Supplementary Notes – Lec. #5

Mathematical Background for the Kalman Filter

Hyungtae Lim, Ph.D.

Massachusetts Institute of Technology (MIT)

Outline

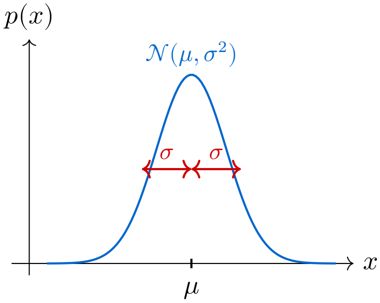
What Is a Random Variable?

A **random variable** x is a quantity whose value we do not know in advance – it is determined by chance.

Two numbers fully describe a Gaussian random variable:

- **Mean** $\mu = \mathbb{E}[x]$: the center (most likely value).
- **Variance** $\sigma^2 = \mathrm{Var}(x)$: how spread out the values are.

We write $x \sim \mathcal{N}(\mu, \sigma^2)$ and say x follows a **normal (Gaussian) distribution**.



Intuition

If you measure the height of a person many times with a ruler, the results scatter around the true height μ with spread σ.

Mean and Variance: Formal Definitions

Mean (Expected Value)

$$\mu = \mathbb{E}[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

- The average value over all possible outcomes.
- “Center of mass” of the distribution.

Variance

$$\sigma^2 = \mathsf{Var}(x) = \mathbb{E}\big[(x-\mu)^2\big]$$

- Average squared distance from the mean.
- Large $\sigma^2 \Rightarrow$ wide spread
- Small $\sigma^2 \Rightarrow$ concentrated.

Key linearity property of expectation (used everywhere):

$$\mathbb{E}[c_1 f(x) + c_2 g(x)] \;=\; c_1\ \mathbb{E}[f(x)] + c_2\ \mathbb{E}[g(x)]$$

for any constants c_1, c_2. This holds for any distribution, not just Gaussian.

Expectation is a **linear operator**: it “passes through” sums and constant multiples.

Effect of Adding a Constant: $y = x + b$

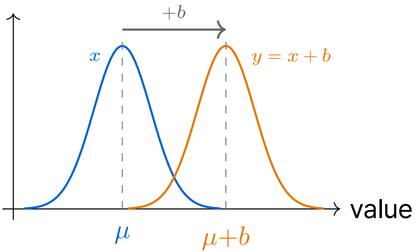
Question: If $x \sim \mathcal{N}(\mu, \sigma^2)$ and $y = x + b$ (shift by a constant b), what are $\mathbb{E}[y]$ and Var(y)?

Mean shifts by b:

$$\mathbb{E}[y] = \mathbb{E}[x + b] = \mathbb{E}[x] + b = \mu + b$$

Variance is unchanged:

$$\begin{aligned} \text{Var}(y) &= \mathbb{E}\big[(y - \mathbb{E}[y])^2\big] \\ &= \mathbb{E}\big[(x + b - \mu - b)^2\big] \\ &= \mathbb{E}\big[(x - \mu)^2\big] = \sigma^2 \end{aligned}$$



Intuition

Shifting all values by the same amount moves the center but does not change the spread.

Effect of Scaling: $y = ax$

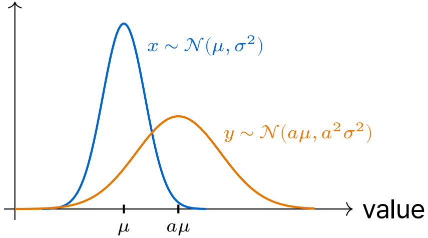
Question: If $x \sim \mathcal{N}(\mu, \sigma^2)$ and $y = ax$ (scale by constant a), what are $\mathbb{E}[y]$ and Var(y)?

Mean scales by a:

$$\mathbb{E}[y] = \mathbb{E}[ax] = a\ \mathbb{E}[x] = a\mu$$

Variance scales by a^2:

$$\begin{aligned}\text{Var}(y) &= \mathbb{E}\big[(ax - a\mu)^2\big] \\ &= \mathbb{E}\big[a^2(x - \mu)^2\big] \\ &= a^2\ \mathbb{E}\big[(x - \mu)^2\big] = a^2\sigma^2\end{aligned}$$



Key point

The variance scales by $a^{\mathbf{2}}$, not a. Doubling x quadruples the variance.

