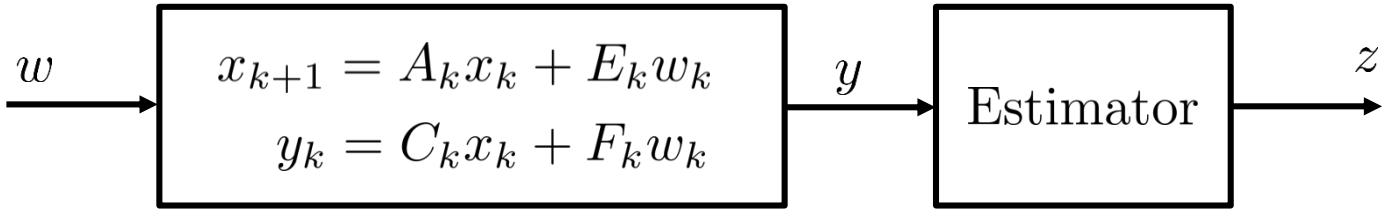
ME C231B/EECS C220C

UC Berkeley

Spring 2018

Estimation Setup

Linear dynamical system

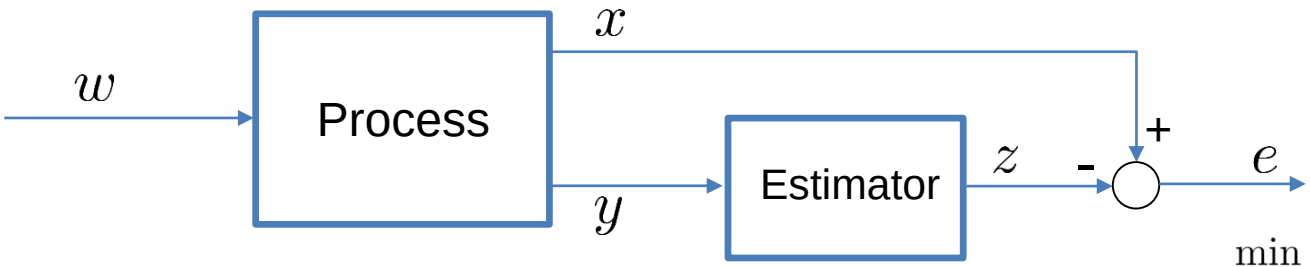


- Discrete-time model
- Internal state, $x_k \in \mathbb{R}^n$, not measured, but to-be-estimated
- Driven by noise, $w_k \in \mathbb{R}^p$
- Measured output, $y_k \in \mathbb{R}^m$, is noisy, linear combinations of x_k
- Known matrices, $\{A_k, E_k, C_k, F_k\}_{k=0}^{\infty}$ defining relationships between (x, y, w)
- Some known (perhaps vague) properties of $\{w_k\}_{k=0}^{\infty}$, and x_0

Build “optimal estimator”

yields z_k, the “best” estimate of x_k, given $y_0, y_1, \ldots, y_{k-1}$

Basic picture of filtering/estimation



$$\min_{L \in \mathbf{L}} \operatorname*{avg}_{\text{scenarios}} \|e(L; \text{scenario})\|^2$$

$$\min_{L \in \mathbf{L}} \max_{\text{scenarios}} \|e(L; \text{scenario})\|^2$$

Estimator is a dynamical system, parametrized by $L \in \mathbf{L}$

Scenarios ("training" data)

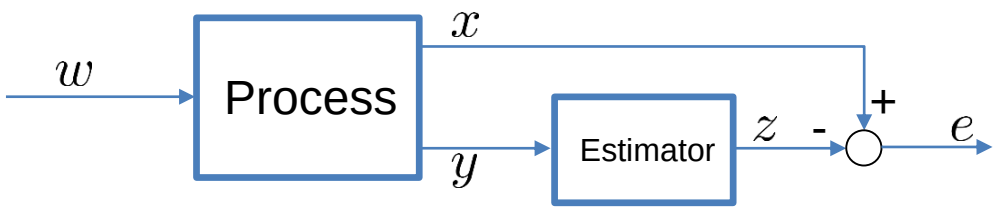
- a collection of representative initial conditions for the process
- a collection of representative noise/disturbance signals over a finite horizon
- a collection of models for the process (representing process uncertainty)

Estimator design is optimization over the parametrization

- minimize some norm of the estimation error
- considering all scenarios

Basic picture of filtering/estimation

Some possible structures for estimator/$\mathbf{L}$



Luenberger observer structure

$$z_{k+1} = Az_k + L(Cz_k - y_k)$$

FIR (finite-impulse response)

$$z_k = L_0 y_k + L_1 y_{k-1} + \cdots + L_N y_{k-N}$$

$$\min_{L \in \mathbf{L}} \ \operatorname*{avg}_{\text{scenarios}} \ \|e(L; \text{scenario})\|^2$$

$$\min_{L \in \mathbf{L}} \ \max_{\text{scenarios}} \ \|e(L; \text{scenario})\|^2$$

FIR (finite-impulse response) with nonlinear basis functions

$$z_k = L_0\phi_0(y_k, \ldots, y_{k-N}) + L_1\phi_1(y_k, \ldots, y_{k-N}) + \cdots + L_m\phi_m(y_k, \ldots, y_{k-N})$$

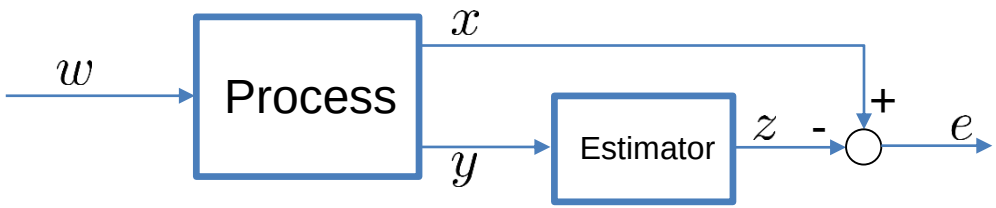
Time-varying FIR (finite-impulse response) with nonlinear basis functions

$$z_k = L_{k,0}\phi_0(y_k, \ldots, y_{k-N}) + L_{k,1}\phi_1(y_k, \ldots, y_{k-N}) + \cdots + L_{k,m}\phi_m(y_k, \ldots, y_{k-N})$$

Neural network (network weights represented by L)

$$z_k = \mathcal{N}(y_k, y_{k-1}, \ldots, y_{k-N}; L)$$

What is the Kalman Filter?



