugate family simplicity.

# Covariance matrix properties (brief review)

Covariance between variables  $X_i$  and  $X_j$ :

$$\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])].$$

**Covariance matrix:**

$$\Sigma = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \begin{bmatrix} \sigma_1^2 & \text{Cov}(X_1, X_2) & \cdots \\ \text{Cov}(X_2, X_1) & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

**Correlation:**  $\rho_{ij} = \text{Cov}(X_i, X_j)/(\sigma_i\sigma_j) \in [-1, 1]$ .

## Why precision for Bayesian inference?

- Precisions in likelihood and prior **add directly**, not reciprocals.
- Posterior precision is always non-negative (automatic numerical stability).
- Computational benefits for high-dimensional problems.

## Setup: Univariate Gaussian, known $\sigma^2$

**Model:**  $x_1, \dots, x_n \mid \mu, \sigma^2 \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2$  known.

**Likelihood (in precision form):**

$$p(x_1, \dots, x_n \mid \mu) \propto \exp \left( -\frac{\lambda}{2} \sum_{i=1}^n (x_i - \mu)^2 \right), \quad \lambda = \frac{1}{\sigma^2}.$$

Using the sum-of-squares identity:

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2.$$

First term is data-independent, so:

$$p(x_1, \dots, x_n \mid \mu) \propto \exp \left( -\frac{n\lambda}{2} (\bar{x} - \mu)^2 \right).$$

**Interpretation:** Likelihood is Gaussian in  $\mu$  with precision  $n\lambda = n/\sigma^2$ .

# Uniform (non-informative) prior

**Prior:**  $p(\mu) \propto 1$  for  $\mu \in \mathbb{R}$  (improper, but posterior is proper).

**Posterior (via Bayes):**

$$p(\mu \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \mu) \cdot p(\mu) \propto \exp\left(-\frac{n\lambda}{2}(\bar{x} - \mu)^2\right).$$

Recognizing this as Gaussian:

$$\mu \mid x_1, \dots, x_n \sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right).$$

**In precision form:**

$$\lambda_n = \frac{n}{\sigma^2} = n\lambda, \quad \mu_n = \bar{x}.$$

## Key observations

- Posterior mean = sample mean  $\bar{x}$  (frequentist MLE).
- Posterior variance shrinks as  $1/n$ .
- Posterior precision grows linearly with sample size.

## Normal prior (conjugate case)

**Prior:**  $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$  with precision  $\lambda_0 = 1/\sigma_0^2$ .

**Posterior:** Also Gaussian with

$$\lambda_n = \lambda_0 + n\lambda, \quad \mu_n = \frac{\lambda_0\mu_0 + n\lambda\bar{x}}{\lambda_n}.$$

**Rewritten (cleaner form):**

$$\mu_n = \frac{\lambda_0\mu_0 + (n/\sigma^2)\bar{x}}{\lambda_0 + n/\sigma^2} = w\bar{x} + (1 - w)\mu_0,$$

where

$$w = \frac{\lambda_{\text{data}}}{\lambda_0 + \lambda_{\text{data}}} = \frac{n/\sigma^2}{\lambda_0 + n/\sigma^2}.$$

The golden rule: Precision Additivity

$$\lambda_{\text{posterior}} = \lambda_{\text{prior}} + \lambda_{\text{data}}$$

Posterior mean is precision-weighted average of prior and data.

# Intuition: Precision-weighted averaging

- **High data precision** ( $n/\sigma^2$  large):  $w \approx 1$ , posterior tracks data.
- **High prior precision** ( $\lambda_0$  large):  $w \approx 0$ , posterior tracks prior.
- **Competing uncertainties:** Weight determined by relative precisions.
- **Posterior variance:**  $\sigma_n^2 = 1/\lambda_n$ , monotonically shrinks.

## Example interpretation:

- Prior: expert opinion (uncertain,  $\sigma_0 = 5$ ).
- Data: 10 precise measurements (noisy,  $\sigma = 3$ ).
- Data precision  $0.2/\sigma^2 > 0.04 = \text{prior precision} \Rightarrow \text{data dominates}$ .