Combined: $y = ax + b$

Combining the two results:

If $x \sim \mathcal{N}(\mu, \sigma^2)$ **and** $y = ax + b$**, then**

$$\mathbb{E}[y] = a\mu + b$$

$$\text{Var}(y) = a^2\sigma^2$$

$$y \sim \mathcal{N}(a\mu + b,\ a^2\sigma^2)$$

Summary: Linear Transformation Rules

Rules (1D)

Transform	Mean	Variance
$y = x + b$	$\mu + b$	σ^2
$y = ax$	$a\mu$	$a^2\sigma^2$
$y = ax + b$	$a\mu + b$	$a^2\sigma^2$

Quick memory aid

- Mean: follows the same arithmetic as x.
- Variance: only scaling matters, and it squares.
- A constant shift b never affects variance.

Why does variance square? Variance measures squared deviations. Scaling x by a scales each deviation by a, and the squared deviation by a^2; hence the variance multiplies by a^2.

Let’s Proof Prediction Step of Kalman filter in 1D

Connection to the Kalman Filter Prediction Step

The motion model is $x_k = x_{k-1} + u_k + w_k$ where $w_k \sim \mathcal{N}(0, \sigma_w^2)$.

Treating x_{k-1} as a Gaussian random variable with mean $\hat{x}_{k-1}$ and variance P_{k-1}:

- $a = 1$, $b = u_k + w_k \Rightarrow$ mean propagates: $\bar{x}_k = \hat{x}_{k-1} + u_k$
- Variance grows by σ_w^2: $\bar{P}_k = P_{k-1} + \sigma_w^2$

Independence and Why It Matters

Two random variables x and w are **independent** if knowing the value of one tells you nothing about the other.

Independent (in the Kalman Filter)

- The robot's current position x_{k-1}.
- The process noise w_k (random jitter in motion).
- These are independent: the noise does not depend on where the robot is.

Not independent (counterexample)

- Temperature T and ice cream sales S.
- Knowing T is high tells us S is likely high too – they are correlated.

Formal condition: x and w are independent if and only if

$$p(x, w) = p(x)\, p(w).$$

For independent variables: $\mathbb{E}[xw] = \mathbb{E}[x]\ \mathbb{E}[w]$.

Mean of a Sum

Claim: For any two random variables x and w (independent or not),

$$\mathbb{E}[x + w] = \mathbb{E}[x] + \mathbb{E}[w].$$

Proof (using linearity of expectation):

$$\begin{aligned}\mathbb{E}[x + w] &= \int\!\!\int (x + w)\, p(x, w)\, dx\, dw \\ &= \int\!\!\int x\, p(x, w)\, dx\, dw \;+\; \int\!\!\int w\, p(x, w)\, dx\, dw = \mathbb{E}[x] + \mathbb{E}[w].\end{aligned}$$

Kalman Filter application

In the motion model $x_k = x_{k-1} + u_k + w_k$, taking the expectation:

$$\mathbb{E}[x_k] = \mathbb{E}[x_{k-1}] + u_k + \underbrace{\mathbb{E}[w_k]}_{=0} \quad \Longrightarrow \quad \bar{x}_k = \hat{x}_{k-1} + u_k.$$

The noise term vanishes because $w_k \sim \mathcal{N}(0, \sigma_w^2)$ has zero mean.

Variance of a Sum (Independent Case)

Claim: If x and w are independent,

$$\text{Var}(x+w) = \text{Var}(x) + \text{Var}(w).$$

Proof: Let $\mu_x = \mathbb{E}[x]$, $\mu_w = \mathbb{E}[w]$.

$$\begin{aligned} \text{Var}(x+w) &= \mathbb{E}\big[((x+w)-(\mu_x+\mu_w))^2\big] = \mathbb{E}\big[(x-\mu_x)^2 + 2(x-\mu_x)(w-\mu_w) + (w-\mu_w)^2\big] \\ &= \text{Var}(x) + 2\ \underbrace{\mathbb{E}[(x-\mu_x)(w-\mu_w)]}_{\text{cross term}} + \text{Var}(w). \end{aligned}$$