What estimation problem does the *Kalman Filter* solve?

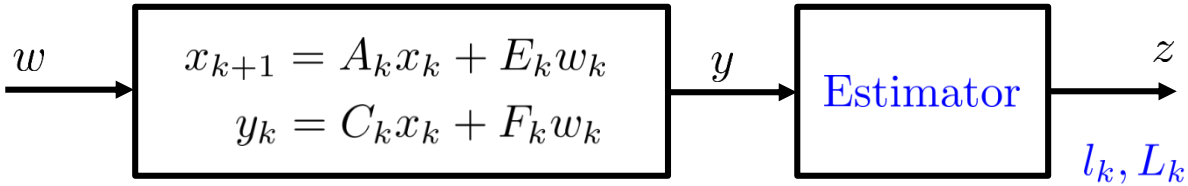
- an infinite family of initial condition and noise/disturbance scenarios, defined in terms of the mathematical language of *probability*
- the estimate at time k is restricted to be a linear combination of measurements $\{y_0, y_1, \ldots, y_{k-1}\}$ (can also include y_k)

$$z_k = l_k + \sum_{m=0}^{k-1} L_{k,m} y_m = l_k + L_{k,0} y_0 + L_{k,1} y_1 + \cdots + L_{k,k-1} y_{k-1}$$

- **optimal estimate**, using the average error (over all scenarios) as the cost
- the process dynamics are linear, **but can be time-varying**. The estimate at time k depends on the process model from 0 to k.

Kalman filter structure and dependencies

$$z_k = l_k + \sum_{m=0}^{k-1} L_{k,m} y_m$$



$l_k, L_{k,:}$ depend on $(A, E, C, G)_{0:k-1}$
and assumptions on $x_0, w_0, \ldots, w_{k-1}$

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_{k-1} \\ z_k \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ L_{1,0} & 0 & 0 & \cdots & 0 \\ L_{2,0} & L_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ L_{k-1,0} & L_{k-1,1} & L_{k-1,2} & \cdots & 0 \\ L_{k,0} & L_{k,1} & L_{k,2} & \cdots & L_{k,k-1} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{k-1} \end{bmatrix} + \begin{bmatrix} l_0 \\ l_1 \\ l_2 \\ \vdots \\ l_{k-1} \\ l_k \end{bmatrix}$$

Goals of the next 3 lectures

Addressing observer-design in the presence of measurement noise and process disturbances is the subject of "optimal filtering."

The usual (but not essential) approach involves making probabilistic models of the noises and disturbances. For any observer, the performance is assessed on the distribution of errors (between actual state and estimated state) that would occur given the distribution of noise/disturbances

This course does not have a probability prerequisite, and the experience with probability ideas is likely wide across the class members

We will introduce the minimal amount of probability theory in order to precisely state and derive the **Kalman Filter**. Depending on your interest and what directions you pursue, you may want to learn more about this topic. In these 3+ lectures, we will give you a good introduction, and you will learn many important facts. But, like most everything, there is always more to learn.

Function approximation (a related mathematical problem)

Data: function $f_i : \mathcal{H} \to \mathbb{R}, i = 0, 1, \ldots, N$

Find "best" linear combination of $f_1, f_2, \ldots, f_N$ to approximate f_0

Error criterion (integrated, square error)

$$e^2(\alpha) := \int_{x \in \mathcal{H}} \left(f_0(x) - \sum_{i=1}^{N} \alpha_i f_i(x) \right)^2 dx$$

fixed $\alpha \in \mathbb{R}^N$
average squared-error over $\mathcal{H}$

$$= \int_{x \in \mathcal{H}} f_0^2(x) - 2 \sum_{i=1}^{N} \alpha_i f_0(x) f_i(x) + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j f_i(x) f_j(x) dx$$

$$= c_0 - 2\alpha^T c + \alpha^T C \alpha$$

$$c_i := \int_{x \in \mathcal{H}} f_0(x) f_i(x) dx$$

$$C_{ij} := \int_{x \in \mathcal{H}} f_i(x) f_j(x) dx$$

$$\Rightarrow \quad \min_{\alpha} \; c_0 - 2\alpha^T c + \alpha^T C \alpha$$

Least Squares gives optimal α
doesn't explicitly depend on f_i, only on C and c

Linear Algebra: background

There is a separate “linearAlgebraPrimer.pdf” handout. We will review this material as it arises.

- eigenvalues and eigenvectors
- Schur decomposition
  - connection to eigenvalues/eigenvectors
- Normal matrices (includes Hermitian and symmetric)
- Definite and semi-definite matrices
- Singular Value Decomposition (SVD)
- Matrix square roots of PSD matrices

Matrix-Valued Completion of Squares

Data: $S = S^T \in \mathbb{R}^{n\times n},\ S \succ 0 \qquad T \in \mathbb{R}^{n\times m} \qquad W = W^T \in \mathbb{R}^{m\times m}$

