Module 2

Probability Distributions

Spring 2026

Learning goals

- Review key probability distributions, activities focus on Gaussian PDFs in multiple dimensions.
  - Discrete probability distributions such as Bernoulli or Binomial, or continuous PDFs such as Student-t, are discussed in chapter 2. They will not be used in the current set of activities, but are relevant to future work.
- Reflect on why the Gaussian Distribution is so widely used to model the distribution of continuous variables. We will introduce PDF representing correlated variables

Why probability distributions matter

- ❑ Measurements are noisy
- ❑ Models are uncertain
- ❑ Predictions require uncertainty

Probability as uncertainty, not randomness

- Describes incomplete information
- Applies to:
  - experiments
  - Parameters
  - predictions

- Probability is bookkeeping for uncertainty!

Discrete Probability

Example: coin toss

$$p(x = \text{heads}|\mu) = \mu$$
$$p(x = \text{tails}|\mu) = 1 - \mu$$

Described by Bernoulli distribution

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

$$\sum_i p(x_i) = 1$$

Binomial Distribution

- How many times would we expect m heads in N coin flips?

$$\mathrm{B}in(m|N,\mu) = \binom{N}{m} \mu^{m} (1-\mu)^{N-m}$$

$$\binom{N}{m} = \frac{N!}{(N-m)!m!}$$

-> number of ways of choosing m objects out of N identical objects

From probabilities to densities

- What if parameters are continuous? e.g. voltage, position, etc
- Probability of *exact* value is *zero*
- We must define a probability *density* instead

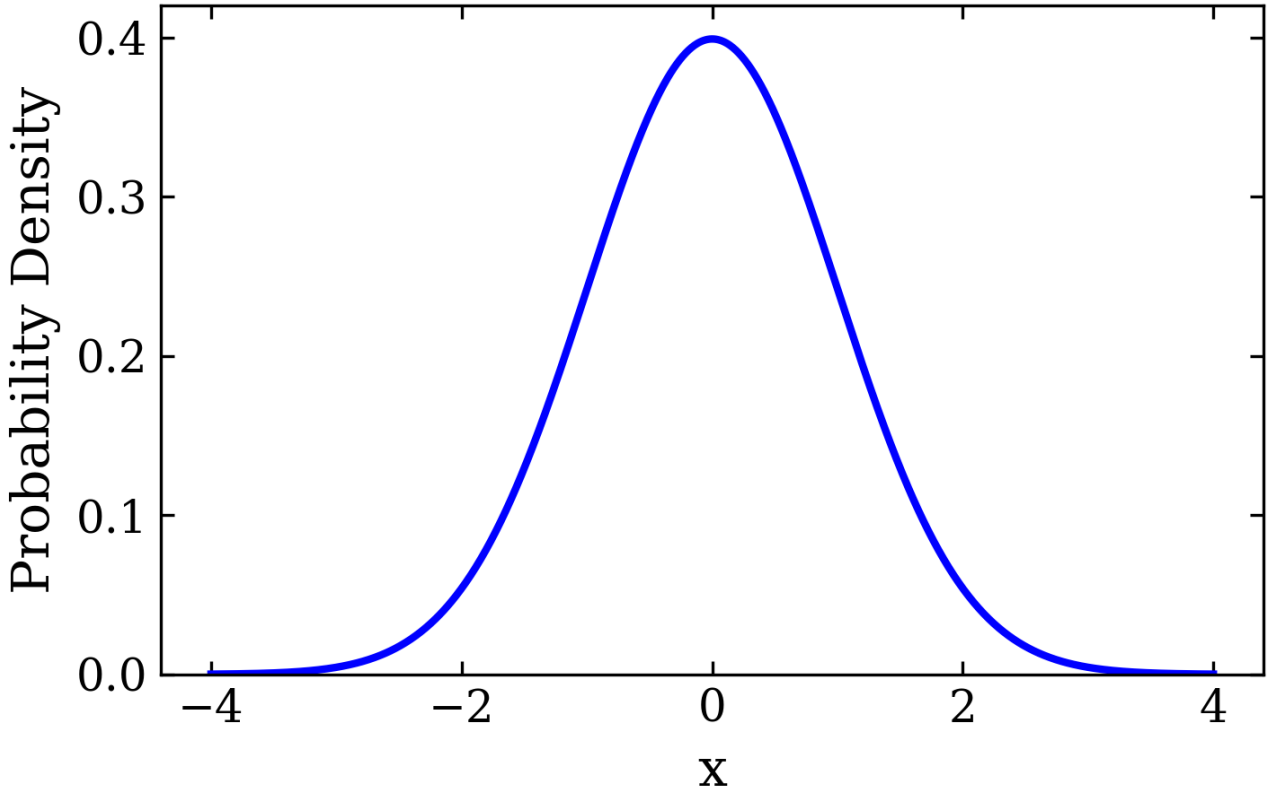
- Think about probability not at fixed point, but over an interval

Probability density function (PDF)

Probability density function

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$



Probability density function (PDF)

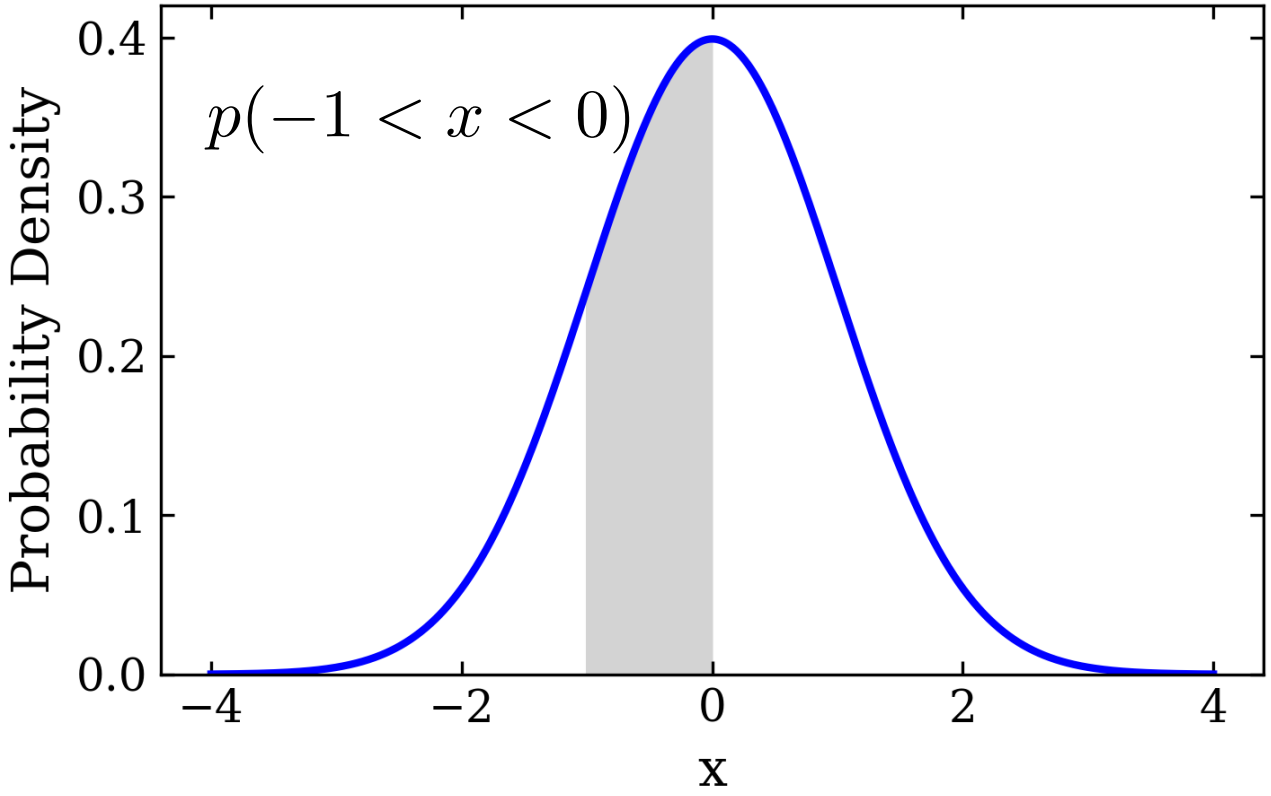
Probability density function

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Area = probability

Hight = density (NOT probability)



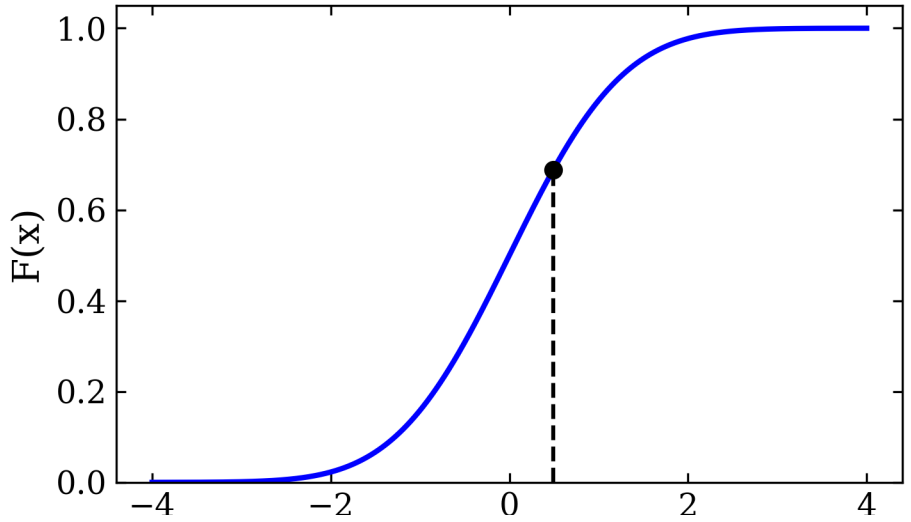
Cumulative distribution function (CDF)

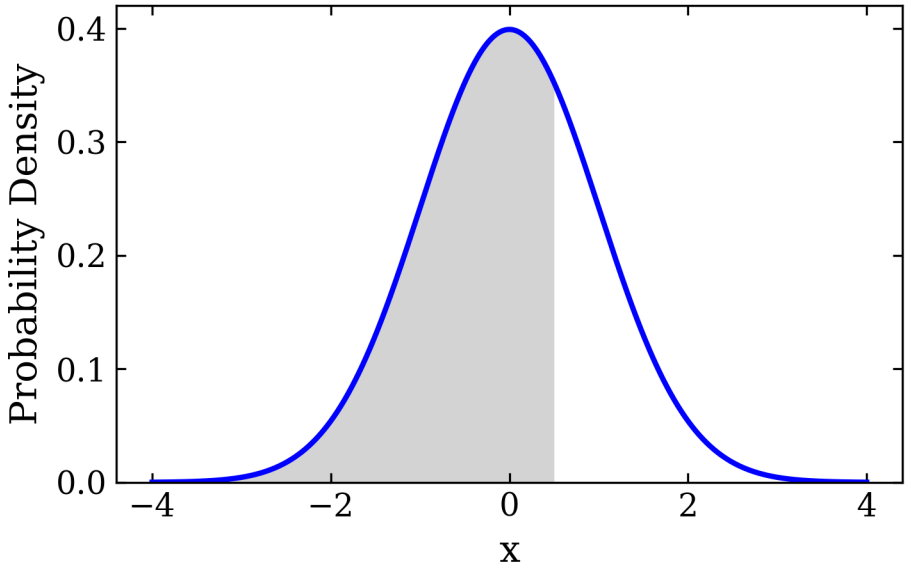
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x')dx'$$

- Monotonic function describing total of probabilities

- Useful for calculating probability within an interval, hypothesis testing, etc

$$P(a < X \leq b) = F(b) - F(a)$$





Expectation values

- Expectation value: weighted average of a function $f(x)$ under probability distribution $p(x)$

Discrete: $$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Continuous: $$\mathbb{E}[f] = \int p(x) f(x) \mathrm{d}x$$

Variance

- Variance: measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f]$

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

Covariance

- Covariance quantifies the extent in which the variance of multiple variables is related

$$\operatorname{cov}[X_i, X_j] = \mathbb{E}[\{X_i - \mathbb{E}[X_i]\}\{X_j - \mathbb{E}[X_j]\}]$$

- Variance can be thought of as a specific case of the covariance matrix

$$\operatorname{cov}[\mathbf{X}] = \begin{bmatrix} \operatorname{var}(X_i) & \operatorname{cov}(X_i, X_j) \\ \operatorname{cov}(X_j, X_i) & \operatorname{var}(X_j) \end{bmatrix}$$

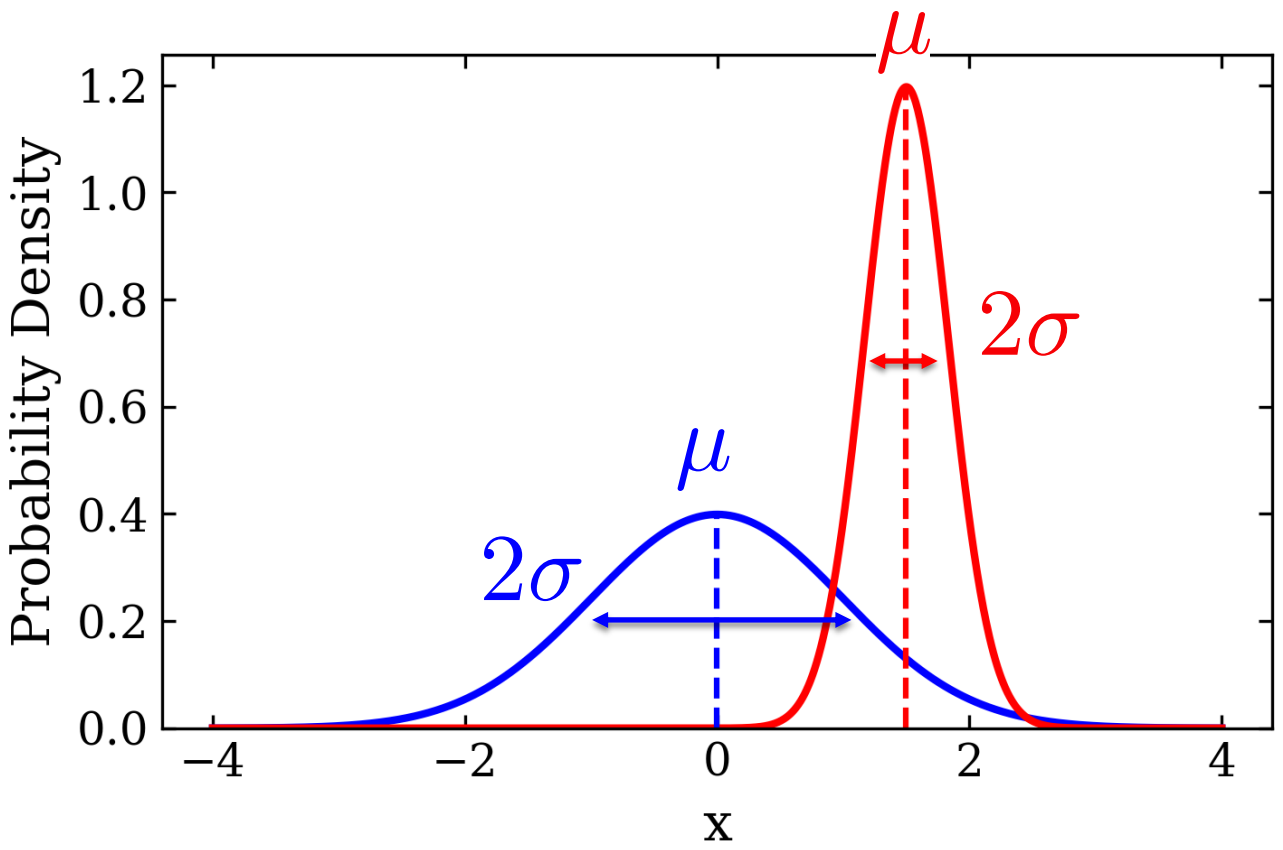
Normal (Gaussian) Distribution

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

- Parameterized by μ and σ^2
- μ: center of distribution
- σ: width of distribution

$$\mu = \mathbb{E}[x]$$

$$\sigma^2 = \mathrm{var}[x]$$



Why do we care about Gaussians?

- ❑Gaussian distributions occur everywhere in physics
  - Thermal noise
  - Electronics noise
  - Measurement errors
  - etc…
- ❑But why?