The cross term is the **covariance**. For independent variables:

$$\mathbb{E}[(x-\mu_x)(w-\mu_w)] = \mathbb{E}[x-\mu_x]\ \mathbb{E}[w-\mu_w] = 0 \cdot 0 = 0.$$

Therefore $\text{Var}(x+w) = \text{Var}(x) + \text{Var}(w)$. $\square$

Why the Predicted Variance Grows

Recall the motion model:

$$x_k = \underbrace{x_{k-1}}_{\mathcal{N}(\hat{x}_{k-1}, P_{k-1})} + u_k + \underbrace{w_k}_{\mathcal{N}(0, \sigma_w^2)}$$

where x_{k-1} and w_k are **independent**.

Applying the sum rules step by step:

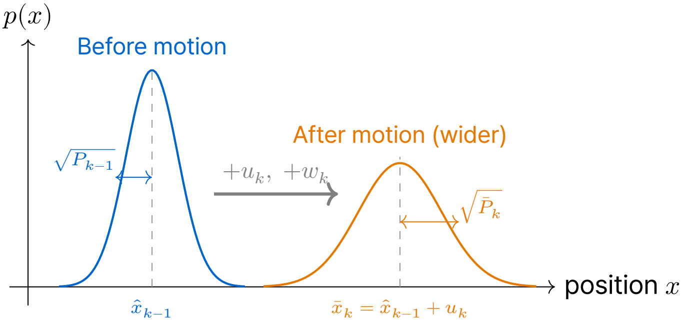
Step	Expression	Result
$y_1 = x_{k-1} + u_k$	adding constant u_k	$\text{Var}(y_1) = P_{k-1}$
$x_k = y_1 + w_k$	adding independent noise	$\text{Var}(x_k) = P_{k-1} + \sigma_w^2$

Predicted variance:

$$\bar{P}_k = P_{k-1} + \sigma_w^2$$

Each motion step adds uncertainty from the process noise. This is why the Kalman Filter prediction step always increases P.

Visual Intuition: Variance Accumulates



The orange (post-motion) distribution is **wider** because motion noise σ_w^2 has been added to the variance.

Summary: Rules for Sum of Independent Random Variables

For independent $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $w \sim \mathcal{N}(\mu_w, \sigma_w^2)$:

$$x + w \;\sim\; \mathcal{N}(\mu_x + \mu_w,\;\; \sigma_x^2 + \sigma_w^2)$$

What adds

- Means always add: $\mathbb{E}[x + w] = \mathbb{E}[x] + \mathbb{E}[w]$.
- Variances add only when independent.

What does NOT add in general

- Variances of dependent variables: there is an extra cross-term (i.e., covariance).
- Standard deviations σ do not add: $\sigma_{x+w} \neq \sigma_x + \sigma_w$.

Kalman Filter summary so far:

$$\bar{x}_k = \hat{x}_{k-1} + u_k, \qquad \bar{P}_k = P_{k-1} + \sigma_w^2.$$

These are exact consequences of the two rules above.

The Core Problem: Two Noisy Estimates

After the prediction step we have **two estimates** of the same quantity x_k:

Prediction (prior)

$$\bar{x}_k \sim \mathcal{N}(\bar{x}_k,\ \bar{P}_k)$$

- Comes from the motion model.
- Has variance $\bar{P}_k$ (uncertainty from motion noise).

Measurement (likelihood)

$$z_k \sim \mathcal{N}(x_k,\ \sigma_v^2)$$

- Comes from the sensor.
- Has variance σ_v^2 (sensor noise).

Goal: Combine them into a single best estimate $\hat{x}_k$ that is more accurate than either alone.

The answer is a **weighted average** – but what weights should we use?

Precision = Inverse Variance

The key idea: weight each estimate by how **reliable** (precise) it is.

Precision ω

$$\omega = \frac{1}{\sigma^2}$$

- Large σ^2 (uncertain) $\Rightarrow$ small ω (low weight).
- Small σ^2 (precise) $\Rightarrow$ large ω (high weight).

Analogy

Trust a careful grader more than a sloppy one: give a higher weight proportional to reliability.

Define the two precisions: $\omega_{\text{pred}} = 1/\bar{P}_k$, $\omega_{\text{meas}} = 1/\sigma_v^2$.