Define $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times m}$

$$f(K) := KSK^T - KT - T^TK^T + W$$

$f(k) := sk^2 - 2tk + w$

$= (\sqrt{s}k - \frac{t}{\sqrt{s}})^2 + w - \frac{t^2}{s}$

minimum occurs at $k = \frac{t}{s}$

minimum value is $w - \frac{t^2}{s}$

scalar version

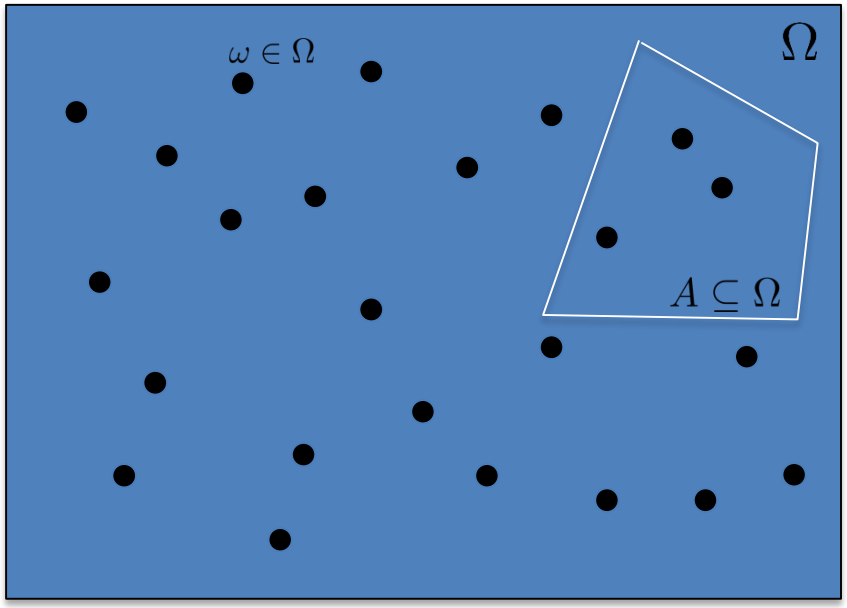
Theorem: there exists $K^* \in \mathbb{R}^{m\times n}$ such that

$$f(K^*) \preceq f(K) \quad \text{for all } K \in \mathbb{R}^{m\times n}$$

In fact, $K^* = T^TS^{-1}$, $f(K^*) = W - T^TS^{-1}T$

Proof: $f(K) = \underbrace{\left(KS^{\frac{1}{2}} - T^TS^{-\frac{1}{2}}\right)}_{=:Q^T}\underbrace{\left(S^{\frac{1}{2}}K^T - S^{-\frac{1}{2}}T\right)}_{=:Q} + W - T^TS^{-1}T$

Sample space



Sample space Ω of all *outcomes* of an "experiment" (or situation). # of outcomes can be

- Finite
- Countable (1-to-1 correspondence with integers)
- Uncountable ($\mathbb{R}$, subset in $\mathbb{R}$, $\mathbb{R}^n$ …)

$\omega \in \Omega$ is an outcome

Subsets of the sample space are called *events*

Experiment: walking from home to BART station, 1 mile away

uncountable outcomes

$\Omega = \{\text{all time-dependent paths from home to BART}\}$

$A = \{\text{all paths that pass by McDonalds}\}$

$A = \{\text{all paths of length less than 1.65 miles}\}$

Experiment: Flipping a coin twice

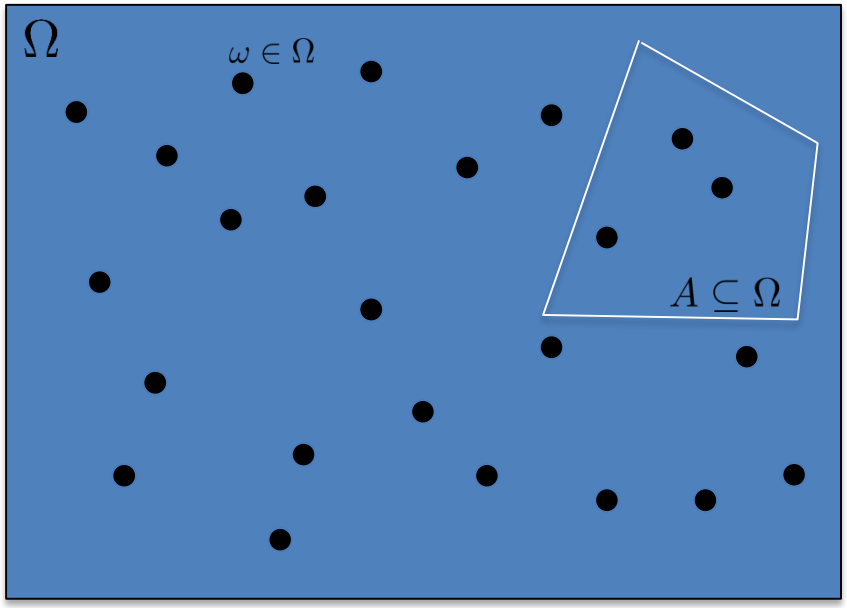
4 outcomes

$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$A = \{\text{outcomes with at least one T}\}$

$A = \{\text{outcomes with one H, one T}\}$

Probability law



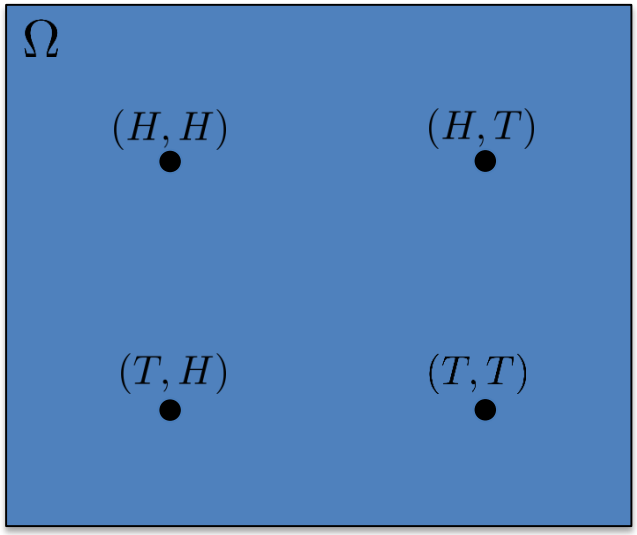
Probability law, $\mathbf{P}$

- assigns to every event A, a nonnegative number $\mathbf{P}(A)$
- $\mathbf{P}(A)$ is the **probability** of A
- $\mathbf{P}$ must satisfy some consistency rules A

1. $\mathbf{P}(A) \geq 0$
2. $\{A_i\}_{i=1}^{N}$, all disjoint, then
$$\mathbf{P}\left(\cup_{i=1}^{N} A_i\right) = \sum_{i=1}^{N} \mathbf{P}(A_i)$$
3. $\mathbf{P}(\Omega) = 1$

The sample space (all outcomes) and Probability law, $\mathbf{P}$, constitute the *probability model* of the experiment. It may require some "art" and "ingenuity" to pick an appropriate model for any specific experiment.