Central Limit Theorem

- **The scaled sum of a sequence of i.i.d. random variables with finite mean and variance converges in distribution to the normal distribution.**

Let $x_1, \ldots, x_N$ be independent random variables with finite mean μ and variance σ^2. Then the distribution of the sample mean:

$$\bar{x} = \frac{1}{N}\sum_i x_i$$

approaches a Gaussian distribution as $N \to \infty$, regardless of the original distribution.

Central Limit Theorem

Assume you have N independent measurements $x_1, \ldots, x_N$ drawn from a distribution with $\mathbb{E}[x_i] = \mu \quad \mathrm{Var}(x_i) = \sigma^2$

Central Limit Theorem

Assume you have N independent measurements $x_1, \ldots, x_N$ drawn from a distribution with $\mathbb{E}[x_i] = \mu \quad \mathrm{Var}(x_i) = \sigma^2$

Define the sample mean: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

Central Limit Theorem

Assume you have N independent measurements $x_1, \ldots, x_N$ drawn from a distribution with $\mathbb{E}[x_i] = \mu \quad \mathrm{Var}(x_i) = \sigma^2$

Define the sample mean: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

Calculate the variance of the mean

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) = \frac{1}{N^2} \mathrm{Var}\left(\sum_{i=1}^{N} x_i \right)$$

Central Limit Theorem

$$\mathrm{Var}\left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} \mathrm{Var}(x_i) + 2\sum_{i<j} \mathrm{Cov}(x_i, x_j)$$

Central Limit Theorem

$$\mathrm{Var}\left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} \mathrm{Var}(x_i) + 2\sum_{i<j} \mathrm{Cov}(x_i, x_j)$$

If the x_i are independent, $\mathrm{Cov}(x_i, x_j) = 0 \; for \; i \neq j$. Then:

$$\mathrm{Var}\left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} \sigma^2 = N\sigma^2 \longrightarrow \mathrm{Var}(\bar{x}) = \frac{1}{N^2}(N\sigma^2) = \frac{\sigma^2}{N}$$

Central Limit Theorem

$$\mathrm{Var}\left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} \mathrm{Var}(x_i) + 2\sum_{i<j} \mathrm{Cov}(x_i, x_j)$$

If the x_i are independent, $\mathrm{Cov}(x_i, x_j) = 0 \; for \; i \neq j$. Then:

$$\mathrm{Var}\left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} \sigma^2 = N\sigma^2 \longrightarrow \mathrm{Var}(\bar{x}) = \frac{1}{N^2}(N\sigma^2) = \frac{\sigma^2}{N}$$

Averaging N independent noisy measurements cancels fluctuations, so the noise amplitude shrinks like $\sqrt{N}$

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}}$$

Central Limit Theorem

- Many phenomena in physics are not necessarily Gaussian

Central Limit Theorem

- Many phenomena in physics are not necessarily Gaussian

  However, most *measurements* are

Central Limit Theorem

- Example: voltage noise

Central Limit Theorem

- Example: voltage noise

For a given sample of noise V_{noise} in an electrical system, you are really measuring the *sum* of many physical processes

$$V_{\mathrm{noise}} = \delta V_1 + \delta V_2 + \delta V_3 + \cdots$$

Central Limit Theorem

- Example: voltage noise

For a given sample of noise V_{noise} in an electrical system, you are really measuring the *sum* of many physical processes

$$V_{\text{noise}} = \delta V_1 + \delta V_2 + \delta V_3 + \cdots$$

Gaussian random variable

- thermal motion of many electrons
- microscopic scattering
- phonons
- shot noise contributions
- …. Not necessarily gaussian

You will see this in the coding activity

Lecture 2

What about sums of Gaussians?

What if we want to add multiple Gaussian variables?

E.g. $Z = X + Y$

What about sums of Gaussians?

What if we want to add multiple Gaussian variables?

E.g. $Z = X + Y$

- Expectation is linear operator -> $\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(Y)$

What about sums of Gaussians?

What if we want to add multiple Gaussian variables?

E.g. $Z = X + Y$

- Expectation is linear operator -> $\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(Y)$

$$\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

What about sums of Gaussians?

What if we want to add multiple Gaussian variables?

E.g. $Z = X + Y$

- Expectation is linear operator -> $\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(Y)$

$$\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

- If variables are independent $\mathrm{Cov}(X, Y) = 0$

What about sums of Gaussians?

What if we want to add multiple Gaussian variables?

E.g. $Z = X + Y$

- Expectation is linear operator -> $\mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(Y)$

$$\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

- If variables are independent $\mathrm{Cov}(X, Y) = 0$

$$\boxed{\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2}$$

Random Event Processes

Not all randomness comes from noise

- Many measurements involve discrete random events

Examples:

- radioactive decays
- photon detections
- dark counts in detectors
- cosmic ray hits

We want a statistical model for:

- when events occur
- how many events occur

Constant Rate Assumption

- Let us assume events occur independently
- And assume constant average rate λ

  λ = average events per unit time
- In a small time interval dt: $P(\text{event in } dt) = \lambda\, dt$

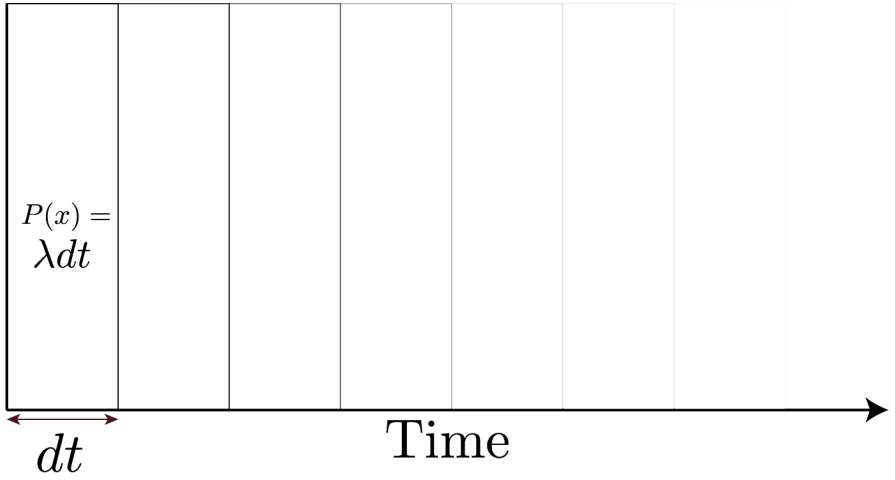
- This is the defining assumption of a ***Poisson process***

  → No memory, no aging, no buildup

We will motivate this as follows:

Discrete Time Approximation

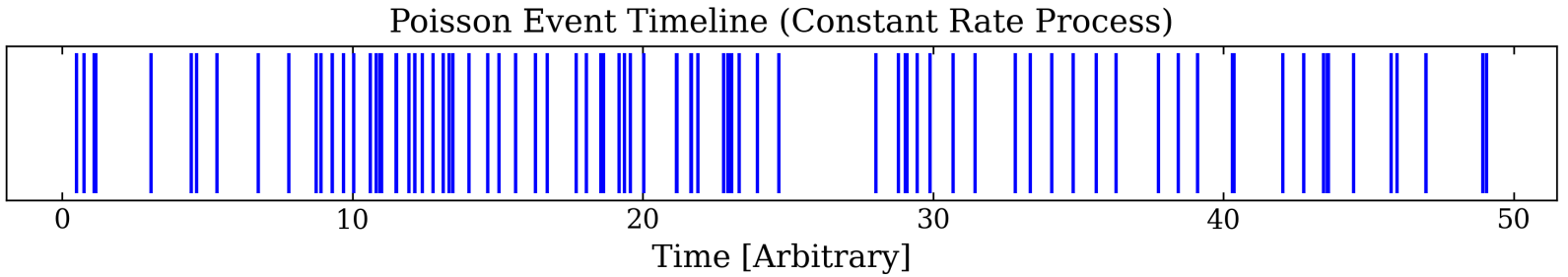
- Let us divide time into small bins dt
- In each bin event occurs with probability λdt



- Each bin is a Bernoulli trial

Simulated Event Timeline

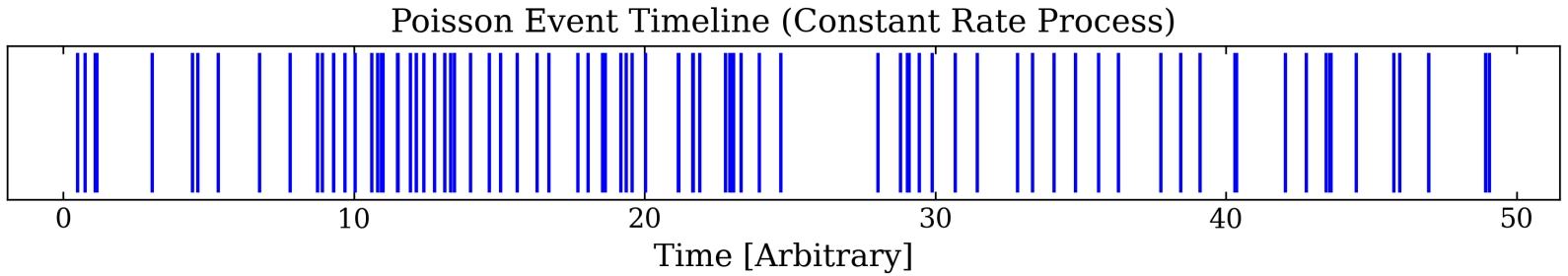
- For every bin dt we sample u from a random uniform distribution (0,1)
- If $u > \lambda\, dt$ then we say an 'event' occurred → repeat for all dt



Each line = detected event

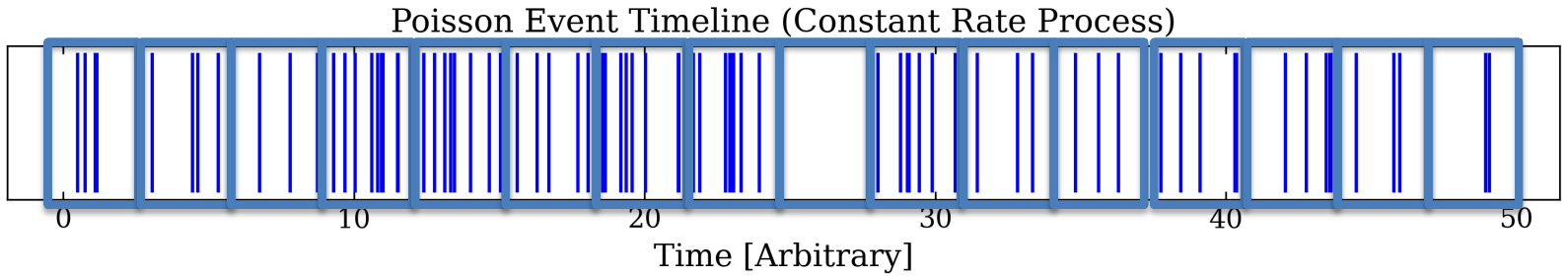
- Arrival times are irregular and unpredictable -> average density is constant

Counting Events in Fixed Windows



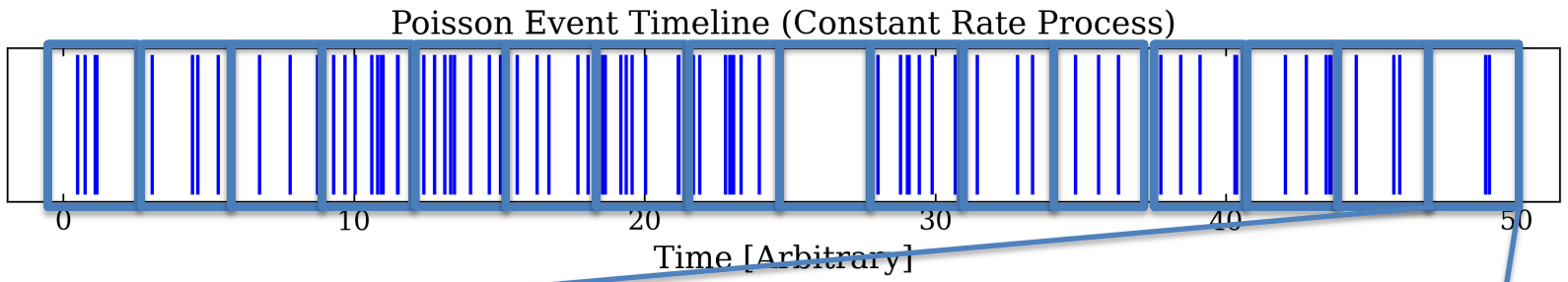
- We now divide into equal time windows

Counting Events in Fixed Windows

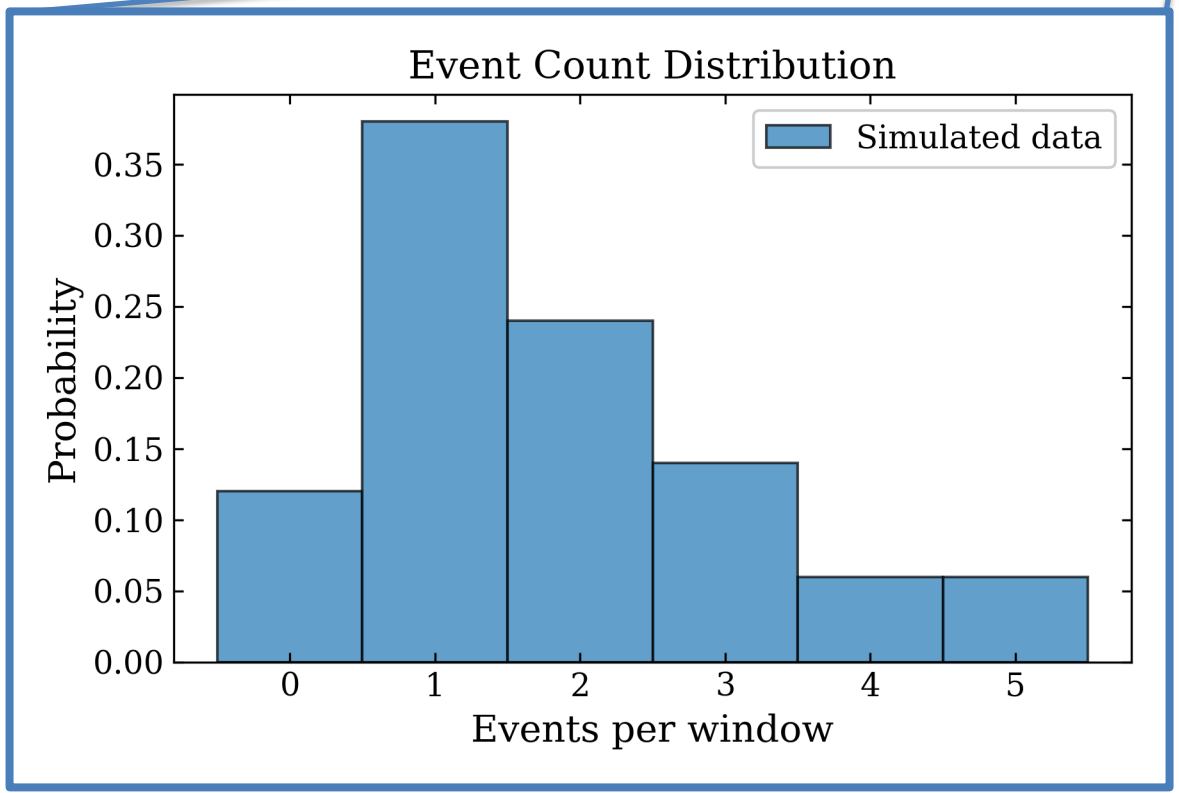


- We now divide into equal time windows

Counting Events in Fixed Windows



- Then build histogram of event counts per time window
- -> repeat and average over all windows to estimate probability



What describes this process?

- Recall, over short interval Δt:
  - probability of **one event:** $p(1) = \lambda \Delta t$

What describes this process?

- Recall, over short interval Δt:
  - probability of **one event:** $p(1) = \lambda \Delta t$
  - probability of no event: $p(0) = 1 - \lambda \Delta t$

What describes this process?

- Recall, over short interval Δt:
  - probability of **one event:** $p(1) = \lambda \Delta t$
  - probability of no event: $p(0) = 1 - \lambda \Delta t$
  - probability of **two or more events:** $p(\geq 2) \approx 0$

What describes this process?