The **precision-weighted average** is:

$$\hat{x}_k = \frac{\omega_{\text{pred}}\, \bar{x}_k + \omega_{\text{meas}}\, z_k}{\omega_{\text{pred}} + \omega_{\text{meas}}} = \frac{\bar{x}_k/\bar{P}_k \ + \ z_k/\sigma_v^2}{1/\bar{P}_k \ + \ 1/\sigma_v^2}.$$

Deriving the Kalman Gain Step by Step

Start from the weighted average and rewrite it in Kalman form.

Step 1. Multiply numerator and denominator by $\bar{P}_k \sigma_v^2$:

$$\hat{x}_k = \frac{\sigma_v^2 \bar{x}_k + \bar{P}_k z_k}{\sigma_v^2 + \bar{P}_k}.$$

Step 2. Add and subtract $\bar{P}_k \bar{x}_k$ in the numerator:

$$\hat{x}_k = \frac{(\sigma_v^2 + \bar{P}_k)\bar{x}_k + \bar{P}_k(z_k - \bar{x}_k)}{\sigma_v^2 + \bar{P}_k} = \bar{x}_k + \frac{\bar{P}_k}{\bar{P}_k + \sigma_v^2}(z_k - \bar{x}_k).$$

Step 3. Define the **Kalman gain**:

$$K_k \; := \; \frac{\bar{P}_k}{\bar{P}_k + \sigma_v^2}.$$

Kalman update:

$$\hat{x}_k = \bar{x}_k + K_k \, (z_k - \bar{x}_k)$$

Interpreting the Kalman Gain

$$K_k = \frac{\bar{P}_k}{\bar{P}_k + \sigma_v^2}, \qquad 0 \le K_k \le 1.$$

$K_k \to 1$ (trust the measurement)

Occurs when $\sigma_v^2 \ll \bar{P}_k$:

$$\hat{x}_k \approx \bar{x}_k + 1 \cdot (z_k - \bar{x}_k) = z_k.$$

The estimate jumps to the measurement.
"The prediction was very uncertain;
thus, the sensor is precise."

$K_k \to 0$ (trust the prediction)

Occurs when $\sigma_v^2 \gg \bar{P}_k$:

$$\hat{x}_k \approx \bar{x}_k + 0 \cdot (z_k - \bar{x}_k) = \bar{x}_k.$$

The estimate stays near the prediction.
"The prediction was good;
the sensor is noisy."

The term $(z_k - \bar{x}_k)$ is called the **innovation term**: how much the measurement "surprises" the prediction.

Updated Variance

After fusing the two estimates, the uncertainty decreases.

Recall: $\hat{x}_k = \bar{x}_k + K_k(z_k - \bar{x}_k) = (1 - K_k)\bar{x}_k + K_k z_k$.
Applying the variance rule for a weighted combination of independent estimates:

$$P_k = \mathrm{Var}(\hat{x}_k) = (1 - K_k)^2 \bar{P}_k + K_k^2 \sigma_v^2.$$

Substituting $K_k = \bar{P}_k / (\bar{P}_k + \sigma_v^2)$ and simplifying (algebra):

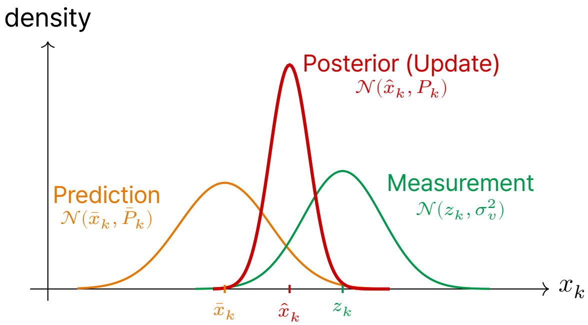
$$P_k = (1 - K_k)\, \bar{P}_k.$$

Since $0 \leq K_k \leq 1$, we have $(1 - K_k) \leq 1$, so:

$$P_k = (1 - K_k)\bar{P}_k \ \leq \ \bar{P}_k$$

The update step always reduces uncertainty.

Visual Summary: Prediction then Update



The posterior (red) is **narrower** than both inputs and lies between them, pulled toward the more precise one.