Probability law: simple example



$A = \{\text{outcomes with at least one T}\}$

fair : $\mathbf{P}(A) = 0.25 + 0.25 + 0.25 = 0.75$
biased : $\mathbf{P}(A) = 0.24 + 0.24 + 0.16 = 0.56$

$A = \{\text{outcomes with one H, one T}\}$

fair : $\mathbf{P}(A) = 0.25 + 0.25 = 0.50$
biased : $\mathbf{P}(A) = 0.24 + 0.24 = 0.48$

Example: Flipping a **fair** coin twice

4 outcomes

$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$\mathbf{P}(\{(H,H)\}) = 0.25$
$\mathbf{P}(\{(H,T)\}) = 0.25$
$\mathbf{P}(\{(T,H)\}) = 0.25$
$\mathbf{P}(\{(T,T)\}) = 0.25$

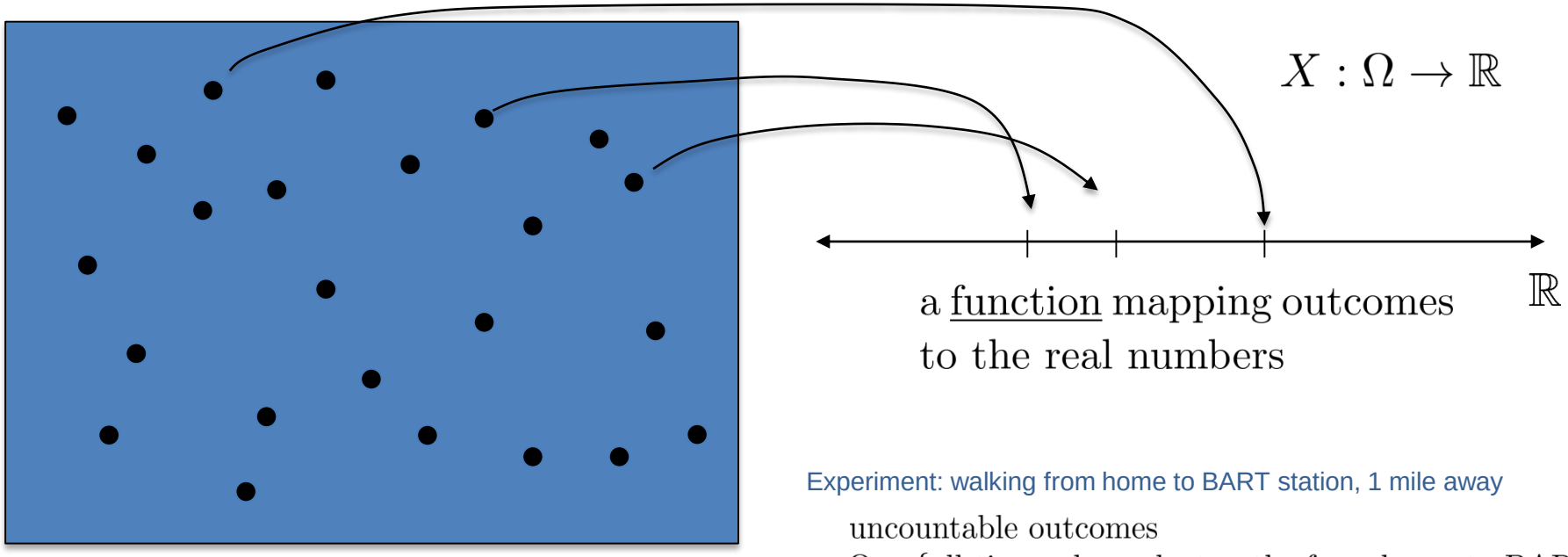
Example: Flipping a **biased** coin twice (h=0.6)

4 outcomes

$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$\mathbf{P}(\{(H,H)\}) = 0.36$
$\mathbf{P}(\{(H,T)\}) = 0.24$
$\mathbf{P}(\{(T,H)\}) = 0.24$
$\mathbf{P}(\{(T,T)\}) = 0.16$

(scalar-valued) Random variables



Experiment: Flipping a coin twice

4 outcomes

$$\Omega = \{\omega_1 = (H,H), \omega_2 = (H,T), \omega_3 = (T,H), \omega_4 = (T,T)\}$$

$$X(\omega_1) := 2,\ X(\omega_2) := 1, X(\omega_3) := 1,\ X(\omega_4) := 0$$

What does this X represent?

Experiment: walking from home to BART station, 1 mile away

uncountable outcomes

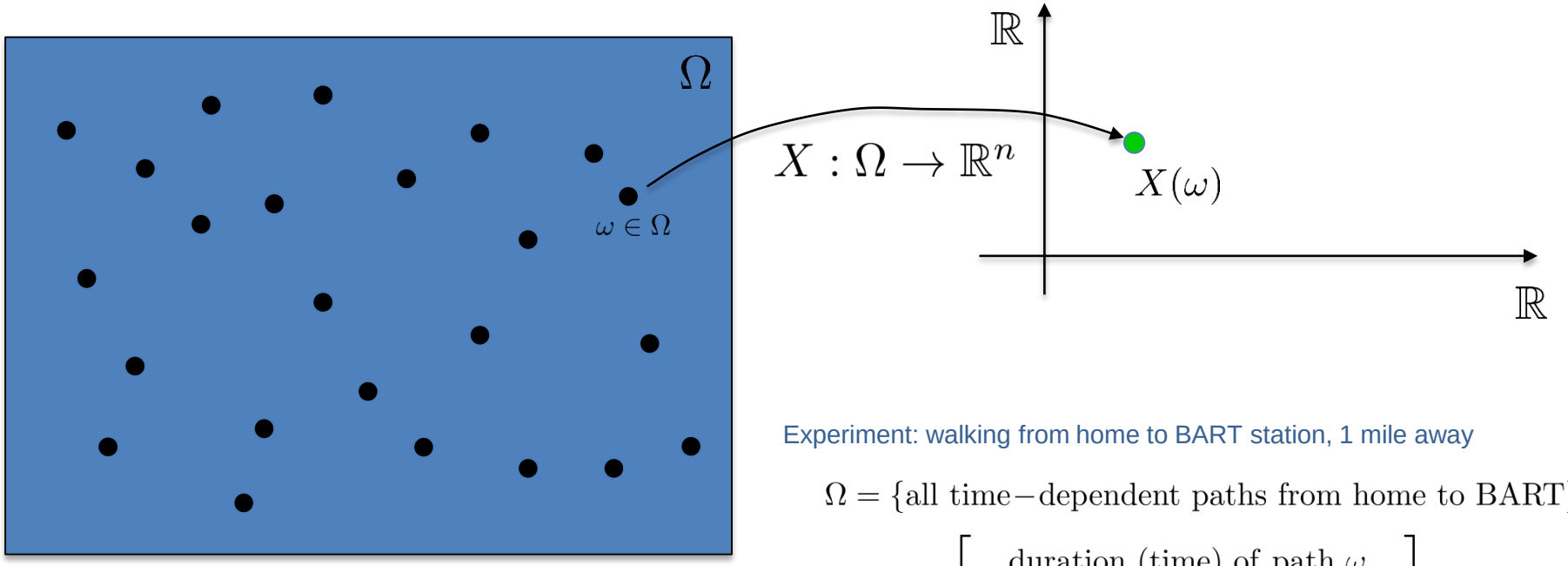
$$\Omega = \{\text{all time}-\text{dependent paths from home to BART}\}$$