## Worked example: Internet speed measurement

**Context:** ISP claims average speed is 20 Mbps. You measure 5 times.

**Data:**  $x = \{15.77, 20.5, 8.26, 14.37, 21.09\}$ ,  $\bar{x} = 16.0$  Mbps.

**Known measurement noise:**  $\sigma = 5$  Mbps  $\Rightarrow \lambda = 0.04$ .

**Prior:**  $\mu_0 = 20$  Mbps,  $\sigma_0 = 5$  Mbps  $\Rightarrow \lambda_0 = 0.04$ .

**Posterior calculation:**

$$\lambda_n = 0.04 + 5 \cdot 0.04 = 0.24 \quad \Rightarrow \quad \sigma_n^2 = 4.17.$$

$$\mu_n = \frac{0.04 \cdot 20 + 0.2 \cdot 16}{0.24} = \frac{0.8 + 3.2}{0.24} = 16.67 \text{ Mbps.}$$

$$w = \frac{0.2}{0.24} = 0.833 \quad \Rightarrow \quad \text{Data dominates (5\times more precise).}$$

**Conclusion:** Posterior  $\Pr(\mu > 20 \mid \text{data}) \approx 0.05$  (ISP claim is dubious).

# Posterior for Internet speed measurement

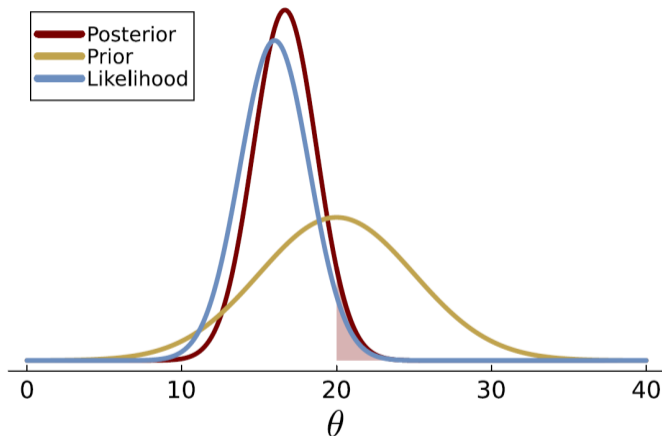


Figure: Prior-Likelihood-Posterior. The shaded region marks out  $Pr(\theta > 20 | x_1, \dots, x_n)$ .

# Sensor fusion example: Temperature + Humidity

**Scenario:** Two correlated measurements (temperature, humidity).

**True state:**  $\mu = \begin{bmatrix} 22C \\ 55\% \end{bmatrix}$ ,  $\Sigma = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 4 \end{bmatrix}$ .

**Data:**  $n = 15$  bivariate readings.

**Prior (weak):**  $\mu_0 = \begin{bmatrix} 20 \\ 50 \end{bmatrix}$ , independent prior on correlation.

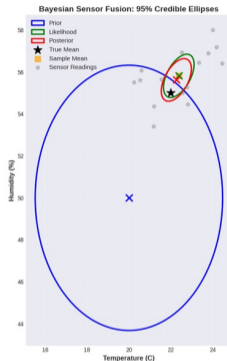
**Matrix precision additivity:**

$$\Lambda_n = \Lambda_0 + n\Sigma^{-1} = \text{prior precision} + \text{data precision}.$$

**Posterior:**

$$\Sigma_n = \Lambda_n^{-1}, \quad \mu_n = \Sigma_n(\Lambda_0\mu_0 + n\Sigma^{-1}\bar{x}).$$

# Posterior for the fusion example



## 2D visualization:

- 1 Prior: Ellipse centered at  $[20, 50]$  (weak, large).
- 2 Likelihood: Ellipse centered at sample mean (from 15 readings).
- 3 Posterior: Ellipse centered at  $\mu_n$  (tighter).
- 4 Correlation improvement (ellipse axes align with data).

# Sequential vs. batch updating

**Batch form:**

$$p(\mu \mid x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n \mid \mu)p(\mu)}{p(x_1, \dots, x_n)}.$$

**Sequential form:**

$$p(\mu \mid x_1, \dots, x_n) = \frac{p(x_n \mid \mu) \cdot p(\mu \mid x_1, \dots, x_{n-1})}{p(x_n \mid x_1, \dots, x_{n-1})}.$$

**Key insight:** Posterior at step  $n - 1$  becomes the prior at step  $n$ .

## Streaming advantage

- Never store all past data, only current posterior parameters.
- Each new observation  $x_n$  triggers one simple update.
- Perfect for sensor networks, real-time monitoring, online learning.
- Memory:  $O(d)$  vs.  $O(nd)$  for batch processing.