Kalman Filter 1D: Complete Equations

Prediction

$$\bar{x}_k = \hat{x}_{k-1} + u_k$$

$$\bar{P}_k = P_{k-1} + \sigma_w^2$$

(By mean adds, variance adds.)

Update

$$K_k = \frac{\bar{P}_k}{\bar{P}_k + \sigma_v^2}$$

$$\hat{x}_k = \bar{x}_k + K_k(z_k - \bar{x}_k)$$

$$P_k = (1 - K_k)\,\bar{P}_k$$

(By precision-weighted average.)

The Kalman Filter is nothing more than **repeated application** of “add Gaussian noise” (predict) and “precision-weighted average” (update).

From 1D to Multi-Dimensional State

So far we tracked a **single number** x_k (1D position).
However, real robots have many unknowns at once:

2D example: position + velocity

$$\boldsymbol{x}_k = \begin{pmatrix} p_k \\ v_k \end{pmatrix} \in \mathbb{R}^2 \quad \text{(a 2D vector)}$$

Motion: $p_k = p_{k-1} + v_{k-1}\Delta t, \quad v_k = v_{k-1}.$

2D robot pose

$$\boldsymbol{x}_k = \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathbb{R}^3$$

Position (x, y) and heading θ.

We now need:

- A **mean vector** $\boldsymbol{\mu}$ (instead of a scalar μ).
- A **covariance matrix** Σ (instead of a scalar σ^2).
- **Matrix arithmetic** to propagate uncertainty.

Matrix-Vector Multiplication: What Does Ax Mean?

A matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ applied to a vector $\boldsymbol{x} \in \mathbb{R}^n$ produces a new vector $\boldsymbol{y} \in \mathbb{R}^m$.

Example (2×2 matrix):

$$\boldsymbol{A} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} p \\ v \end{pmatrix}, \qquad \boldsymbol{Ax} = \begin{pmatrix} 1\cdot p + \Delta t \cdot v \\ 0 \cdot p + 1 \cdot v \end{pmatrix} = \begin{pmatrix} p + v\Delta t \\ v \end{pmatrix}.$$

This encodes the motion model $p_k = p_{k-1} + v_{k-1}\Delta t$, $v_k = v_{k-1}$ compactly as:

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1}.$$

Rule: row · column

The i-th entry of $\boldsymbol{Ax}$ is the dot product of the i-th row of $\boldsymbol{A}$ with the vector $\boldsymbol{x}$:

$$(\boldsymbol{Ax})_i = \sum_{j=1}^{n} A_{ij}\, x_j.$$

Transpose: $\boldsymbol{A}^\top$

The **transpose** of a matrix $\boldsymbol{A}$ flips rows and columns:

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{A}^\top = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Properties used in the Kalman Filter:

- $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$.
- $(\boldsymbol{A}\boldsymbol{B})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$ (order reverses).
- $(\boldsymbol{A}\boldsymbol{x})^\top = \boldsymbol{x}^\top \boldsymbol{A}^\top$.
- If $\boldsymbol{A} = \boldsymbol{A}^\top$, the matrix is called **symmetric**. Covariance matrices are always symmetric.

Why does the covariance propagation formula have $\boldsymbol{A}^\top$?
When computing variance of $\boldsymbol{A}\boldsymbol{x}$, terms like $\mathbb{E}[\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{A}^\top]$ arise naturally – the transpose comes from transposing the outer product.

Matrix Inverse: $\boldsymbol{A}^{-1}$

For a square matrix $\boldsymbol{A}$, its **inverse** $\boldsymbol{A}^{-1}$ satisfies:

$$\boldsymbol{A}\,\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \quad \text{(identity matrix)}.$$

Analogy: $\boldsymbol{A}^{-1}$ is the matrix version of dividing by $\boldsymbol{A}$.

2×2 **inverse formula:**

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Rightarrow \boldsymbol{A}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

The scalar $ad - bc$ is the **determinant**.

In the Kalman Filter

The Kalman gain involves $(\bar{\boldsymbol{\Sigma}}_k + \boldsymbol{R})^{-1}$: the combined-uncertainty precision, analogous to $(\bar{P}_k + \sigma_v^2)^{-1}$ in 1D.