$X(\omega) :=$ duration (time) of path ω

$Y(\omega) :=$ length (meters) of path ω

$Z(\omega) :=$ # of mailboxes along path ω

(vector-valued) Random variables



Extension: can generalize to *matrix-valued* random variables

$$X : \Omega \to \mathbb{R}^{n \times m}$$

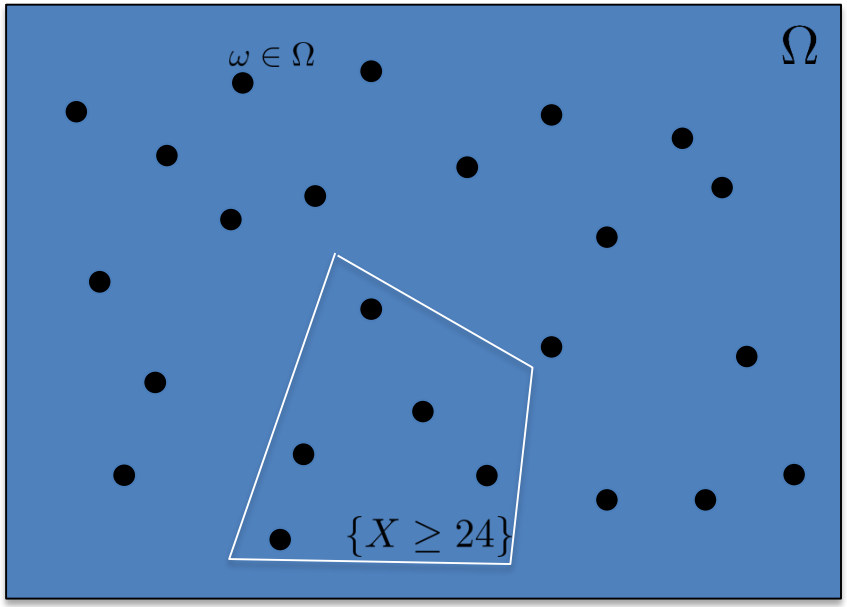
Experiment: walking from home to BART station, 1 mile away

$$\Omega = \{\text{all time}-\text{dependent paths from home to BART}\}$$

$$X(\omega) := \begin{bmatrix} \text{duration (time) of path } \omega \\ \text{length (meters) of path } \omega \\ \#\text{ of mailboxes along path } \omega \end{bmatrix}$$

$$X : \Omega \to \mathbb{R}^3$$

Events defined in terms of Random variables



Experiment: walking from home to BART station, 1 mile away

uncountable outcomes

$\Omega = \{\text{all time}-\text{dependent paths from home to BART}\}$

$X(\omega) :=$ duration (time) of path ω

Random variable $X : \Omega \to \mathbb{R}^n$

If $\mathcal{S} \subseteq \mathbb{R}^n$, then

$$\{\omega \in \Omega : X(\omega) \in \mathcal{S}\}$$

is a subset of Ω (ie., an event)

Shorthand notation

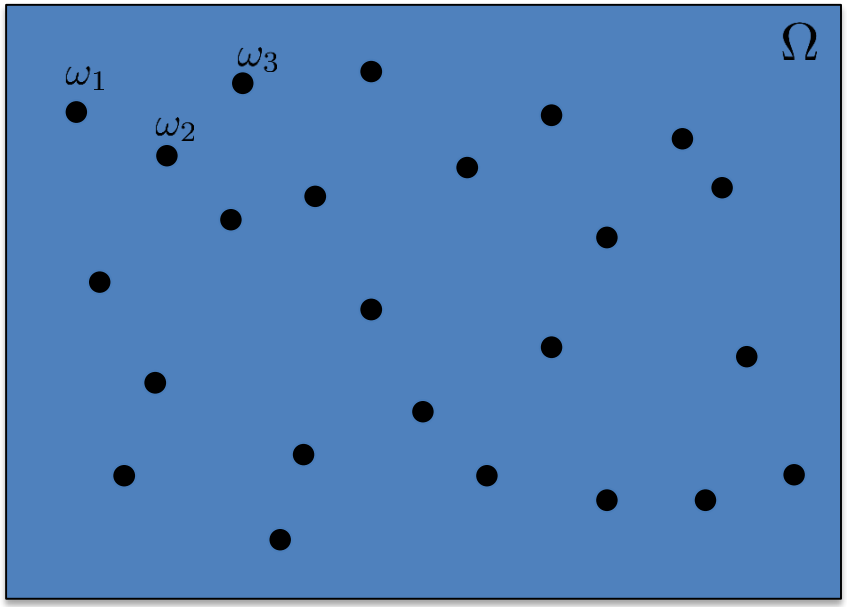
$$\{X \in \mathcal{S}\}$$

Example

$$\{X \geq 24\} \to \{\omega \in \Omega : X(\omega) \geq 24\}$$

all time-dependent paths from Ω with duration ≥ 24 minutes

Our setting: Finite or Countable Sample space



In general: the sums become integrals, and the sample space Ω must be such that you can "integrate over Ω"