# Recursive updating: precision and mean

Starting from  $\mu \mid x_1, \dots, x_{n-1} \sim \mathcal{N}(\mu_{n-1}, \sigma_{n-1}^2)$ , when  $x_n$  arrives:

**Precision update:**

$$\lambda_n = \lambda_{n-1} + \lambda = \lambda_{n-1} + \frac{1}{\sigma^2}.$$

**Mean update (in learning rate form):**

$$\mu_n = \mu_{n-1} + w_n(x_n - \mu_{n-1}),$$

where the learning rate is

$$w_n = \frac{\lambda}{\lambda_n} = \frac{1/\sigma^2}{\lambda_{n-1} + 1/\sigma^2}.$$

## Interpretation

- $w_n$  is the fractional move toward new observation (shrinks over time).
- As  $n$  increases,  $\lambda_n$  grows, so  $w_n$  shrinks (diminishing returns).
- Posterior variance  $\sigma_n^2 = 1/\lambda_n$  monotonically decreases.

# Example: Temperature sensor over time

**Scenario:** Noisy temperature sensor, true temperature  $\mu = 22\text{C}$ .

**Setup:**

- Measurement noise:  $\sigma = 1.5\text{C} \Rightarrow \lambda = 0.444$ .
- Prior:  $\mu_0 = 20\text{C}$ ,  $\sigma_0 = 3\text{C} \Rightarrow \lambda_0 = 0.111$ .
- 20 sequential readings, typical values:  $\{20.1, 22.3, 21.8, 23.1, \dots\}$ .

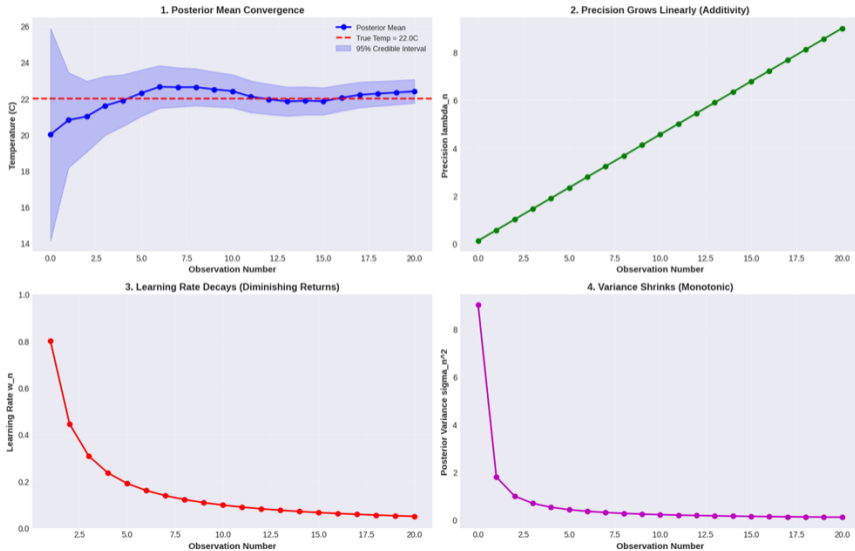
**Dynamics:**

$n$	$\lambda_n$	$\sigma_n$	$w_n$
1	0.556	1.342	0.800
5	2.333	0.655	0.190
10	4.556	0.468	0.097
20	8.889	0.335	0.050

**Trend:** Precision grows linearly, learning rate decays, variance shrinks monotonically.

# Online learning dynamics

## Online Learning Dynamics: Sequential Bayesian Updates



# Why Bayesian online learning beats alternatives

Compare three approaches for 1M sensor readings:

	Storage	Updates	Recompute
Batch (store all)	$O(n)$	Once at end	Full posterior
<b>Bayes online</b>	$O(1)$	Each reading	Full posterior

**Key advantage:** Bayes online gives full posterior & uncertainty, not just point estimate.

## Streaming Bayesian inference

Sequential updates are exact (not approximations). Posterior is proper at every step. No batch reprocessing needed. Natural uncertainty quantification.

# Chi-square, Gamma, and why we care

**Definition (sum of squares).** If  $Z_1, \dots, Z_k \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ , then

$$X = \sum_{i=1}^k Z_i^2 \sim \chi_k^2.$$

This makes  $\chi^2$  the natural distribution for *energy*, *squared error*, and *quadratic forms*.