Scalar $\longleftrightarrow$ matrix analogy:

1D scalar	Matrix
$K_k = \bar{P}_k(\bar{P}_k + \sigma_v^2)^{-1}$	$\boldsymbol{K} = \bar{\boldsymbol{\Sigma}}_k(\bar{\boldsymbol{\Sigma}}_k + \boldsymbol{R})^{-1}$

The Covariance Matrix Σ

For a random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\top$ with mean $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$, the **covariance matrix** is:

$$\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}) = \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\right].$$

The (i, j) entry is $\Sigma_{ij} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = \text{Cov}(x_i, x_j)$.

Diagonal entries

$\Sigma_{ii} = \text{Var}(x_i) = \sigma_i^2$

The variance of each component.

Off-diagonal entries

$\Sigma_{ij} = \text{Cov}(x_i, x_j)$

How much x_i and x_j vary together.

$= 0$ if they are independent.

Example: position + velocity

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_p^2 & \sigma_{pv} \\ \sigma_{pv} & \sigma_v^2 \end{pmatrix}$$

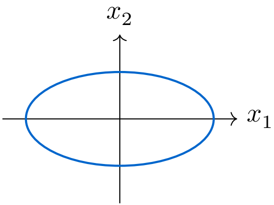
- σ_p^2: uncertainty in position.
- σ_v^2: uncertainty in velocity.
- σ_{pv}: correlation (if v is high, is p likely to be high too?).

Reading a Covariance Matrix

Diagonal (independent components):

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

$\sigma_x = 2$, $\sigma_y = 1$, no correlation.

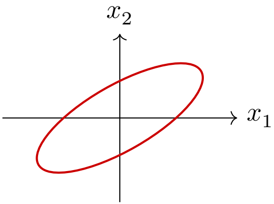


axis-aligned ellipse

Off-diagonal (correlated components):

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1.5 \\ 1.5 & 1 \end{pmatrix}$$

Positive correlation: both large or both small.



tilted ellipse (correlated)

$\boldsymbol{\Sigma}$ is always **symmetric** ($\Sigma_{ij} = \Sigma_{ji}$) and **positive semi-definite** (all eigenvalues ≥ 0). The ellipse shape shows the uncertainty region.

Linear Transformation of a Random Vector: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

Claim: If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, then:

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{\mu}, \qquad \mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top.$$

Proof of the mean:

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\ \mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu}.$$

Proof of the covariance:

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{y}) &= \mathbb{E}\big[(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu})^\top\big] = \mathbb{E}\big[\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{A}^\top\big] \\ &= \boldsymbol{A}\ \underbrace{\mathbb{E}\big[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\big]}_{\boldsymbol{\Sigma}}\ \boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top.\end{aligned}$$

Scalar analogy: if $y = ax$ then $\mathrm{Var}(y) = a^2\sigma^2$.

Vector version: if $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ then $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$.

The matrix $\boldsymbol{A}$ plays the role of a, and $\boldsymbol{A}^\top$ of a again (since $a^2 = a \cdot a$).

Multivariable Prediction Step Explained

The multivariable motion model is:

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k, \qquad \boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q}).$$

Apply our rules step by step:

Step	Operation	Result
Mean of $\boldsymbol{A}\boldsymbol{x}_{k-1}$	linear map of mean	$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}_{k-1}] = \boldsymbol{A}\hat{\boldsymbol{x}}_{k-1}$
Add $\boldsymbol{B}\boldsymbol{u}_k$	deterministic shift	$\bar{\boldsymbol{x}}_k = \boldsymbol{A}\hat{\boldsymbol{x}}_{k-1} + \boldsymbol{B}\boldsymbol{u}_k$
Cov of $\boldsymbol{A}\boldsymbol{x}_{k-1}$	linear map of cov	$\boldsymbol{A}\boldsymbol{\Sigma}_{k-1}\boldsymbol{A}^\top$
Add $\boldsymbol{w}_k$ (independent)	variances add	$\bar{\boldsymbol{\Sigma}}_k = \boldsymbol{A}\boldsymbol{\Sigma}_{k-1}\boldsymbol{A}^\top + \boldsymbol{Q}$

$$\bar{\boldsymbol{x}}_k = \boldsymbol{A}\hat{\boldsymbol{x}}_{k-1} + \boldsymbol{B}\boldsymbol{u}_k$$
$$\bar{\boldsymbol{\Sigma}}_k = \boldsymbol{A}\boldsymbol{\Sigma}_{k-1}\boldsymbol{A}^\top + \boldsymbol{Q}$$

Identical structure to the 1D case, now with matrices.