- Recall, over short interval Δt:
  - probability of **one event:** $p(1) = \lambda \Delta t$
  - probability of no event: $p(0) = 1 - \lambda \Delta t$

  } Bernoulli trial

  - probability of **two or more events:** $p(\geq 2) \approx 0$

What describes this process?

- Consider total measurement time T

What describes this process?

- Consider total measurement time T

- Small time interval is now defined as: $\Delta t = \frac{T}{N}$

What describes this process?

- Consider total measurement time T
- Small time interval is now defined as: $\Delta t = \frac{T}{N}$
- Probability of one event in a bin is now:

$$p(1) = \lambda \Delta t = \frac{\lambda T}{N}$$

Binomial Selection

- Question becomes: What is the probability of observing exactly k events across all bins?

Binomial Selection

- Question becomes: What is the probability of observing exactly k events across all bins?

→ use the binomial distribution! $P_N(k) = \binom{N}{k} p^k (1-p)^{N-k}$

Binomial Selection

- Question becomes: What is the probability of observing exactly k events across all bins?

$\rightarrow$ use the binomial distribution! $P_N(k) = \binom{N}{k} p^k (1-p)^{N-k}$

Substitute in $p = \frac{\lambda T}{N}$

$$P_N(k) = \binom{N}{k} \left(\frac{\lambda T}{N}\right)^k \left(1 - \frac{\lambda T}{N}\right)^{N-k}$$

Large N limit

- Large N limit is dividing a fixed observation time into arbitrarily fine time resolution: continuum-time limit

Large N limit

- Large N limit is dividing a fixed observation time into arbitrarily fine time resolution: continuum-time limit
- Binomial coefficient → $$\binom{N}{k} \approx \frac{N^k}{k!}$$
- Power term → $$\left(\frac{\lambda T}{N}\right)^k = \frac{(\lambda T)^k}{N^k}$$
- Exponential term → $$\lim_{N\to\infty} \left(1 - \frac{\lambda T}{N}\right)^N = e^{-\lambda T}$$

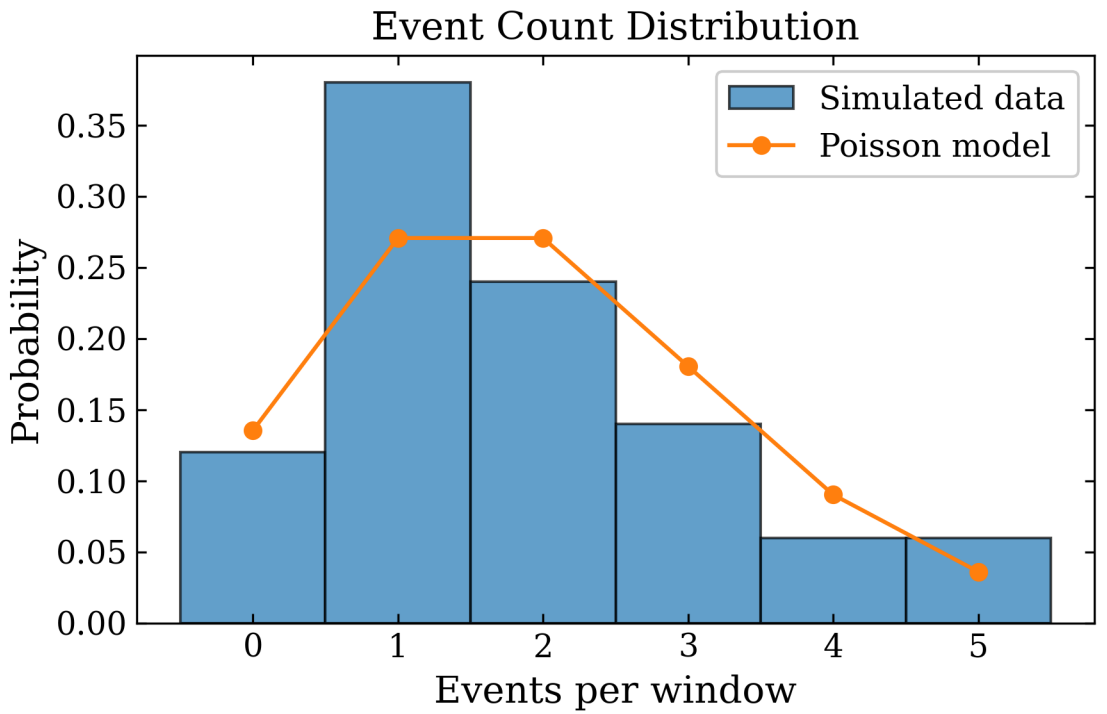
Large N limit

Let $\mu = \lambda T$

Arrive at the Poisson distribution:

$$P(k \mid \mu) = \frac{\mu^k}{k!} e^{-\mu}$$

$$\mathbb{E}(k) = \mathrm{var}(k) = \mu$$



Timing Distribution

Similar concept: the distribution of waiting times between events

Timing Distribution

Similar concept: the distribution of waiting times between events

- For Poissonian events, consider probability that the first event occurs at $t_1 > \mathrm{t}$ -> no event in the time window $[0, t]$

Timing Distribution

Similar concept: the distribution of waiting times between events

- For Poissonian events, consider probability that the first event occurs at $t_1 > \mathrm{t}$ -> no event in the time window $[0, t]$
- $p(t_1 > t) = p(k = 0) = e^{-\lambda t}$

Timing Distribution

Similar concept: the distribution of waiting times between events

- For Poissonian events, consider probability that the first event occurs at $t_1 > \mathrm{t}$ -> no event in the time window $[0, t]$
- $p(t_1 > t) = p(k = 0) = e^{-\lambda t}$
- CDF tells us probability an event occurred at or before t

Timing Distribution

Similar concept: the distribution of waiting times between events

- For Poissonian events, consider probability that the first event occurs at $t_1 > \mathrm{t}$ -> no event in the time window $[0, t]$
- $p(t_1 > t) = p(k = 0) = e^{-\lambda t}$
- CDF tells us probability an event occurred at or before t
- $F(t) = \ p(t_1 \leq t) = 1 - p(t_1 > t) = 1 - \ e^{-\lambda t}$

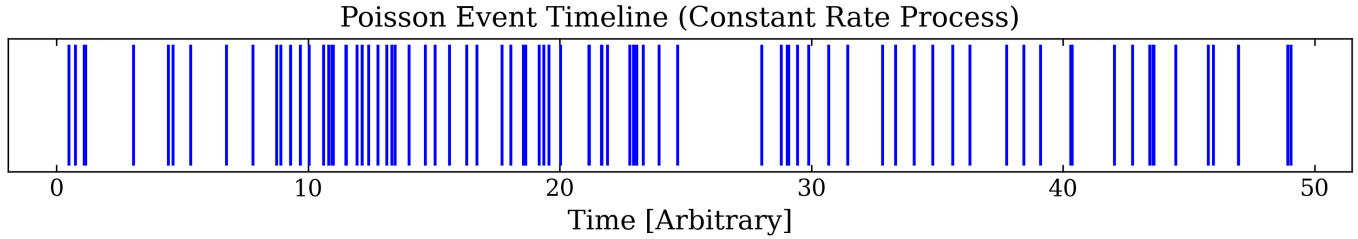
Timing Distribution

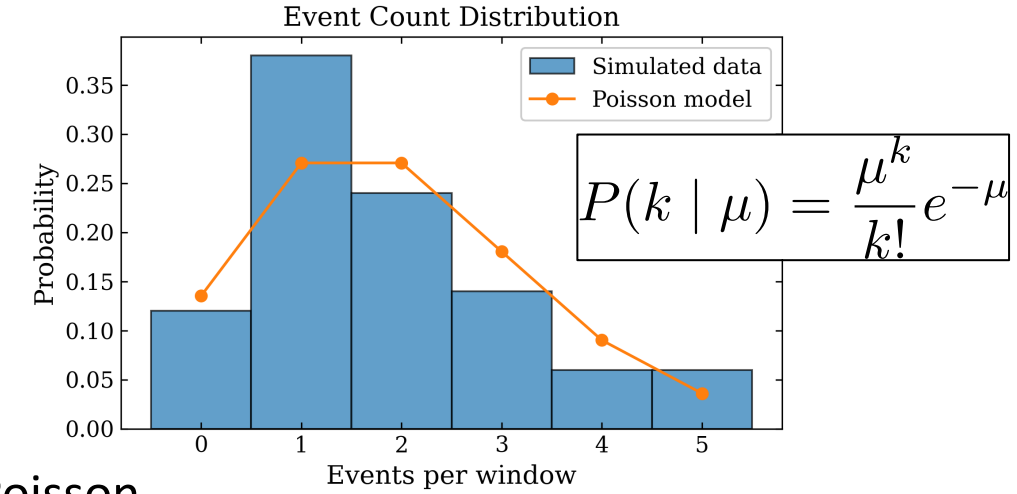
Similar concept: the distribution of waiting times between events

- For Poissonian events, consider probability that the first event occurs at $t_1 > \mathrm{t}$ -> no event in the time window $[0, t]$
- $p(t_1 > t) = p(k = 0) = e^{-\lambda t}$
- CDF tells us probability an event occurred at or before t
- $F(t) = \ p(t_1 \leq t) = 1 - p(t_1 > t) = 1 - \ e^{-\lambda t}$
- PDF is thus:

$$p(t) = \frac{d}{dt} F(t) = \lambda e^{-\lambda t}$$

Poisson and Exponential Distribution



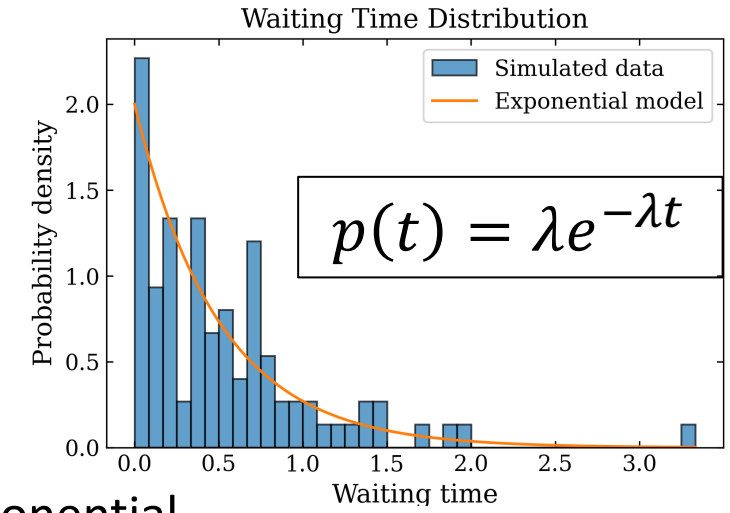


$$P(k \mid \mu) = \frac{\mu^k}{k!} e^{-\mu}$$

Poisson

- Discrete distribution
- "**how many** events in fixed time"

$$\mathbb{E}(k) = \mathrm{var}(k) = \mu$$



$$p(t) = \lambda e^{-\lambda t}$$

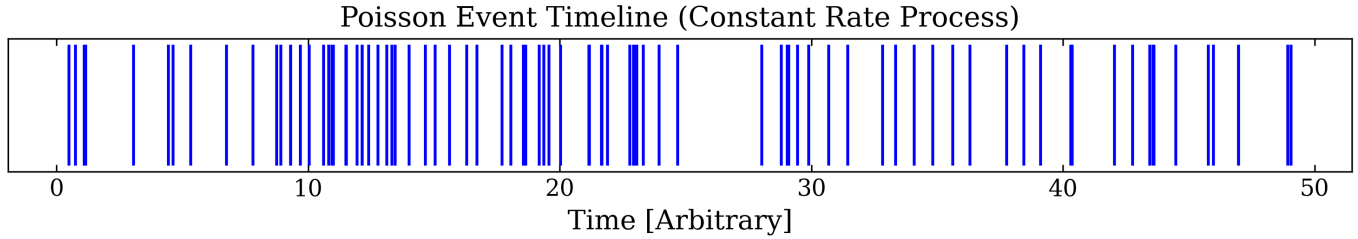
Exponential

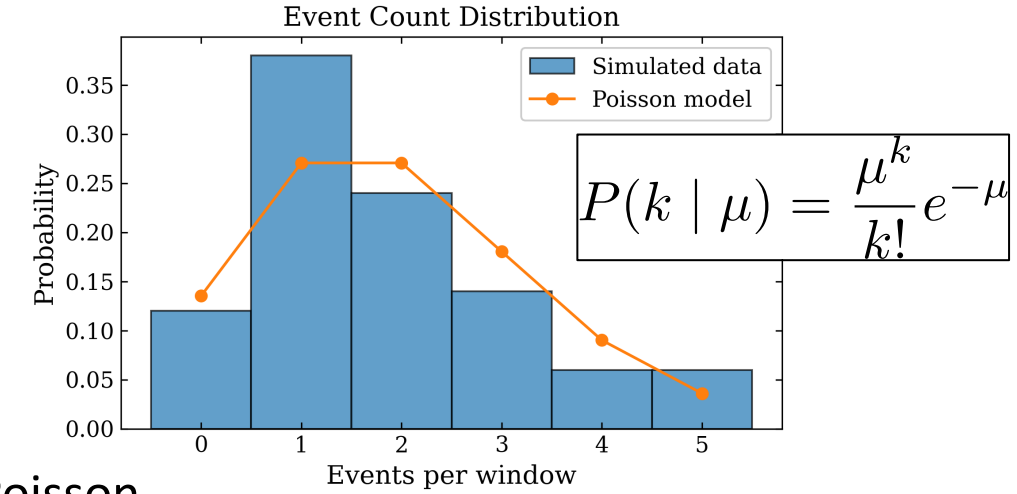
- Continuous distribution
- **"when** the next event happens"

$$\mathbb{E}(t) = \frac{1}{\lambda}$$

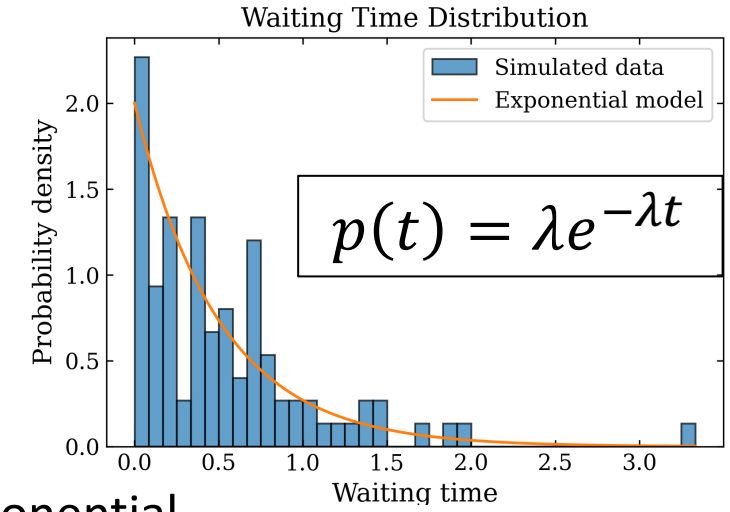
$$\mathrm{var}(t) = \frac{1}{\lambda^2}$$

Poisson and Exponential Distribution





$$P(k \mid \mu) = \frac{\mu^k}{k!} e^{-\mu}$$



$$p(t) = \lambda e^{-\lambda t}$$

Poisson

- Discrete distribution
- "**how many** events in fixed time"

Exponential

- Continuous distribution
- "**when** the next event happens"

$$\mathbb{E}(t) = \frac{1}{\lambda}$$

$$\mathrm{var}(t) = \frac{1}{\lambda^2}$$

$$\mathbb{E}(k) = \mathrm{var}(k) = \mu$$

Big deal!

CLT to the rescue

- When the average number of Poissonian events is large, the distribution tends to a Gaussian… why?