The possible outcomes are either

- finite in number, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$; or
- countable (meaning an infinite sequence)
  Label the outcomes as $\omega_1, \omega_2, \ldots$

The probability law dictates the probability of each single-element set, $\{\omega_k\}$, namely

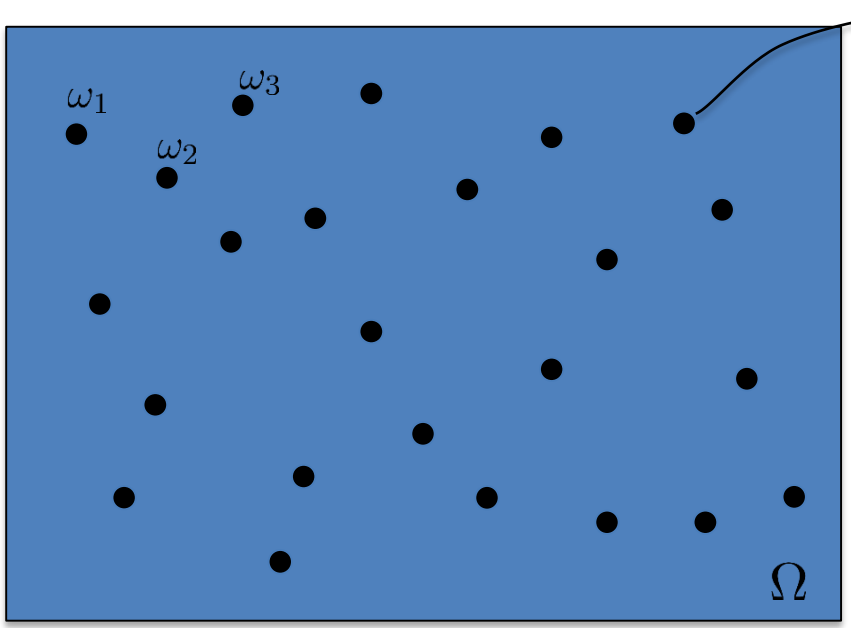
$$\mathbf{P}\left(\{\omega_k\}\right) = p_k \geq 0$$

with $\sum_{k=1}^{N} p_k = 1$ or $\sum_{k=1}^{\infty} p_k = 1$

Write $\sum_k$, to indicate the sum over the entire sample space,

$$\sum_{k=1}^{N} \quad \text{(for finite)}, \qquad \sum_{k=1}^{\infty} \quad \text{(for countable)}$$

Expectation, expected value of random variable



$X : \Omega \to \mathbb{R}$

a function mapping outcomes to the real numbers

$\mathbb{R}$

Expected-value of X

$$\mathbb{E}(X) := \sum_k p_k X(\omega_k)$$

Weighted sum (by the p_k) of all of the values that X takes on across the sample space. The "average value of X"

Experiment: Flipping the biased coin twice

$$\Omega = \{\omega_1 = (H,H), \omega_2 = (H,T), \omega_3 = (T,H), \omega_4 = (T,T)\}$$

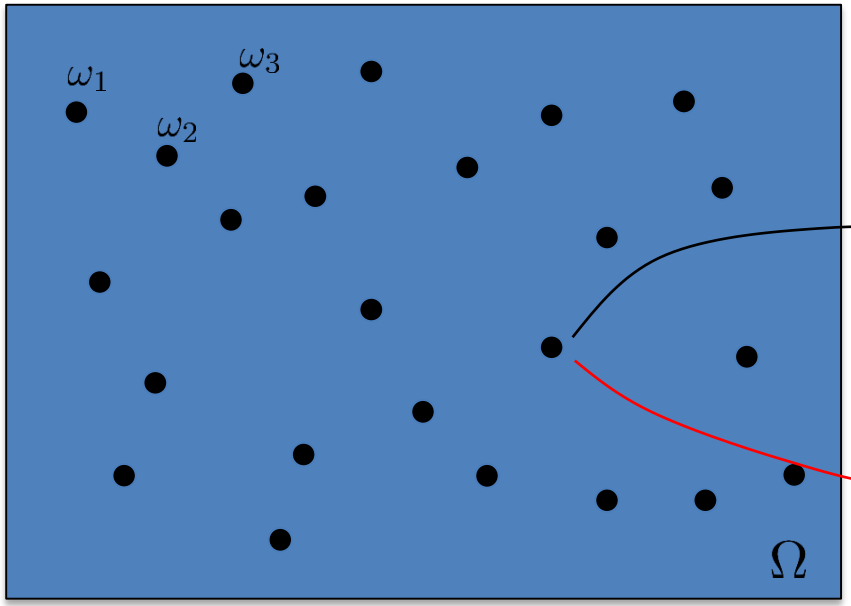
$$\mathbf{P}(\{\omega_1\}) = 0.36 \quad \mathbf{P}(\{\omega_2\}) = 0.24 \quad \mathbf{P}(\{\omega_3\}) = 0.24 \quad \mathbf{P}(\{\omega_4\}) = 0.16$$

$$X(\omega) := (\#\text{ of } H \text{ in } \omega)$$

$$X(\omega_1) = 2, \; X(\omega_2) = X(\omega_3) = 1, \; X(\omega_4) = 0$$

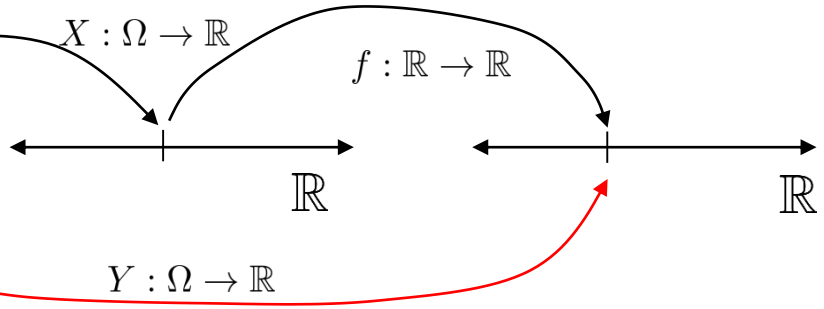
$$\mathbb{E}(X) = 0.36 \cdot 2 + 0.24 \cdot 1 + 0.24 \cdot 1 + 0.16 \cdot 0 = 1.20$$