**Key relationship:** Chi-square is a special case of the Gamma distribution.

$$X \sim \chi_k^2 \iff X \sim \text{Gamma}\left(\alpha = \frac{k}{2}, \theta = 2\right)$$

(using  $\text{Gamma}(\text{shape } \alpha, \text{scale } \theta)$ ). Equivalently, with rate  $\beta = 1/\theta$ :

$$X \sim \text{Gamma}\left(\alpha = \frac{k}{2}, \beta = \frac{1}{2}\right).$$

**Why it matters (canonical results).** If  $x_1, \dots, x_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ , then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2,$$

which drives: (i) confidence intervals for  $\sigma^2$ , (ii) chi-square tests, (iii) Gaussian-conjugate updates (via Gamma on precision).

Visualization: [probstats.org](http://probstats.org) (interactive PDFs/CDFs for  $\chi^2$  and Gamma),

# Inverse- $\chi^2$ , scaled inverse- $\chi^2$ , and Gamma priors on precision

**Inverse- $\chi^2$  (basic transform).** If  $X \sim \chi_\nu^2$ , then

$$V = \frac{1}{X} \sim \text{Inv-}\chi_\nu^2.$$

**Scaled inverse- $\chi^2$  (Bayesian variance prior).** A very common conjugate prior for a Normal variance  $\sigma^2$  is:

$$\sigma^2 \sim \text{Scale-inv-}\chi^2(\nu, \tau^2), \quad \text{equivalently} \quad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu}{2}, \frac{\nu\tau^2}{2}\right),$$

(Inv-Gamma(shape, scale) parameterization).

**Precision view (often cleaner).** Let  $\lambda = 1/\sigma^2$  (precision). Then:

$$\sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu}{2}, \frac{\nu\tau^2}{2}\right) \iff \lambda \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu\tau^2}{2}\right)$$

(Gamma(shape, rate) parameterization for  $\lambda$ ).

## Visualizationn

Inverse- $\chi^2$  / scaled inverse- $\chi^2$  have heavy right tails (variance can be large). Interactive plots: [probstats.org](http://probstats.org) (Gamma,  $\chi^2$ ); Wolfram: `InverseChiSquareDistribution`.

# Student- $t$ distribution and its relationship to the Gaussian

**Definition (location–scale Student- $t$ ).** A random variable  $T$  has a Student- $t$  distribution with  $\nu$  degrees of freedom, location  $\mu$ , and scale  $s$  if

$$T \sim t_{\nu}(\mu, s) \iff \frac{T - \mu}{s} \sim t_{\nu}.$$

Its density has heavier tails than a Gaussian (more mass far from the mean), especially for small  $\nu$ . **Relationship to Gaussian.**

- As  $\nu \rightarrow \infty$ ,  $t_{\nu}(\mu, s) \Rightarrow \mathcal{N}(\mu, s^2)$  (tails become Gaussian).
- For  $\nu > 1$ ,  $\mathbb{E}[T] = \mu$ ; for  $\nu > 2$ ,  $\text{Var}(T) = \frac{\nu}{\nu - 2} s^2$  (larger than  $s^2$ ).
- Small  $\nu$  (e.g.,  $\nu \in [3, 10]$ ) yields strong robustness: outliers are less influential than under a Gaussian likelihood.

**Where it appears in Bayesian inference.** In the Normal model with unknown variance, integrating out  $\sigma^2$  (with an Inv-Gamma / scaled-inv- $\chi^2$  prior) yields a Student- $t$  marginal for  $\mu$  and a Student- $t$  posterior predictive for new observations.

## Quick visualization (interactive)

Overlay  $t_{\nu}$  vs Gaussian for different  $\nu$ : [probstats.org](http://probstats.org). (Use as a screenshot figure or live demo.)

## Challenge: Both $\mu$ and $\sigma^2$ unknown

**Model:**  $x_1, \dots, x_n \mid \mu, \sigma^2 \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with both unknown.

**Problem:** Prior must specify joint  $p(\mu, \sigma^2)$  with possible dependence.

**Solution:** Use conjugate factorization

$$p(\mu, \sigma^2) = p(\mu \mid \sigma^2) \cdot p(\sigma^2).$$

**Conjugate choice:**

- **Conditional:**  $\mu \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$  (precision scales with  $\sigma^2$ ).
- **Marginal:**  $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$  (Inverse-Chi-squared prior).