CLT to the rescue

- When the average number of Poissonian events is large, the distribution tends to a Gaussian… why?
- $\mu = \lambda T \gg 1$ means:
  - High event rate
  - Long observation period

CLT to the rescue

- When the average number of Poissonian events is large, the distribution tends to a Gaussian… why?
- $\mu = \lambda T \gg 1$ means:
  - High event rate
  - Long observation period
- →accumulating many statistically independent event opportunities

CLT to the rescue

- When the average number of Poissonian events is large, the distribution tends to a Gaussian… why?
- $\mu = \lambda T \gg 1$ means:
  - High event rate
  - Long observation period
- →accumulating many statistically independent event opportunities
- Total observed events is a *sum*

$$\boxed{\mathrm{Poisson}(\mu)_{\mu\to\infty} \to \mathcal{N}(\mu,\mu)}$$

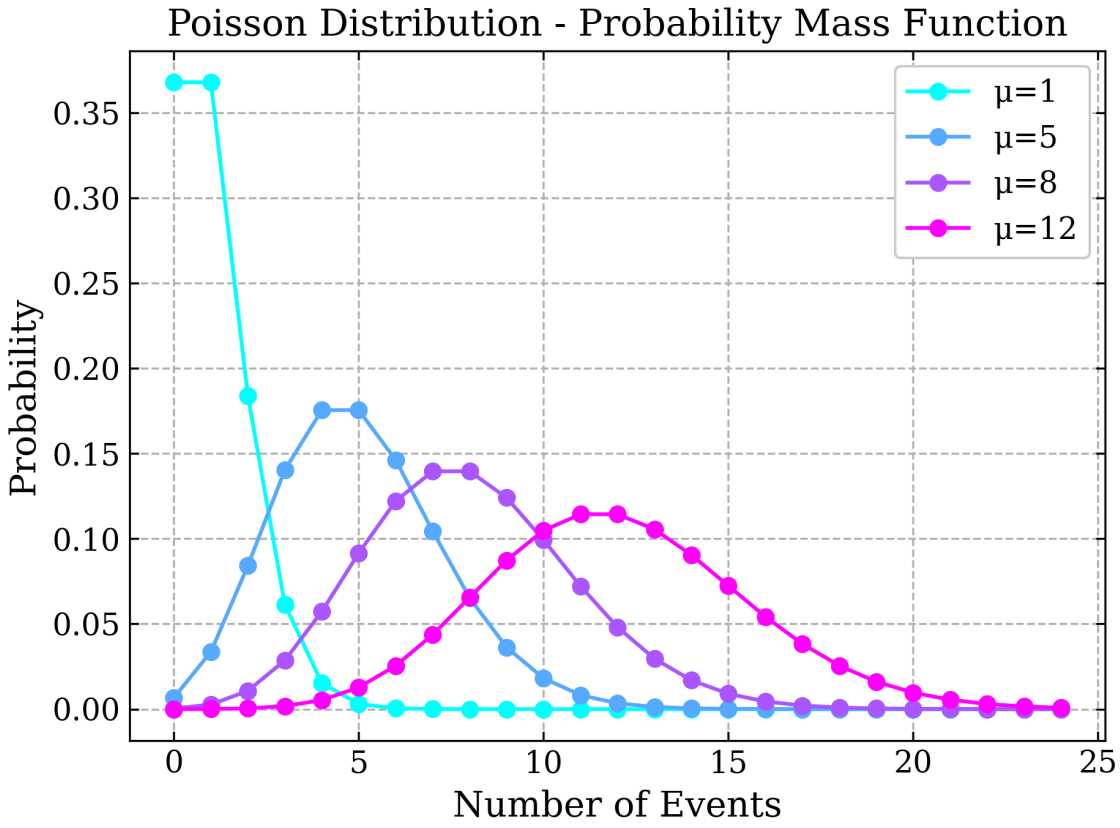
Large μ limit

$$\text{Poisson}(\mu)_{\mu \to \cancel{\infty}} \to \mathcal{N}(\mu, \mu)$$

> 10

Now we get the best of both words

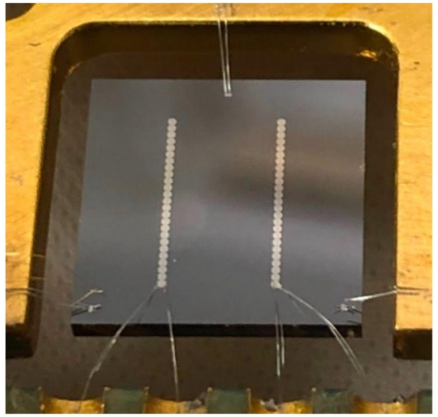
- Can take advantage of techniques for gaussian estimation
- **Mean and variance however still describe underlying Poisson physics!!**

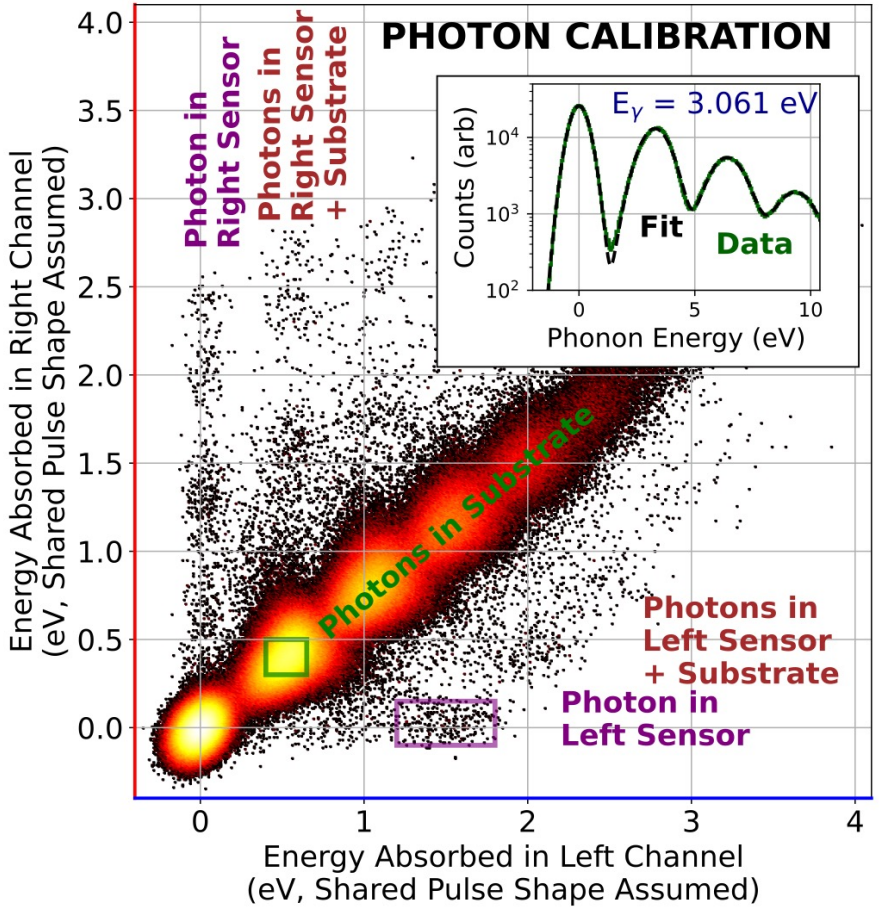


From Scalar to Vector-Valued Data

From Scalar to Vector-Valued Data

- Many measurements are vector-valued
- We need probability distributions over vectors
- Examples: detector hit positions, correlated sensor channels, regression parameter vectors





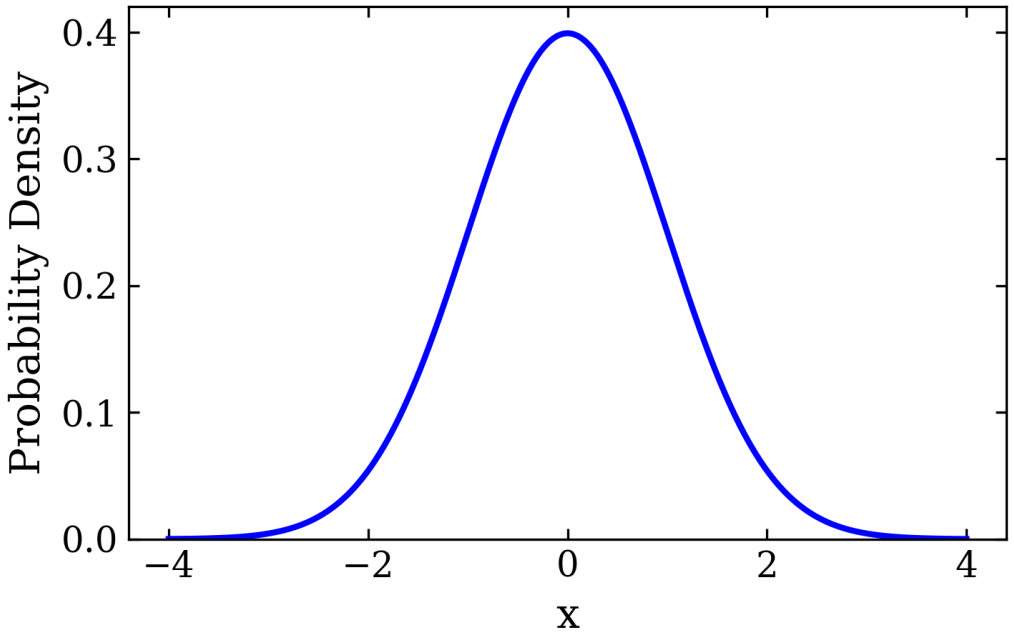
Multivariate Gaussian Distribution

- Characterized by a mean vector μ and covariance matrix Σ
- Notation: $p(x) = \mathcal{N}(\mu, \Sigma)$
- Extension of the familiar 1D Gaussian to higher dimensions

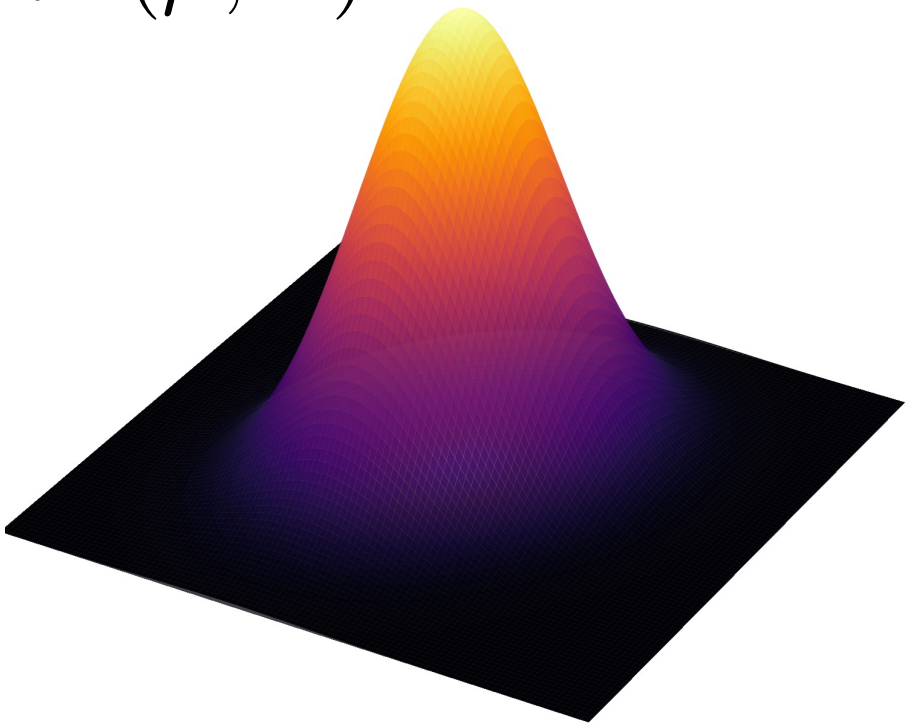
Multivariate Gaussian Distribution

Extension of the familiar 1D Gaussian to higher dimensions

$$\mathcal{N}(\mu, \sigma^2)$$



$$\mathcal{N}(\mu, \Sigma)$$



Multivariate Gaussian Distribution

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

$\Sigma =$ covariance matrix

$D =$ Dimension of covariance matrix

$\mu =$ array defining location of center

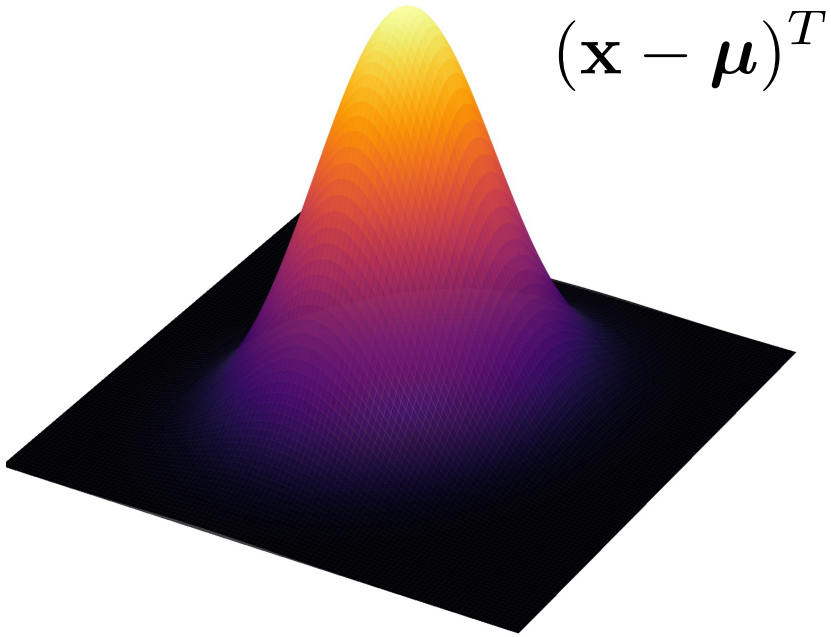
$$\Sigma = \begin{pmatrix} \sigma_x^2 & \mathrm{Cov}(x,y) \\ \mathrm{Cov}(x,y) & \sigma_y^2 \end{pmatrix}$$

$\mathrm{Cov}(x,y) = \rho\sigma_x^2\sigma_y^2 = \sigma_{xy}^2$

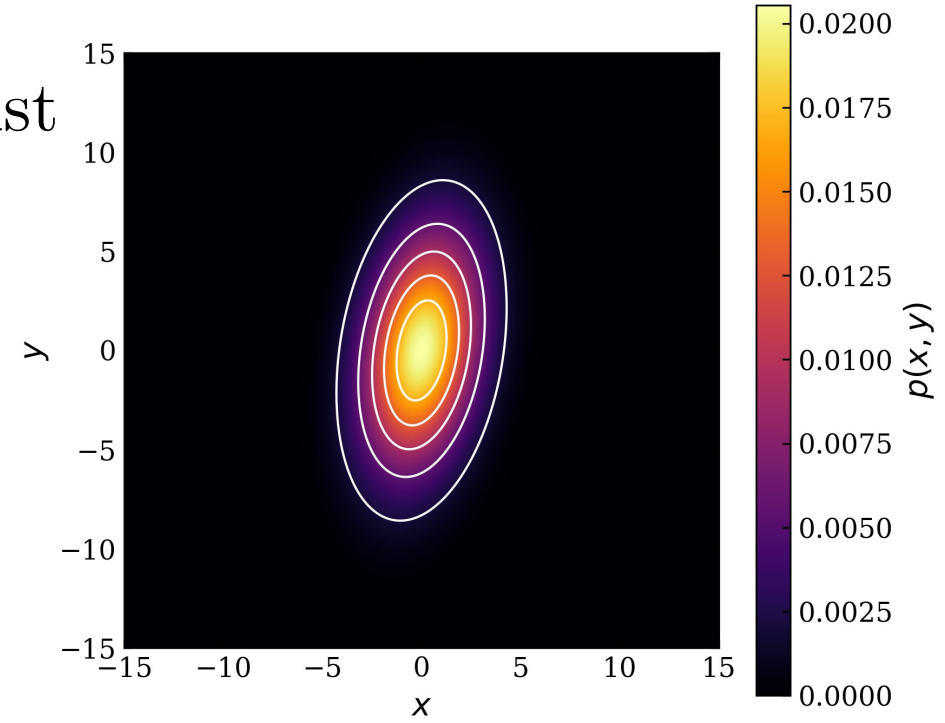
Where $\rho \in [-1,1]$ is the correlation coefficient

Gaussian Geometry: Ellipses in 2D

- Constant probability contours form ellipses
- Mean sets the center of the ellipse
- Covariance controls size, orientation, and shape