The Measurement Matrix $\boldsymbol{H}$

In the multivariable case, the sensor may only observe part of the state vector:

$$\boldsymbol{z}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}_k, \qquad \boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R}).$$

Example: State $\boldsymbol{x} = (p, v)^\top$ but sensor only measures position p:

$$\boldsymbol{H} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad \boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} p \\ v \end{pmatrix} = p.$$

What $\boldsymbol{H}$ does

- Projects the full state into the measurement space.
- Rows of $\boldsymbol{H}$ select which combinations of state variables are observed.
- $\boldsymbol{H} = \boldsymbol{I}$ (identity) means all states are directly measured.

Measurement noise $\boldsymbol{R}$

- $\boldsymbol{R}$ is the covariance of sensor noise $\boldsymbol{v}_k$.
- Analogous to σ_v^2 in the 1D case.
- Diagonal $\boldsymbol{R}$: independent noise on each sensor axis.

Summary: 1D vs. Multivariable

	1D	Multivariable
State	scalar x	vector $\boldsymbol{x} \in \mathbb{R}^n$
Mean	μ	$\boldsymbol{\mu} \in \mathbb{R}^n$
Uncertainty	variance σ^2	covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$
State transition	$x_k = a x_{k-1} + \ldots$	$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \ldots$
Cov. propagation	$a^2\sigma^2$	$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$
Variance of sum	$\sigma^2 + \sigma_w^2$	$\boldsymbol{\Sigma} + \boldsymbol{Q}$
Kalman gain	$K = \bar{P}/(\bar{P} + \sigma_v^2)$	$\boldsymbol{K} = \bar{\boldsymbol{\Sigma}}\boldsymbol{H}^\top(\boldsymbol{H}\bar{\boldsymbol{\Sigma}}\boldsymbol{H}^\top + \boldsymbol{R})^{-1}$
State update	$\hat{x} = \bar{x} + K(z - \bar{x})$	$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{K}(\boldsymbol{z} - \boldsymbol{H}\bar{\boldsymbol{x}})$
Cov. update	$(1 - K)\bar{P}$	$(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\bar{\boldsymbol{\Sigma}}$

Every 1D rule has a direct matrix analogue: scalars $\rightarrow$ vectors/matrices, division $\rightarrow$ matrix inverse.

Why Multiply Two Gaussians?

In the Kalman update step, Bayes' rule says:

$$p(x_k \mid z_k) \;\propto\; \underbrace{p(z_k \mid x_k)}_{\text{Gaussian: } \mathcal{N}(x_k;\, z_k,\, \sigma_v^2)} \cdot \underbrace{p(x_k)}_{\text{Gaussian: } \mathcal{N}(x_k;\, \bar{x}_k,\, \bar{P}_k)} .$$

Both the likelihood and the prior are Gaussian.
Their product is also Gaussian – this is the magical property that makes the Kalman Filter work.

Key Question

If I multiply $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ together (as functions of x), what Gaussian do I get?

Note: we multiply the functions (PDFs), not the random variables. The result will be proportional to a Gaussian in x.

The Gaussian PDF as an Exponential

Write the two Gaussians in their exponential form:

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \text{exp}\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right),$$

$$f_2(x) = \frac{1}{\sqrt{2\pi}\sigma_2} \text{exp}\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

Their product:

$$f_1(x) \cdot f_2(x) \propto \text{exp}\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

The exponent is a **quadratic** in x, so the product is proportional to a Gaussian. We just need to find its mean and variance by completing the square.

Completing the Square

Expand the exponent:

$$-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_2)^2}{2\sigma_2^2} = -\frac{1}{2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)x^2 + \left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right)x - (\text{const}).$$

A quadratic $-\frac{1}{2}\alpha x^2 + \beta x + \ldots$ corresponds to $\mathcal{N}(\beta/\alpha,\ 1/\alpha)$.