### Hyperparameters

$\mu_0$ : prior mean for  $\mu$ .  $\kappa_0$ : prior “effective sample size” (scales prior variance).  $\nu_0$ : degrees of freedom for  $\sigma^2$  prior.  $\sigma_0^2$ : prior location guess for variance.

# Gamma distribution for precision

**Key shift:** Instead of Inverse- $\chi^2$  for variance, use Gamma for precision.

**Gamma distribution (shape-rate parametrization):**

$$\lambda \sim \text{Gamma}(\alpha, \beta) \quad \Rightarrow \quad p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}.$$

**Properties:**

- Mean:  $\mathbb{E}[\lambda] = \alpha/\beta$ .
- Variance:  $\text{Var}[\lambda] = \alpha/\beta^2$ .
- Support:  $\lambda > 0$  (necessary for precision).
- Conjugate: Gamma is conjugate to Normal likelihood.

**Relationship to Inverse-Gamma:**

$$\sigma^2 = 1/\lambda \sim \text{Inv-}\Gamma, \quad \text{if } \lambda \sim \text{Gamma}.$$

## Why Gamma for precision?

More numerically stable, directly expresses prior beliefs about precision (not variance). Simpler update equations.

# Normal-Gamma conjugate prior

**Joint prior (factorized):**

$$p(\mu, \lambda) = p(\mu \mid \lambda) \cdot p(\lambda),$$

where

$$\begin{aligned}\mu \mid \lambda &\sim \mathcal{N}(\mu_0, (\kappa_0 \lambda)^{-1}), \\ \lambda &\sim \text{Gamma}(\alpha_0, \beta_0).\end{aligned}$$

**Hyperparameters:**

- $\mu_0, \kappa_0$ : prior location and strength for  $\mu$ .
- $\alpha_0, \beta_0$ : shape and rate for precision.

**Why this works:** The likelihood factors as

$$p(x_1, \dots, x_n \mid \mu, \lambda) = (\text{term in } \mu) \times (\text{term in } \lambda),$$

so posterior is also Normal-Gamma with updated hyperparameters.

# Posterior: Normal-Gamma family

## Result (conjugate update):

$$p(\mu, \lambda \mid \text{data}) = \mathcal{N}(\mu \mid \mu_n, (\kappa_n \lambda)^{-1}) \cdot \text{Gamma}(\lambda \mid \alpha_n, \beta_n).$$

## Update equations:

$$\kappa_n = \kappa_0 + n,$$

$$\mu_n = \frac{\kappa_0 \mu_0 + n \bar{x}}{\kappa_n},$$

$$\alpha_n = \alpha_0 + \frac{n}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2} \left[ \sum (x_i - \bar{x})^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{x} - \mu_0)^2 \right].$$

## Interpretation:

- Posterior precision  $\kappa_n$  is prior + data (mixture weight).
- Posterior  $\alpha_n, \beta_n$  encode both data variance and mean discrepancy.

# Marginal posterior for $\mu$ : Student- $t$ distribution

When we marginalize out the unknown  $\lambda$ :

$$p(\mu \mid \text{data}) = \int p(\mu \mid \lambda, \text{data}) p(\lambda \mid \text{data}) d\lambda \sim t_{\nu_n}(\mu_n, \sigma_n^2).$$

**Degrees of freedom:**  $\nu_n = 2\alpha_n$ .

**Scale:**  $\sigma_n^2 = \frac{\beta_n}{\alpha_n \kappa_n}$ .

**Student- $t$  density:**

$$p(\mu \mid \text{data}) \propto \left( 1 + \frac{\kappa_n(\mu - \mu_n)^2}{\nu_n \sigma_n^2} \right)^{-(\nu_n+1)/2}.$$

## Why Student- $t$ , not Gaussian?

Uncertainty in  $\sigma^2$  creates heavier tails. Extreme values of  $\mu$  are less surprising when variance is also unknown. Robustness to outliers.

# Intuition: Heavy tails from variance uncertainty

**Known  $\sigma^2$ :**

Posterior for  $\mu$  is Gaussian with variance  $\sigma^2/n$  (tight).

**Unknown  $\sigma^2$ :**

Posterior marginalizes over plausible  $\sigma^2$  values:

$$p(\mu \mid \text{data}) = \int p(\mu \mid \sigma^2, \text{data}) p(\sigma^2 \mid \text{data}) d\sigma^2.$$

Some posterior draws have  $\sigma^2 > \hat{\sigma}^2$  (true value underestimated)  $\Rightarrow$  wider credible intervals.  
This averaging manifests as Student- $t$  tails (heavier than Gaussian).