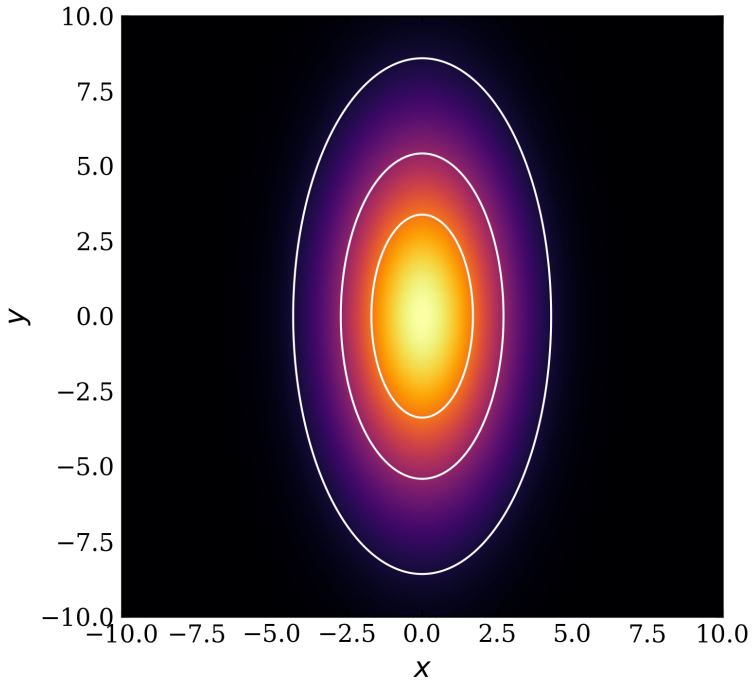
$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \text{const}$$

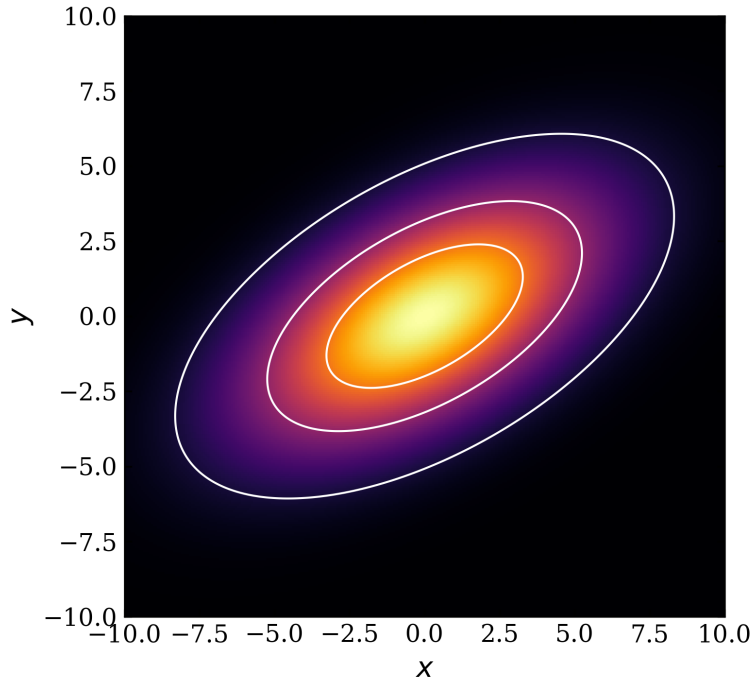


Interpreting the Covariance Matrix

$$\Sigma_{\text{uncorr}} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$



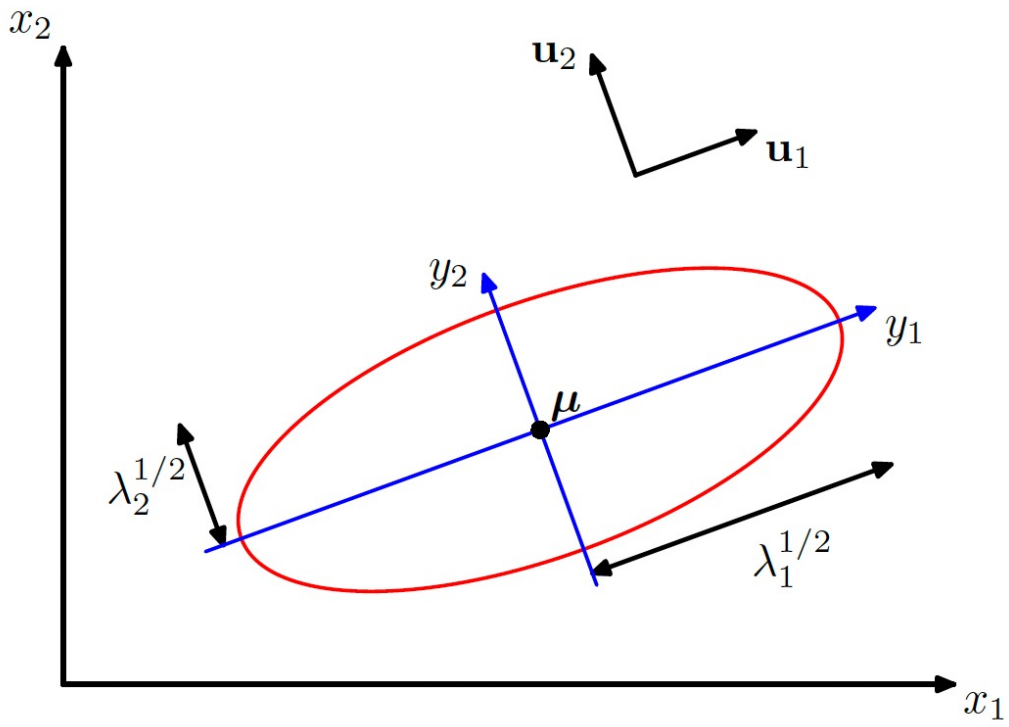


Interpreting the Covariance Matrix

Covariance matrices are symmetric and positive semidefinite

Geometry of ellipses described by eigenvalues/eigenvectors

$$\Sigma \mathbf{u_i} = \lambda_\mathbf{i} \mathbf{u_i} \qquad \Sigma = \sum_i^D \lambda_i \mathbf{u_i} \mathbf{u_i}^\mathrm{T}$$



Whitening Gaussian Data

- It is often easier to work with non-correlated data (key for many ML algorithms)

- We can remove correlations from our *description* of the data

 Change of basis

Removing Correlations

For dataset of vectors $\mathbf{x}_n \in \mathbb{R}^D, \quad n = 1, \ldots, N$

Removing Correlations

For dataset of vectors $\mathbf{x}_n \in \mathbb{R}^D, \quad n = 1, \ldots, N$

Step 1: Center the data: $\mathbf{y}_n = \mathbf{x}_n - \boldsymbol{\mu}$

Removing Correlations

For dataset of vectors $\mathbf{x}_n \in \mathbb{R}^D, \quad n = 1, \ldots, N$

Step 1: Center the data: $\mathbf{y}_n = \mathbf{x}_n - \boldsymbol{\mu}$

Covariance matrix is then given as

$$\Sigma = \frac{1}{N} Y^T Y \qquad Y = \begin{pmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_N^T \end{pmatrix}$$

Removing Correlations

Step 2: Eigen-decomposition

Removing Correlations

Step 2: Eigen-decomposition

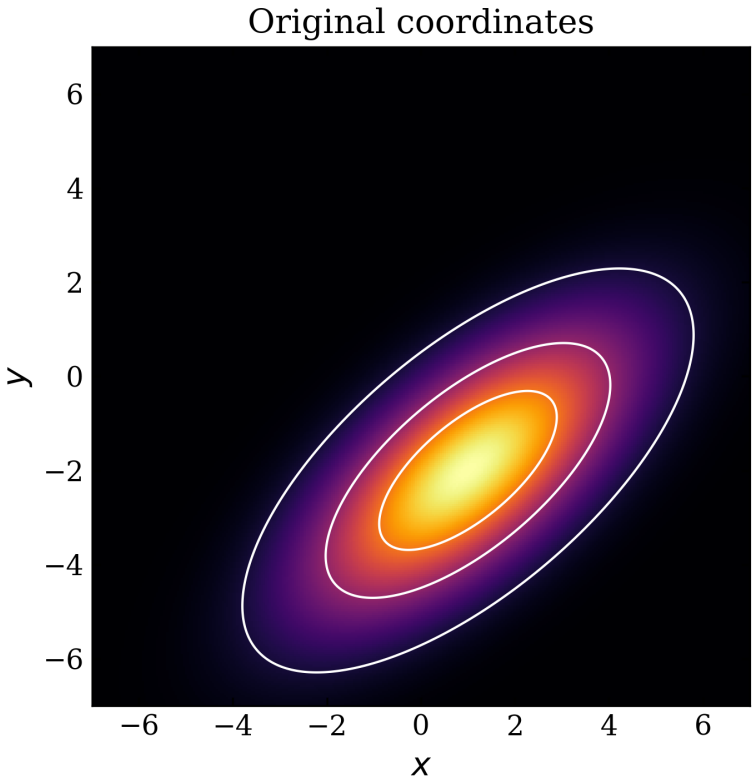
$$\Sigma = Q \boldsymbol{\lambda} Q^T$$

Q orthonormal eigenvectors

$\boldsymbol{\lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ eigenvalues

Removing Correlations

Step 3: rotate into principal axes $Z = YQ$

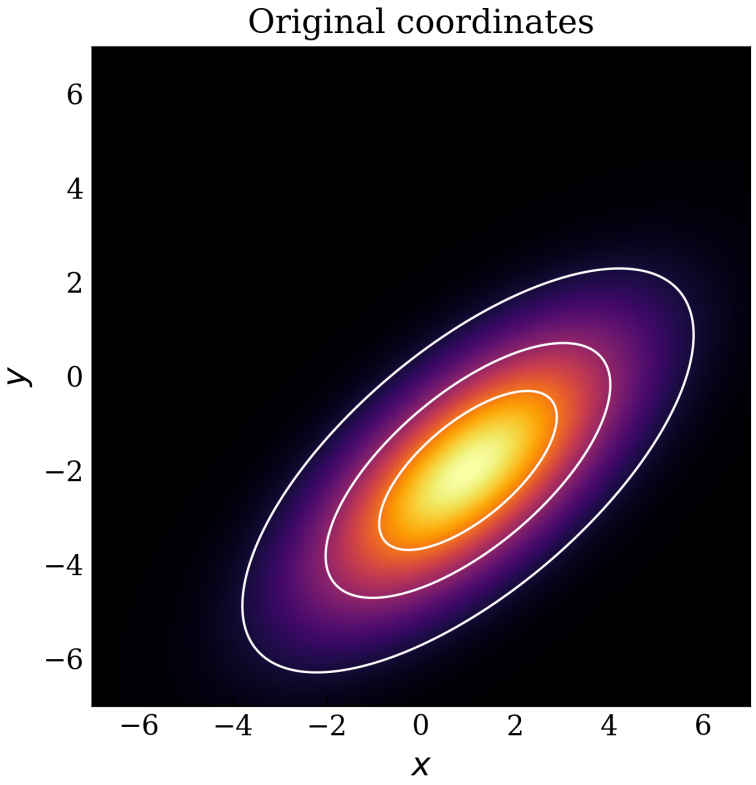


Level sets defined by:

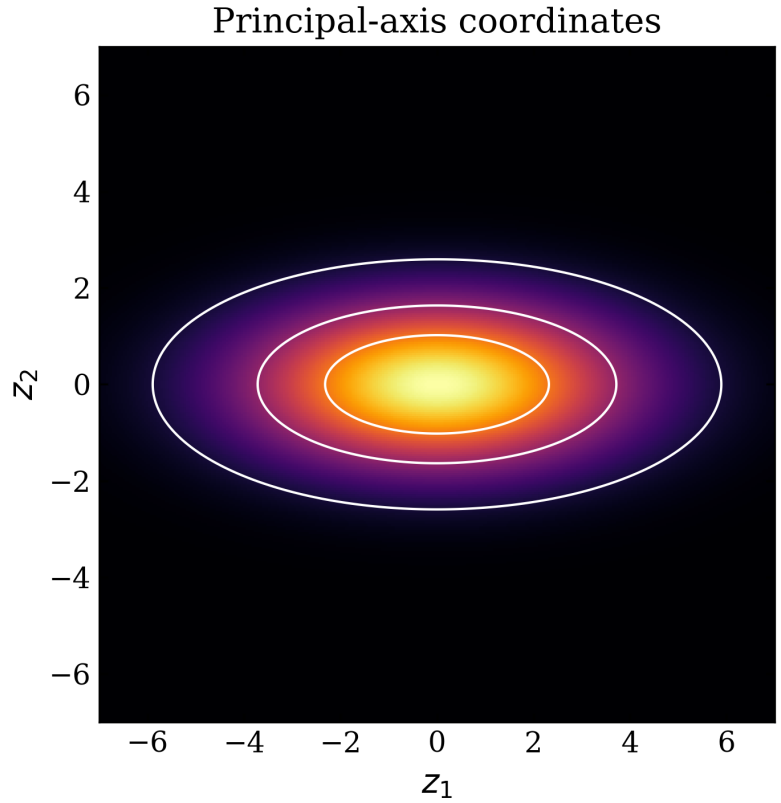
$$q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Removing Correlations

Step 3: rotate into principal axes $Z = YQ$



$$z = Q^T(x - \mu)$$



Level sets defined by:

$$q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$q(\mathbf{z}) = \mathbf{z}^T \boldsymbol{\lambda}^{-1} \mathbf{z}$$

Whitening

As a final step, we can rescale the data such that each axis has unit variance

Whitening

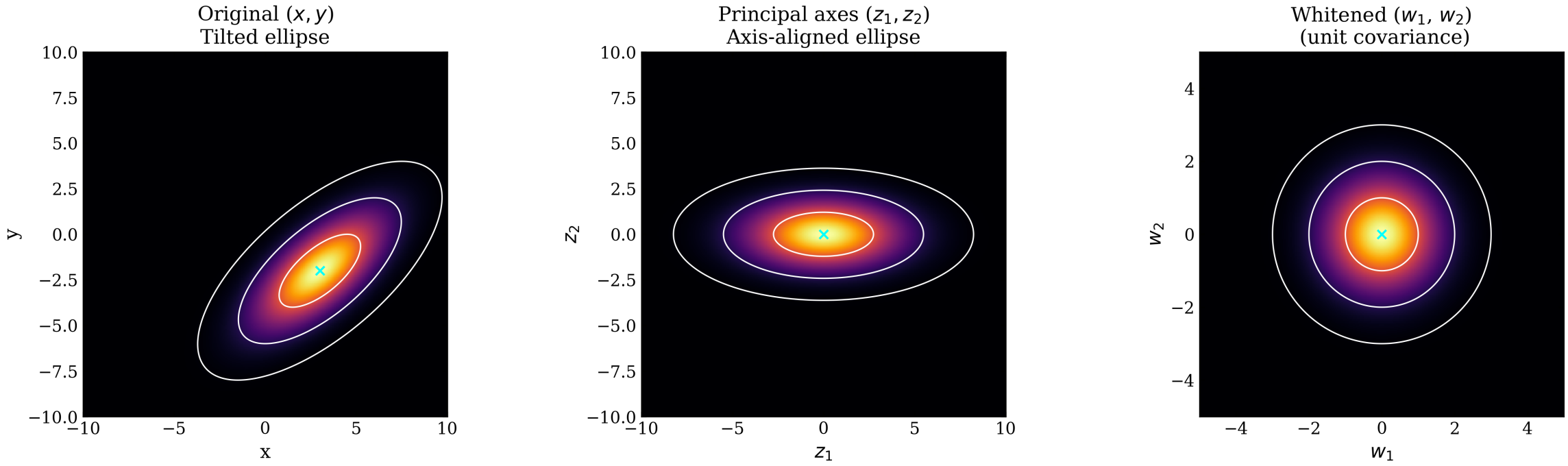
As a final step, we can rescale the data such that each axis has unit variance

- divide each variable by it's corresponding standard deviation:

$$\boldsymbol{\lambda}^{-1/2} = \mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_d}}\right)$$

$$\mathbf{w}_n = \boldsymbol{\lambda}^{-1/2} Q^T (\mathbf{x}_n - \boldsymbol{\mu})$$

Whitening


Original (x, y)
Tilted ellipse
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
y
−10
−5
0
5
10
x
Principal axes (z_1, z_2)
Axis-aligned ellipse
z_2
z_1
Whitened (w_1, w_2)
(unit covariance)
4
2
0
−2
−4
w_2
w_1