Functions of a random variable



random variable $X : \Omega \to \mathbb{R}$

function $f : \mathbb{R} \to \mathbb{R}$



Define new random variable
$Y : \Omega \to \mathbb{R}$ by composition

$$Y(\omega) := f(X(\omega))$$

concise notation

$$Y = f(X)$$

obvious generalization to vector-valued

$X : \Omega \to \mathbb{R}^n$

$f : \mathbb{R}^n \to \mathbb{R}^m$

$Y : \Omega \to \mathbb{R}^m, Y(\omega) := f(X(\omega))$

$$\mathbb{E}(Y) = \mathbb{E}(f(X))$$
$$= \textstyle\sum_k p_k Y(\omega_k) \;=\; \sum_k p_k f(X(\omega_k))$$

Expectation operator is linear

(vector-valued) Random variables X and Y over same probability model $(\Omega, \mathbf{P})$. Fixed matrices/vectors A, B, C

$$\begin{aligned}
\mathbb{E}(AX + BY + C) &= \textstyle\sum_j p_j (AX + BY + C)(\omega_j) \\
&= \textstyle\sum_j p_j (AX(\omega_j) + BY(\omega_j) + C) \\
&= A \textstyle\sum_j p_j X(\omega_j) \; + \; B \sum_j p_j Y(\omega_j) \; + \; C \sum_j p_j \\
&= A\mathbb{E}(X) \; + \; B\mathbb{E}(Y) \; + \; C
\end{aligned}$$

Expectation operator is linear (slight generalization)

(matrix-valued) Random variables X and Y over same probability model $(\Omega, \mathbf{P})$. Fixed matrices/vectors A, B, C, D, E of appropriate dimensions

$$\begin{aligned}
\mathbb{E}(AXB + CYD + E) &= \textstyle\sum_j p_j (AXB + CYD + E)(\omega_j) \\
&= \textstyle\sum_j p_j (AX(\omega_j)B + CY(\omega_j)D + E) \\
&= A\left[\textstyle\sum_j p_j X(\omega_j)\right] B \;+\; C\left[\textstyle\sum_j p_j Y(\omega_j)\right] D \;+\; E\textstyle\sum_j p_j \\
&= A\mathbb{E}(X)B \;+\; C\mathbb{E}(Y)D \;+\; E
\end{aligned}$$

mean of X

$$\mu_X := \mathbb{E}(X) = \sum_k p_k X(\omega_k)$$

weighted (by probabilities) average value of X over Ω

variance of X

$f(t) := (t - \mu_X)^2$

$Z := f(X)$

$\Sigma_{XX} = \mathbb{E}(Z)$

$$\Sigma_{XX} := \mathbb{E}\left[(X - \mu_X)^2\right] = \sum_k p_k (X(\omega_k) - \mu_X)^2$$

weighted (by probabilities) average value of the square of the difference of X from its mean, over Ω

$$\sigma_X := \sqrt{\Sigma_{XX}}$$

"standard deviation" (spread)

covariance of X and Y

$f(t_1, t_2) := (t_1 - \mu_X)(t_2 - \mu_Y)$

$Z := f(X, Y)$ $\Sigma_{XY} = \mathbb{E}(Z)$

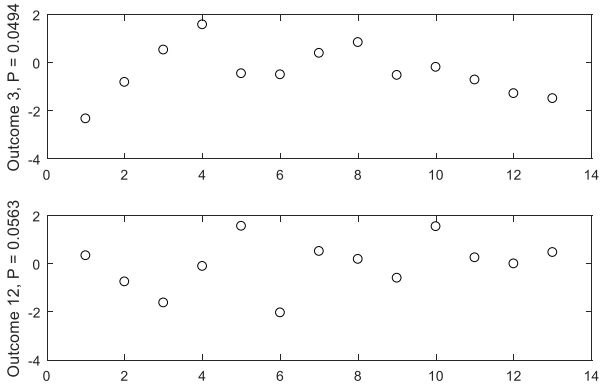
$$\Sigma_{XY} := \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] = \sum_k p_k (X(\omega_k) - \mu_X)(Y(\omega_k) - \mu_Y)$$