## Robustness

Models estimating  $\sigma^2$  are robust to outliers. Heavy tails allow occasional large deviations without dominating posterior.

## Worked example: Internet speed (unknown variance)

**Data:**  $n = 5$ ,  $\bar{x} = 16.0$ ,  $\sum (x_i - \bar{x})^2 = 67.6 \text{ Mbps}^2$ .

**Prior:**  $\mu_0 = 20$ ,  $\kappa_0 = 1$ ,  $\alpha_0 = 2.5$ ,  $\beta_0 = 12.5$  (informative).

**Posterior:**

$$\kappa_n = 1 + 5 = 6,$$

$$\mu_n = \frac{1 \cdot 20 + 5 \cdot 16}{6} = 16.67,$$

$$\alpha_n = 2.5 + 2.5 = 5,$$

$$\beta_n = 12.5 + 33.8 + 3.33 = 49.6.$$

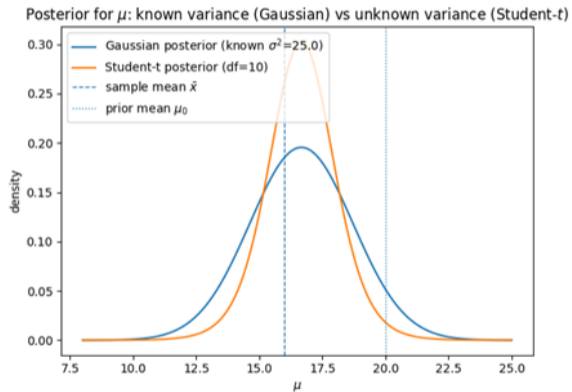
**Marginal posterior for  $\mu$ :**

$$\mu \mid \text{data} \sim t_{10}(16.67, 1.65).$$

Posterior std:  $\sqrt{1.65} \approx 1.28 \text{ Mbps}$  (wider than known- $\sigma$  case).

Heavier tails:  $\Pr(\mu > 20) \approx 0.09$  (vs. 0.05 in known case).

# Student- $t$ vs Gaussian



# Summary: The Gaussian toolbox

- 1 **Known variance:** Posterior is Gaussian. Precision additivity.
- 2 **Online learning:** Sequential updates decay in influence. No batch reprocessing.
- 3 **Unknown variance:** Normal-Gamma posterior. Marginal  $\mu$  is Student- $t$  (heavier tails, robustness).
- 4 **Monte Carlo:** Simulate to compute expectations and functions of parameters.
- 5 **Multivariate:** Matrix precision additivity, conditioning, marginalization.
- 6 **Linear Gaussian:** Conditional, marginal, and predictive distributions all Gaussian.

## Central principle

Precisions (inverse variances) combine additively. This algebraic property drives conjugacy, simplicity, and analytical tractability.

## Exercise ideas for students

- 1 Derive posterior for  $\mu$  when prior is Exponential and likelihood is Gaussian (hint: not conjugate).
- 2 Extend internet speed example to  $n = 50$  measurements. Plot convergence of  $\mu_n$  and learning rate  $w_n$ .
- 3 Prove: For Normal-Gamma posterior,  $\mathbb{E}[\mu \mid \text{data}] = \mu_n$  and  $\text{Var}[\mu \mid \text{data}] = \mathbb{E}_\lambda[\sigma_n^2] + \text{Var}_\lambda[\mu]$ .
- 4 Implement online Bayesian filter for temperature sensor. Compare to exponential moving average (EMA).

## References and further reading

- **Gelman et al. (2013):** “Bayesian Data Analysis” (3rd ed.). Chapters 2, 3.
- **Murphy (2012):** “Machine Learning: A Probabilistic Perspective”. Chapters 2, 3.
- **Bernardo & Smith (2009):** “Bayesian Theory”. Foundational reference.
- **Lecture notebooks:** “Gaussian\_Models\_Extended-1.ipynb” contains worked examples, interactive plots, and exercises.
- **Software:** PyMC3, Stan, NumPyro for Bayesian inference.

This document was converted into LaTeX slides and formatted with assistance from ChatGPT (OpenAI). The instructor provided the source text and verified the final structure and wording.