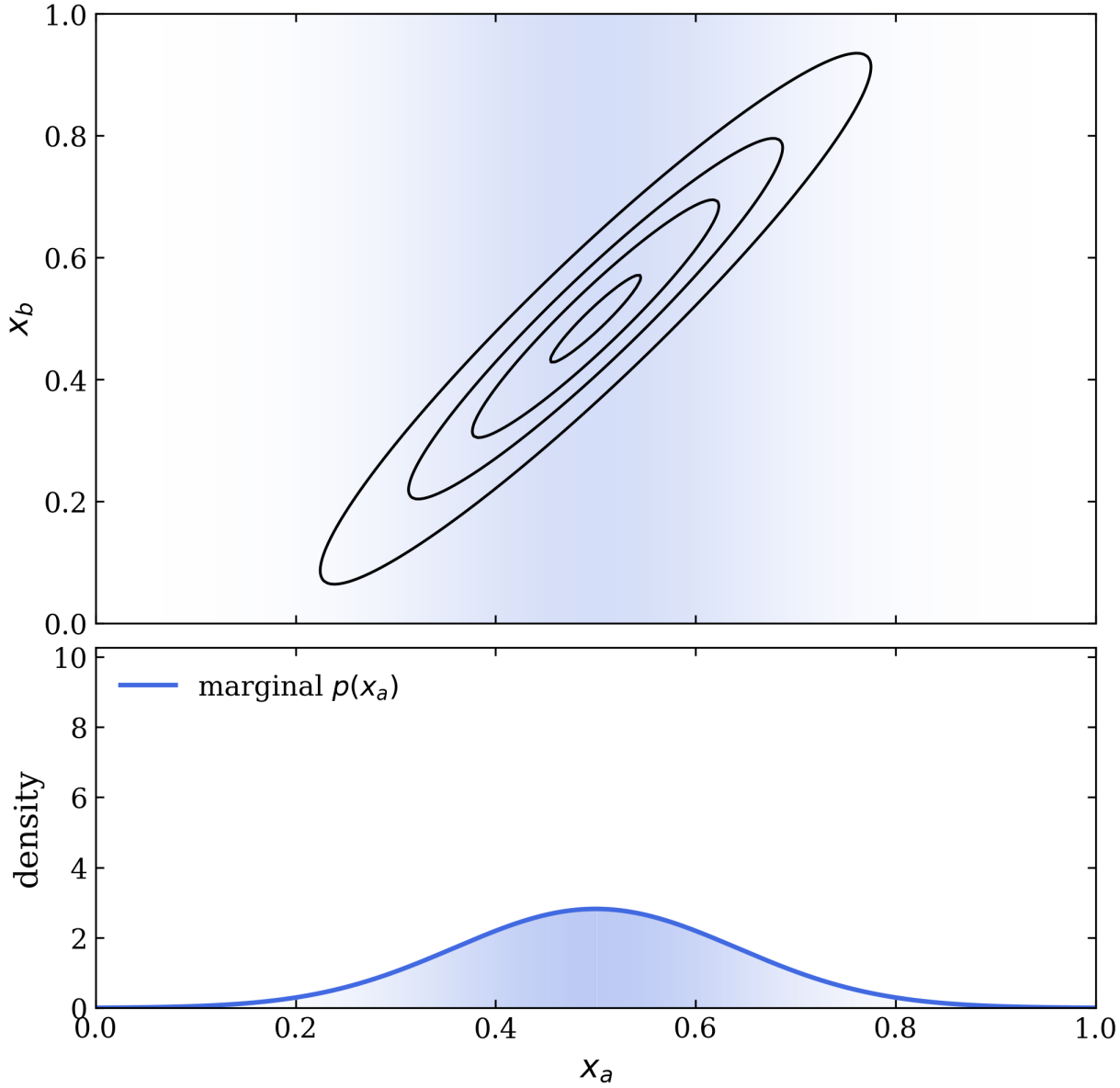
Joint, Marginal, Conditional

- Real systems often involve multiple correlated variables
- We describe them using *joint distributions* $p(x_a, x_b)$
- From the joint, we can form:
  - Marginals
  - Conditionals

Marginalization: forgetting information

- Marginal = distribution over one variable
- Obtained by integrating out the other
- Always increases uncertainty

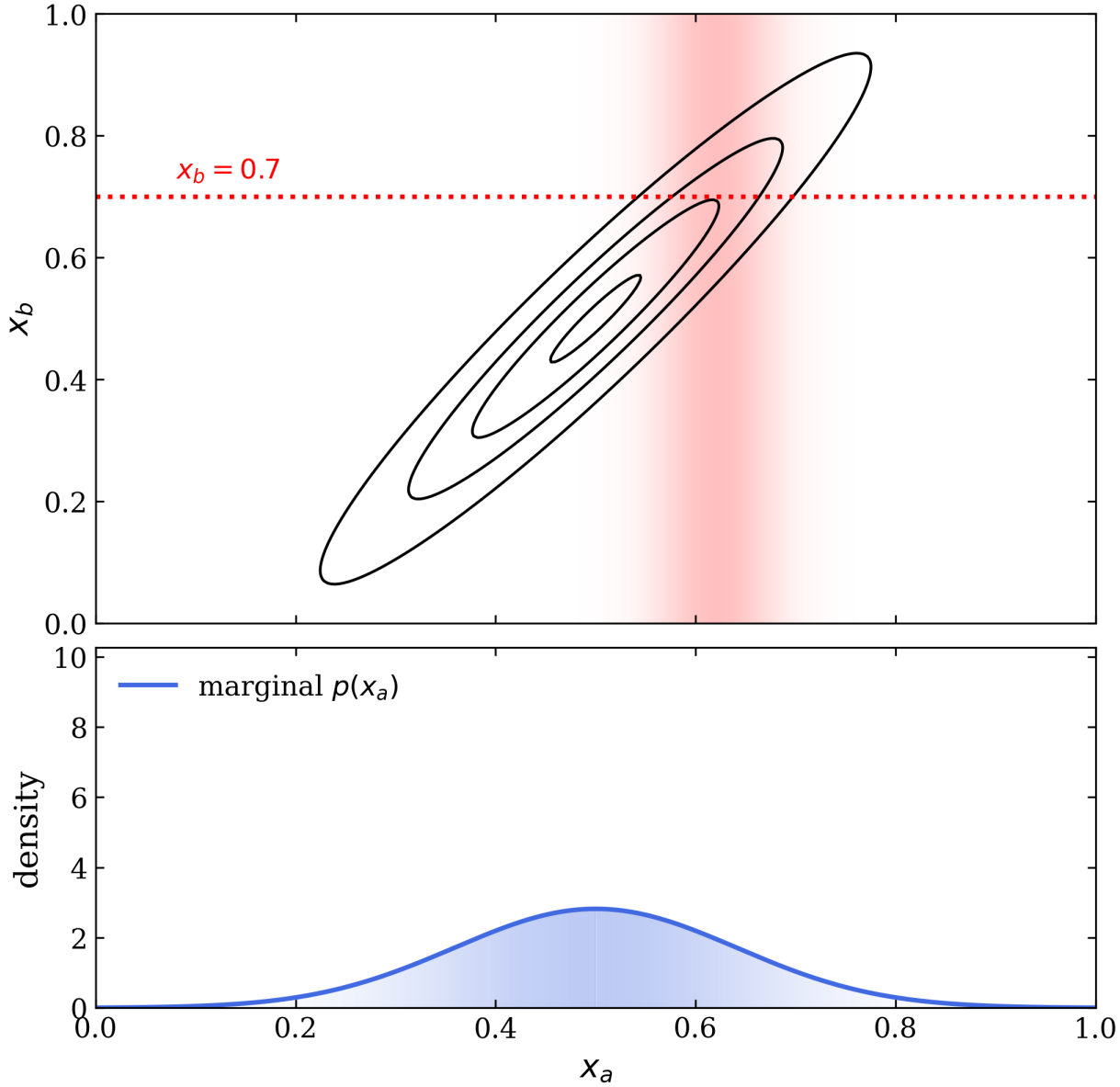
$$p(x_a) = \int p(x_a, x_b)\, dx_b$$



Conditioning: using information

- Conditional = distribution given a known value
- Uses correlations
- Uncertainty typically shrinks

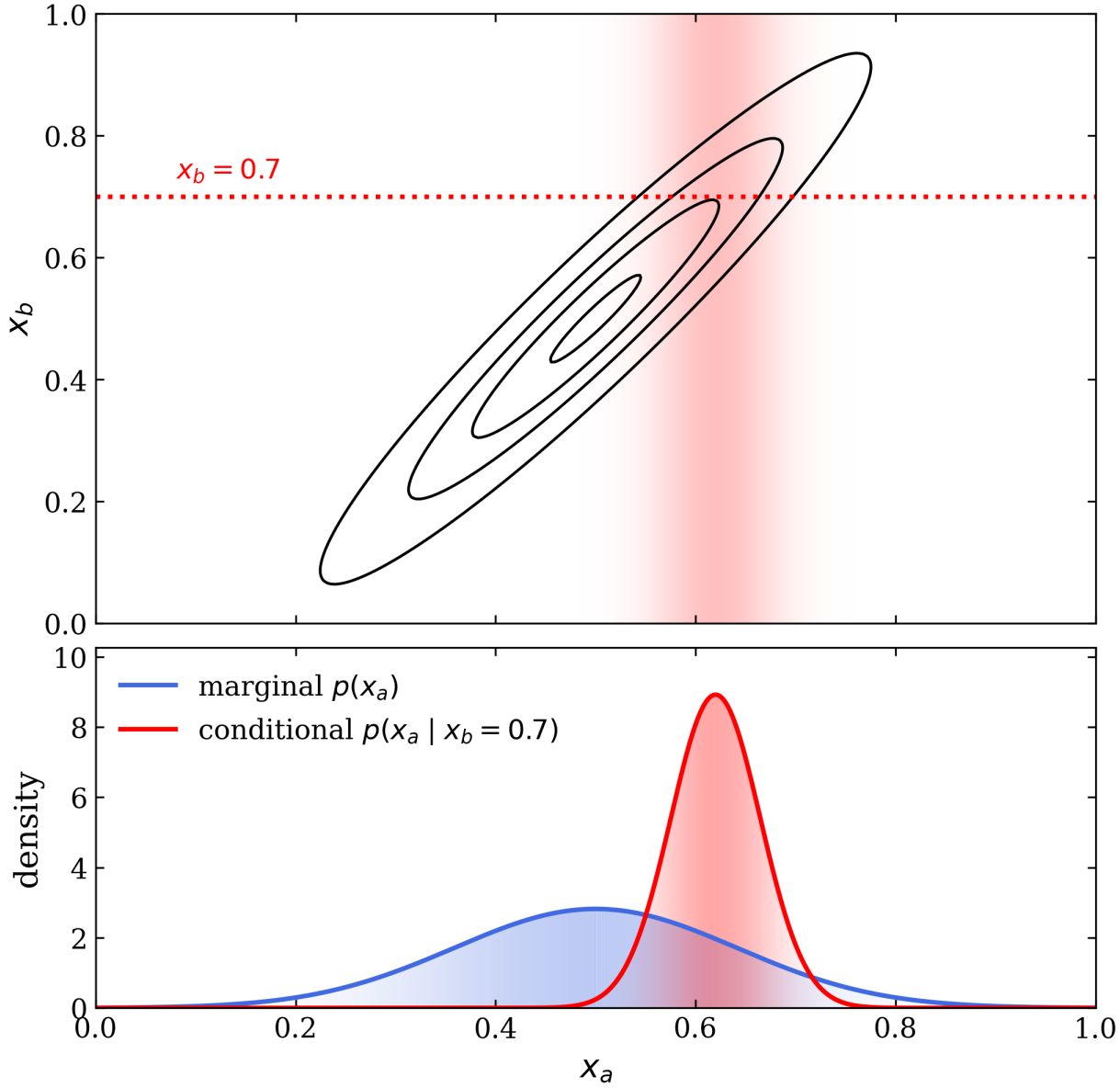
$$p(x_a \mid x_b) = \frac{p(x_a, x_b)}{p(x_b)}$$



Conditioning: using information

- Conditional = distribution given a known value
- Uses correlations
- Uncertainty typically shrinks

$$p(x_a \mid x_b) = \frac{p(x_a, x_b)}{p(x_b)}$$



Conditional Gaussian

For a joint Gaussian: $$\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \right)$$

Conditional Gaussian

For a joint Gaussian: $\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \right)$

Conditional is also Gaussian: $p(x_a \mid x_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$

Conditional Gaussian

For a joint Gaussian: $\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \right)$

Conditional is also Gaussian: $p(x_a \mid x_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

Conditional Gaussian

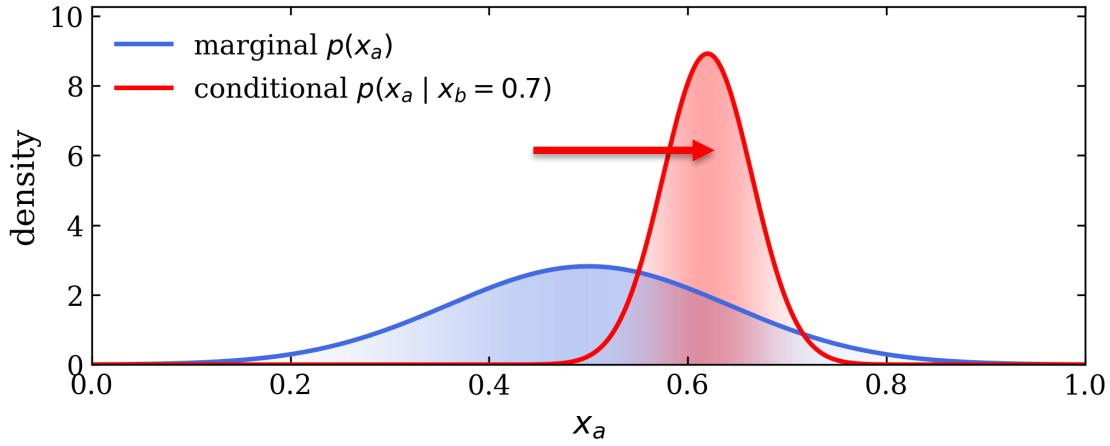
For a joint Gaussian: $$\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \right)$$

Conditional is also Gaussian: $$p(x_a \mid x_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$$

Mean shifts toward correlated value and variance decreases!

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$



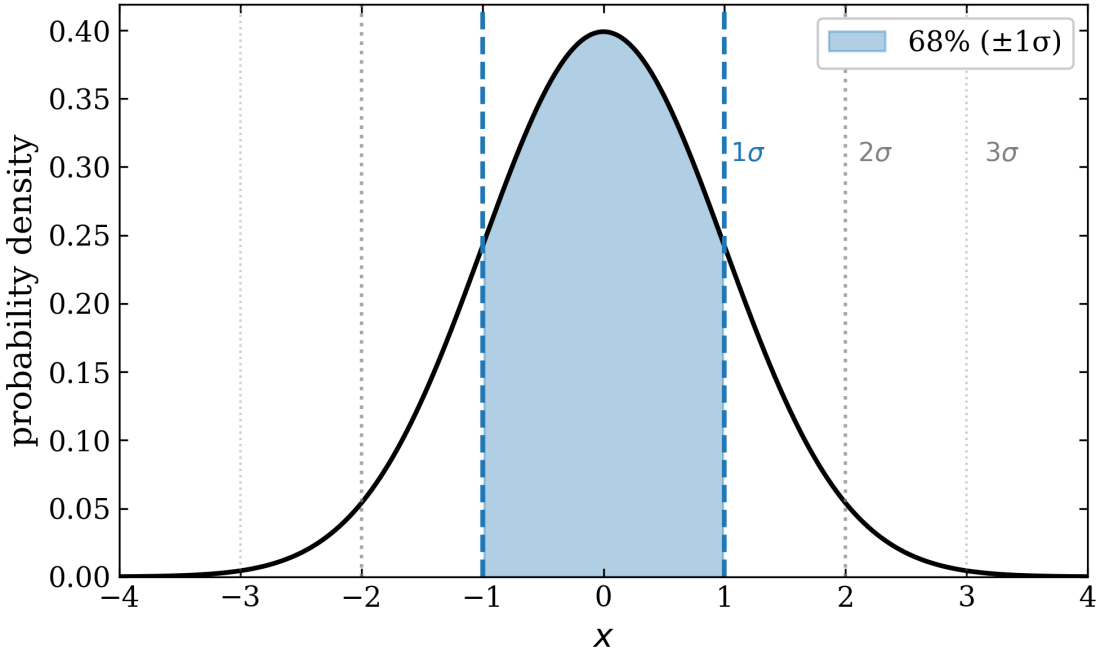
Why Gaussians are special

- ❑ Marginals of Gaussians are Gaussian
- ❑ Conditionals of Gaussians are Gaussian
- ❑ Priors × likelihoods → Gaussian posteriors

This is why Bayesian inference with Gaussians is analytically tractable!