Reading off α and β:

$$\alpha = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2},$$

$$\beta = \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}.$$

So the product $f_1 \cdot f_2 \propto \mathcal{N}(\mu_*, \sigma_*^2)$ with:

$$\frac{1}{\sigma_*^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \qquad \mu_* = \sigma_*^2\left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right).$$

Result: Product of Two Gaussians

Product of $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ is $\propto \mathcal{N}(\mu_*, \sigma_*^2)$ where:

$$\frac{1}{\sigma_*^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \qquad \text{(precisions add)}$$

$$\mu_* = \frac{\mu_1/\sigma_1^2 + \mu_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \qquad \text{(precision-weighted average of means)}$$

Precision interpretation:

Let $\omega_i = 1/\sigma_i^2$. Then:

$$\omega_* = \omega_1 + \omega_2, \quad \mu_* = \frac{\omega_1\mu_1 + \omega_2\mu_2}{\omega_1 + \omega_2}.$$

Connection to Kalman Filter

In the Kalman update:

$\omega_1 = 1/\bar{P}_k$ (prediction)

$\omega_2 = 1/\sigma_v^2$ (measurement)

$\mu_1 = \bar{x}_k$, $\mu_2 = z_k$

$\Rightarrow$ exactly the Kalman formula!

Why the Posterior Is Always Narrower

After multiplying (fusing) the two Gaussians:

$$\frac{1}{\sigma_*^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \quad \Longrightarrow \quad \sigma_*^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Since $\sigma_1^2 + \sigma_2^2 > \sigma_1^2$ and $\sigma_1^2 + \sigma_2^2 > \sigma_2^2$:

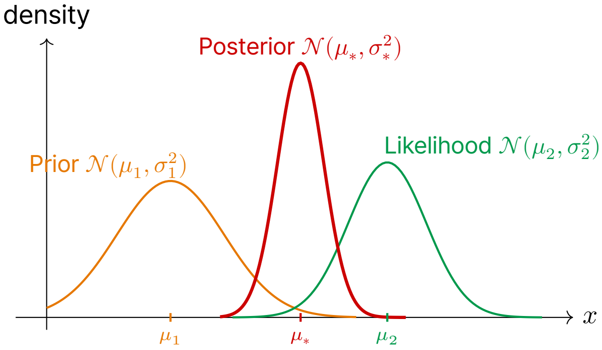
$$\sigma_*^2 < \sigma_1^2 \quad \text{and} \quad \sigma_*^2 < \sigma_2^2.$$

The posterior is always narrower (more certain) than either input.

Every new measurement reduces uncertainty, regardless of how noisy it is.

Analogy: Combining two noisy rulers gives a better estimate than either ruler alone – the errors partially cancel out.

Visual: Product of Two Gaussians



The posterior μ_* lies between μ_1 and μ_2, pulled toward the more precise (narrower) Gaussian. The posterior is narrower than both.

The Full MAP Derivation of the Kalman Update

The Kalman update is the MAP estimate: $\hat{x}_k = \arg\max_x \ p(x_k \mid z_k)$.

Step 1. Bayes: $p(x_k \mid z_k) \propto p(z_k \mid x_k)\, p(x_k)$.

Step 2. Both are Gaussian $\Rightarrow$ product is Gaussian $\mathcal{N}(\mu_*, \sigma_*^2)$.

Step 3. MAP = mode = mean of the posterior Gaussian:

$$\hat{x}_k = \mu_* = \frac{\bar{x}_k/\bar{P}_k + z_k/\sigma_v^2}{1/\bar{P}_k + 1/\sigma_v^2} = \bar{x}_k + \underbrace{\frac{\bar{P}_k}{\bar{P}_k + \sigma_v^2}}_{K_k}(z_k - \bar{x}_k).$$

Step 4. Posterior variance:

$$P_k = \sigma_*^2 = \frac{1}{1/\bar{P}_k + 1/\sigma_v^2} = \frac{\bar{P}_k \sigma_v^2}{\bar{P}_k + \sigma_v^2} = (1 - K_k)\bar{P}_k.$$

The Kalman update is exactly Bayesian fusion of two Gaussians.

The Kalman gain K_k is the fraction of precision coming from the prediction.