How far from the mean is “typical”?

- For a 1D Gaussian: ‘1σ’ contains ~68% of the probability



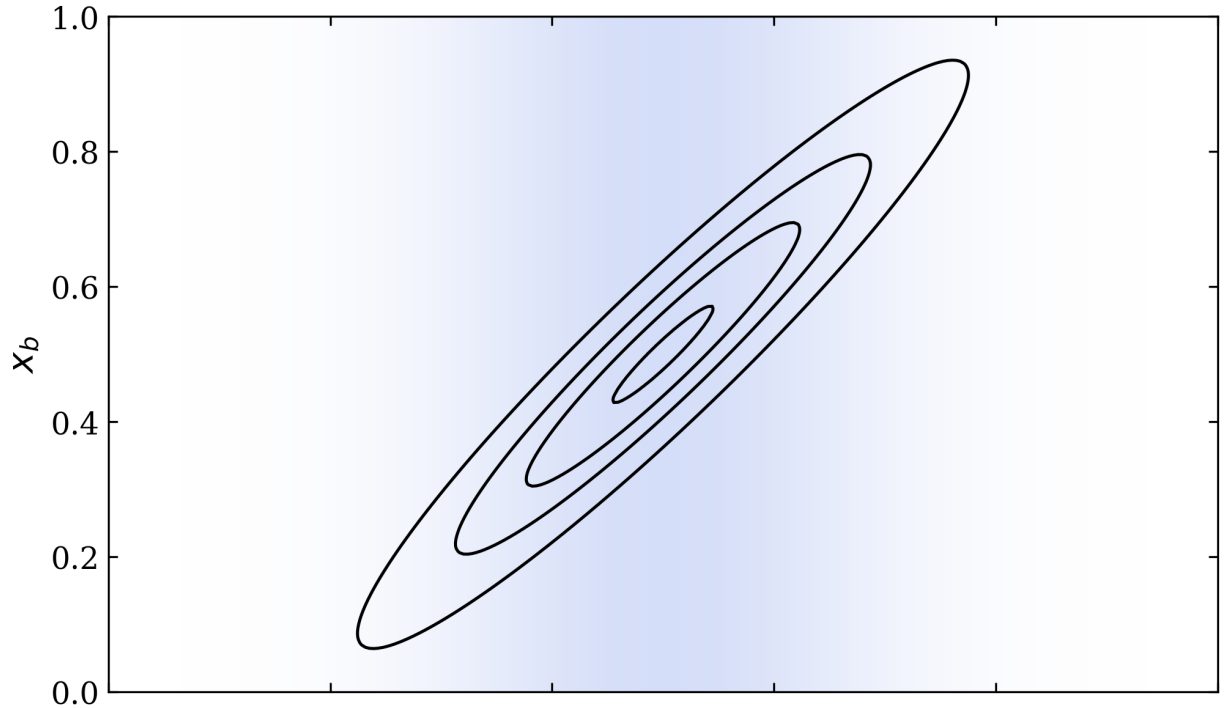
How far from the mean is “typical”?

- For a 1D Gaussian: ‘1σ’ contains ~68% of the probability

- In multiple dimensions…

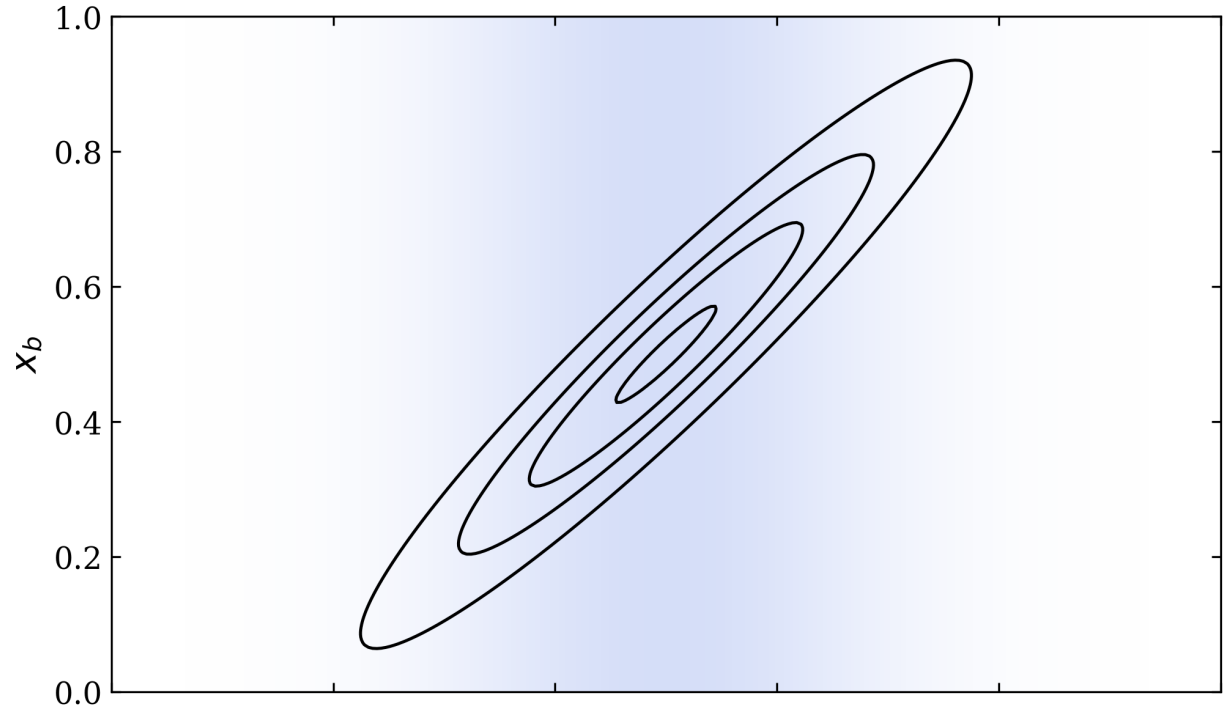
  ….. What replaces ‘1σ’?

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



How far from the mean is “typical”?

- Recall ellipses are defined by quadratic form: $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \text{const}$

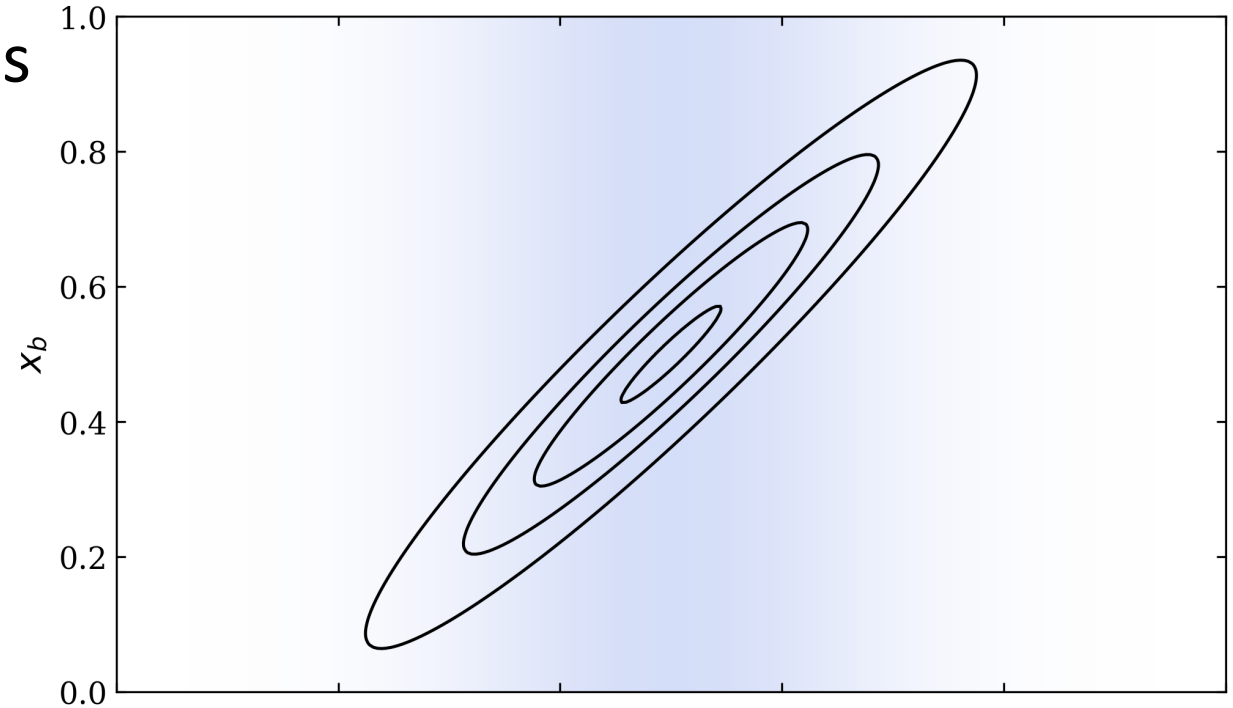


How far from the mean is “typical”?

- Recall ellipses are defined by quadratic form: $(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \text{const}$

- Write in terms of whitened variables

$$\mathbf{w} = \boldsymbol{\lambda}^{-1/2}U^T(\mathbf{x}-\boldsymbol{\mu})$$



How far from the mean is “typical”?

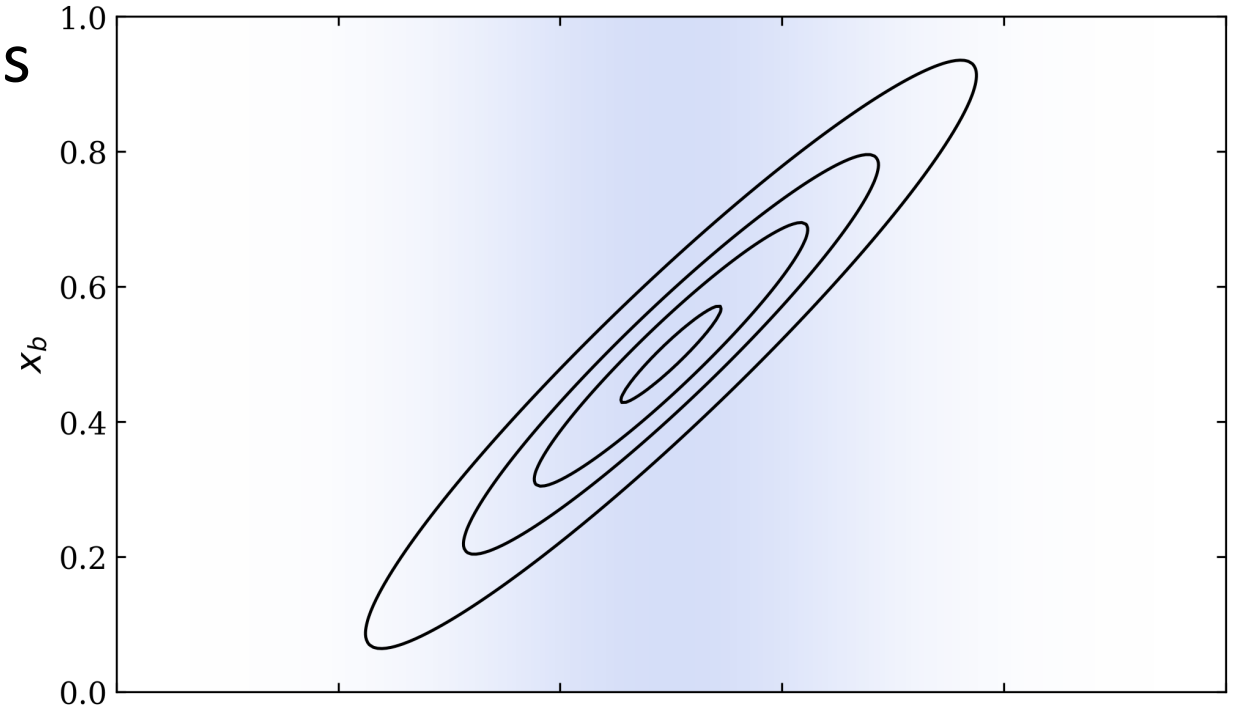
- Recall ellipses are defined by quadratic form: $(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \text{const}$

- Write in terms of whitened variables

$$\mathbf{w} = \boldsymbol{\lambda}^{-1/2}U^T(\mathbf{x}-\boldsymbol{\mu})$$

- The quadratic form can be written:

$$\sum_i w_i^2 = \text{const}$$



How far from the mean is “typical”?

- Recall ellipses are defined by quadratic form: $(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \text{const}$

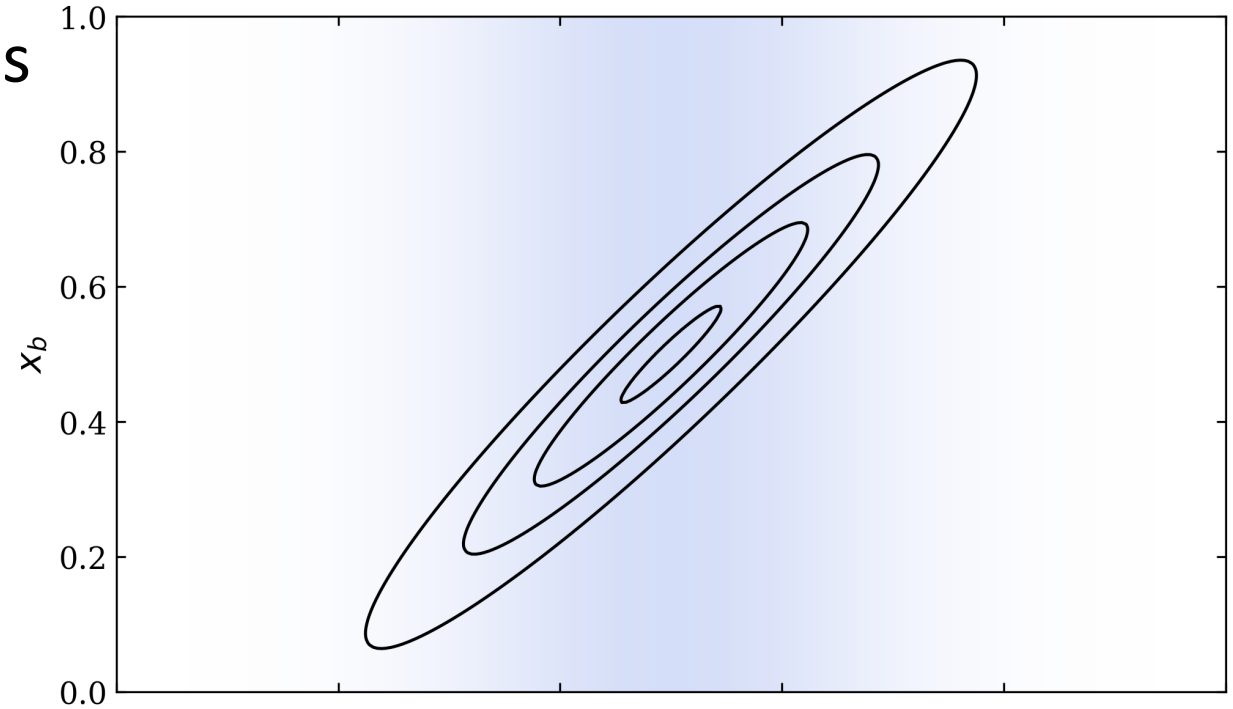
- Write in terms of whitened variables

$$\mathbf{w} = \boldsymbol{\lambda}^{-1/2}U^T(\mathbf{x}-\boldsymbol{\mu})$$

- The quadratic form can be written:

$$\sum_i w_i^2 = \text{const}$$

Notice this looks like Euclidean norm



The χ^2 distribution

- Original question: “How far from the mean is a random data point?”

The χ^2 distribution

- Original question: “How far from the mean is a random data point?”
- Equivalent to: If I draw $\mathbf{x}$ from a Gaussian, how large is $\sum_i w_i^2$?

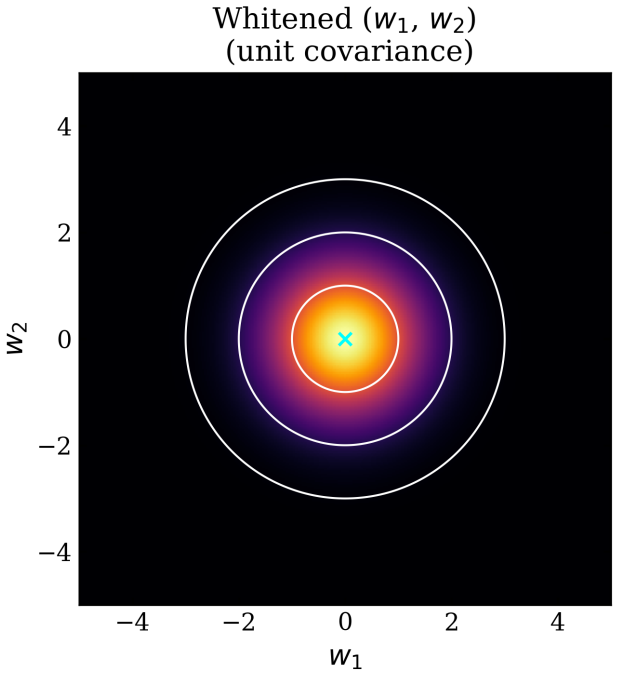
The χ^2 distribution

- Original question: “How far from the mean is a random data point?”
- Equivalent to: If I draw $\mathbf{x}$ from a Gaussian, how large is $\sum_i w_i^2$?
- Distribution of the sum of squares from normally distributed data is described by the ‘chi-squared’ distribution

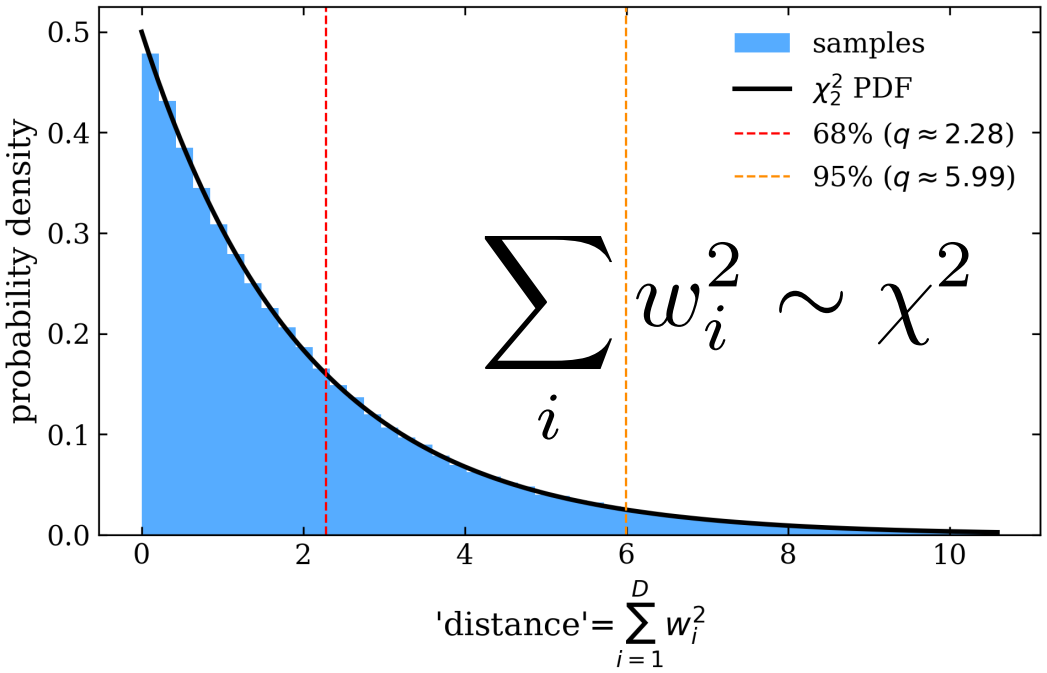
$$\sum_i w_i^2 \sim \chi^2$$

The χ^2 distribution

Two dimensional representation of geometry of covariance matrix



Distribution of ‘distances from the mean’



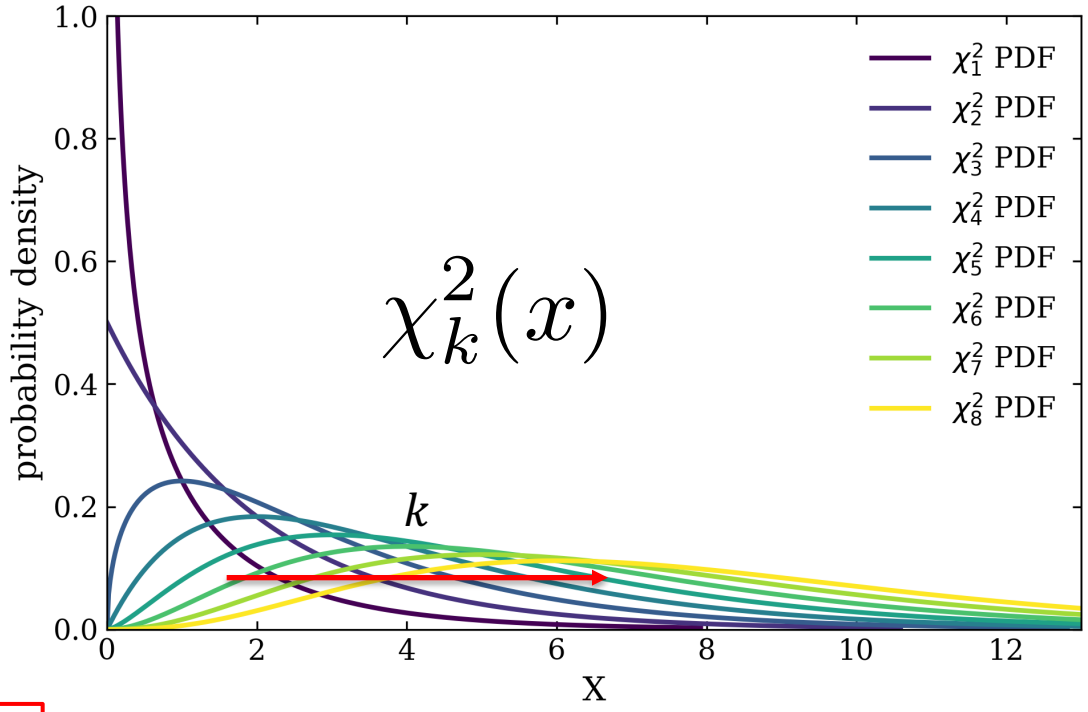
The χ^2 distribution

$$p(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$

- Parameterized by degree's of freedom: k

Mean: k

Variance: $2k$



- Gaussian, exponential, or Poisson describe *how physical quantities are generated*.

- The χ^2 distribution describes *how we analyze data* generated under Gaussian noise.

Confidence Intervals

- we are asking what value of chi-square encloses 68% of the probability.
- Recall CDF tells us the probability

$$F_{\chi^2_k}(c) = P(\chi^2 \leq c) = \int_0^c f_{\chi^2_k}(x)\,dx$$

"what fraction of draws have χ^2 less than c?"

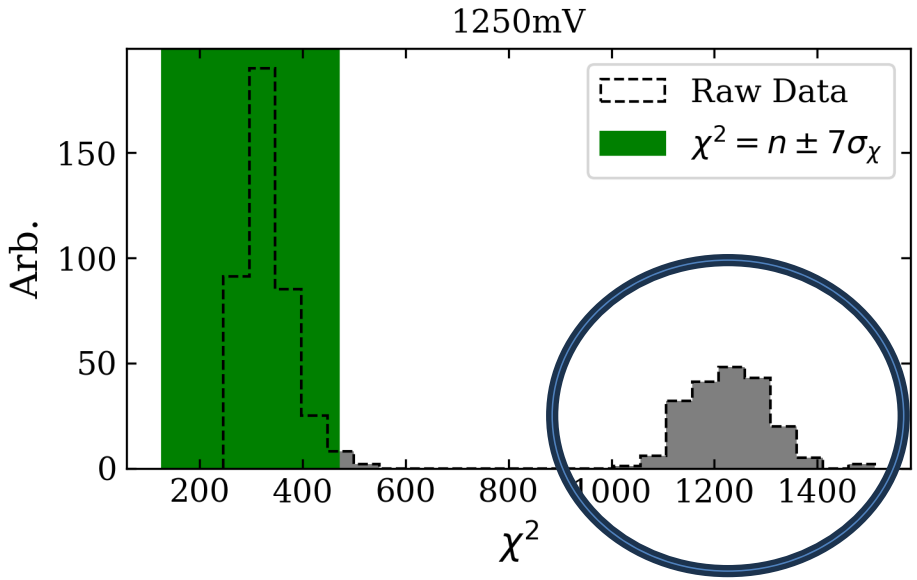
- c is given by evaluating inverse CDF at 0.68

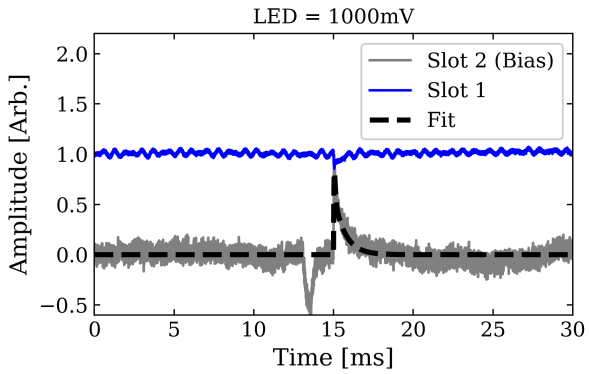
$$c = F^{-1}_{\chi^2_k}(0.68) \longrightarrow (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq c$$

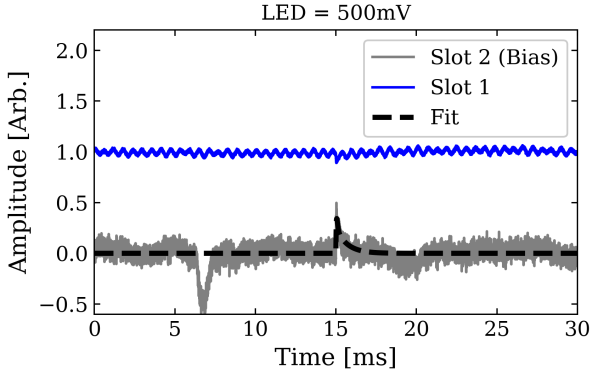
This ellipse defines the 68% confidence interval

The χ^2 metric as a discriminator

- We will see in future modules how the χ^2 metric can be used to judge the quality of your data
- "how likely is an event an expected fluctuation from my model?"







Bayesian Inference: updating beliefs with data

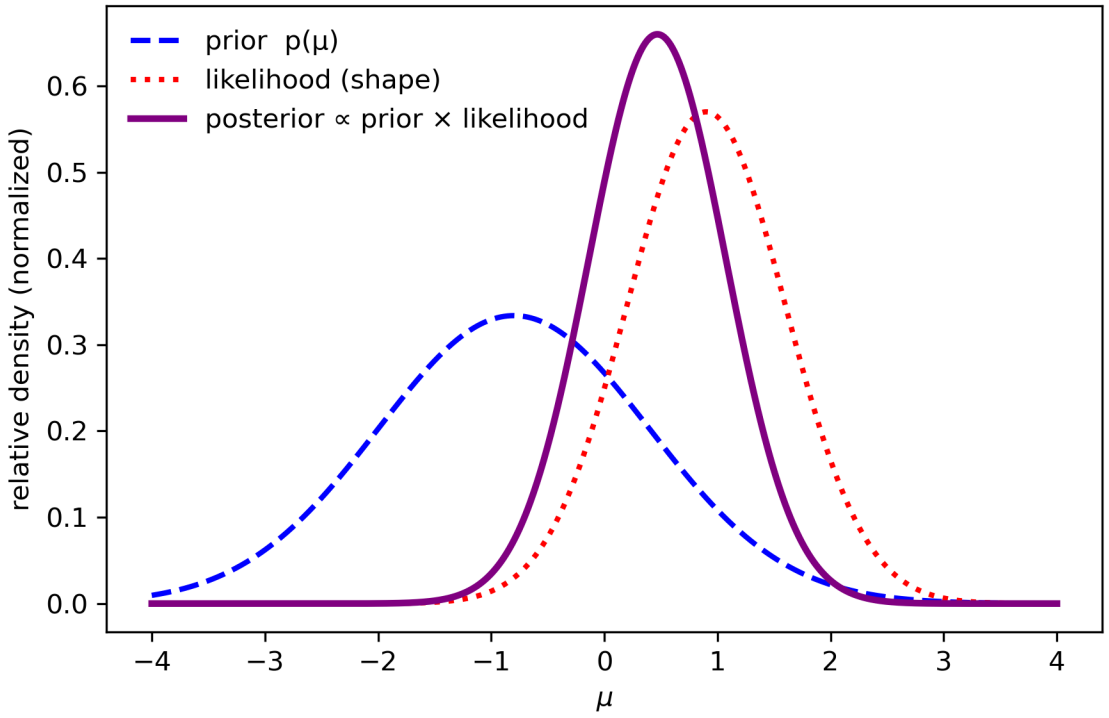
- Prior: what we believe before seeing data
- Likelihood: information from data
- Posterior: updated belief

$$p(\mu \mid \{x_n\}) \propto p(\{x_n\} \mid \mu)\, p(\mu)$$

Gaussian priors + Gaussian data

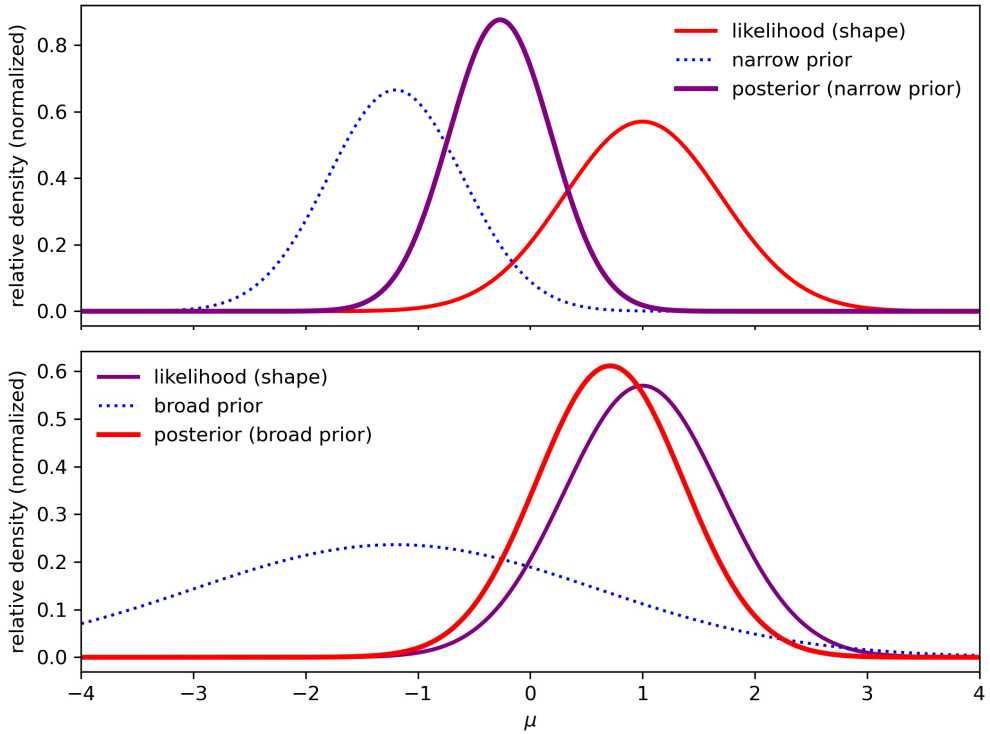
- Gaussian prior × Gaussian likelihood → Gaussian posterior
- Posterior mean = compromise
- Posterior variance shrinks with data
- Data reduces uncertainty, but does not erase prior assumptions instantly.

$$\sigma^2_{\text{post}} < \sigma^2_{\text{prior}}$$



When does the prior matter?

- Small N: prior matters a lot
- Large N: likelihood dominates
  - Posterior becomes data-driven
- As $N \to \infty$, reasonable priors are forgotten.



Next module: regression — turning probability into models