Example: finite Sample Space and 13 random variables

Sample space with 20 outcomes, different probabilities

13 random variables, $X_1, X_2, \ldots, X_{13}$

P	0.091	0.088	0.0494	0.000413	0.0384	0.097	0.0428	0.0158	0.0403	0.0336	0.064	0.0563	0.0252	0.0154	0.0269	0.0875	0.0759	0.0498	0.0829	0.0194
X1	9.6e-01	5.8e-03	-2.3e+00	6.7e+00	-5.0e-01	1.1e+00	-2.0e+00	1.9e+00	1.6e-01	7.5e-01	-5.0e-01	3.4e-01	-1.6e+00	6.3e-01	-1.6e+00	2.3e-01	-5.3e-01	-9.9e-02	9.8e-01	-3.3e-01
X2	-9.8e-02	-2.5e-01	-8.2e-01	1.8e+01	5.0e-01	3.5e-01	2.0e+00	5.2e-01	9.6e-01	-8.8e-02	-1.3e+00	-7.5e-01	1.9e+00	-6.4e-01	1.1e+00	2.1e-01	-2.9e-01	-3.4e-01	5.0e-01	-3.9e+00
X3	-1.5e-01	1.0e+00	5.3e-01	-7.3e+00	-2.6e+00	7.7e-01	-1.3e+00	6.4e-01	-1.1e+00	4.7e-03	5.0e-01	-1.6e+00	3.7e-01	-7.2e-01	1.5e+00	4.9e-01	6.5e-01	4.1e-01	-2.9e-01	-2.6e+00
X4	2.2e-01	-1.7e+00	1.6e+00	-4.6e+00	3.9e-01	1.5e+00	-2.2e-01	1.4e+00	-5.9e-01	2.1e+00	-4.8e-01	-1.1e-01	1.0e+00	-2.5e-01	-3.6e-01	5.6e-02	-6.9e-01	-7.5e-02	-1.1e+00	2.2e-01
X5	2.0e-01	4.2e-01	-4.6e-01	-8.9e+00	-5.6e-02	2.6e-02	-3.8e-01	-1.7e+00	-9.1e-01	5.0e-01	5.5e-01	1.6e+00	1.7e+00	-4.3e+00	1.5e+00	1.0e-01	-1.6e+00	-8.0e-01	6.1e-01	-1.4e-01
X6	-9.9e-01	7.7e-01	-5.0e-01	-4.4e-01	1.6e+00	3.7e-01	2.6e-01	-5.9e-01	-1.5e+00	9.1e-01	4.7e-01	-2.0e+00	1.8e+00	-1.1e-01	-2.5e+00	2.0e-01	-2.2e-01	-2.8e-01	8.3e-01	1.5e+00
X7	-9.3e-02	6.5e-02	3.9e-01	8.0e+00	8.7e-01	3.4e-01	-2.0e-01	-3.9e+00	-3.6e-01	9.8e-01	1.6e+00	5.2e-01	-1.4e+00	2.5e+00	-6.9e-01	-4.3e-01	-6.4e-01	4.4e-01	-2.1e-01	-3.6e+00
X8	4.5e-01	2.0e-02	8.4e-01	-1.3e+01	-8.0e-01	6.2e-01	1.4e+00	7.9e-01	-1.1e-01	-3.3e+00	3.3e-01	1.8e-01	-1.3e-02	1.0e+00	-2.3e+00	6.6e-01	-1.3e+00	-3.2e-01	2.3e-01	-7.8e-01
X9	1.2e+00	-2.8e-01	-5.3e-01	-1.9e+01	1.3e+00	-1.2e+00	-1.1e+00	8.9e-01	-9.2e-02	-1.3e-01	-1.0e-01	-6.0e-01	2.4e+00	1.9e+00	2.4e-01	-3.5e-01	-3.6e-01	1.6e+00	4.5e-02	-1.5e+00
X10	-4.4e-01	6.2e-01	-1.9e-01	-6.5e-01	1.1e+00	4.0e-01	-1.4e+00	-1.4e+00	-1.4e-02	-1.0e+00	-1.2e+00	1.5e+00	1.7e+00	-7.9e-01	-1.9e+00	8.3e-01	1.3e+00	-4.9e-02	-1.2e+00	-1.3e+00
X11	1.2e+00	-6.7e-01	-7.2e-01	-1.2e+00	-9.6e-01	5.1e-01	1.0e+00	-2.6e+00	-2.2e+00	-4.5e-01	-3.0e-01	2.5e-01	1.2e+00	1.3e+00	1.5e-01	-9.6e-01	1.4e+00	-9.6e-01	4.4e-01	6.2e-01
X12	3.0e-01	1.1e+00	-1.3e+00	1.0e+01	1.3e-01	8.1e-01	1.0e+00	5.1e-01	-5.9e-01	-7.6e-01	5.5e-02	-3.2e-03	2.5e-01	-1.7e-01	6.7e-01	-9.9e-01	-8.4e-01	1.8e+00	-1.7e+00	1.6e+00
X13	-1.3e-01	-1.8e-01	-1.5e+00	-1.2e+01	8.7e-01	-5.1e-01	1.7e+00	2.0e+00	-2.0e+00	9.6e-01	3.9e-01	4.7e-01	-2.0e+00	1.1e-01	5.2e-01	1.4e+00	4.1e-01	-3.0e-01	-7.7e-01	-1.2e+00



Data constructed so that:

$$\mathbb{E}(X_1) = 0,\ \mathbb{E}(X_2) = 0,\ \ldots,\ \mathbb{E}(X_{13}) = 0$$

$$\mathbb{E}(X_1 \cdot X_2) = 0,\ \ldots,\ \mathbb{E}(X_{12} \cdot X_{13}) = 0$$

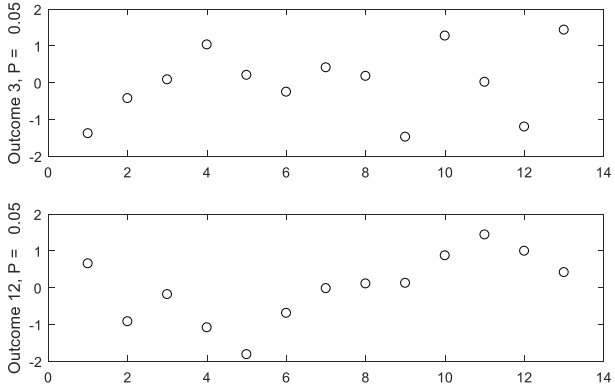
$$\mathbb{E}(X_1 \cdot X_1) = 1,\ \ldots,\ \mathbb{E}(X_{13} \cdot X_{13}) = 1$$

Example: finite Sample Space and 13 random variables

Sample space with 20 outcomes, identical probabilities

13 random variables, $X_1, X_2, \ldots, X_{13}$

P	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
X1	-1.4e+00	-1.0e+00	-1.4e+00	1.2e+00	-1.0e+00	2.0e+00	-2.7e-01	1.5e+00	-1.2e+00	1.0e+00	-4.2e-01	6.5e-01	-7.3e-01	9.1e-01	-6.2e-01	7.7e-02	-6.6e-01	2.4e-01	1.0e+00	1.3e-01
X2	-1.1e+00	-1.5e+00	-4.3e-01	-6.6e-01	-2.6e-01	-1.1e+00	4.9e-01	1.7e+00	7.1e-02	-8.3e-01	-1.2e+00	-9.3e-01	8.5e-01	2.3e-01	2.1e+00	7.1e-01	1.1e+00	1.1e+00	-5.5e-01	2.9e-01
X3	1.0e+00	1.9e-01	7.7e-02	1.7e+00	3.5e-01	-1.1e+00	-1.6e+00	1.7e+00	8.7e-01	7.1e-01	-1.8e-01	-1.9e-01	-3.1e-01	-1.1e-01	6.1e-03	-1.4e+00	-1.3e+00	1.2e+00	-1.7e+00	-1.7e-02
X4	1.9e-01	-2.9e-01	1.0e+00	8.0e-01	-6.1e-01	1.3e+00	-2.5e-01	5.5e-01	1.7e+00	-3.3e-01	-2.4e-01	-1.1e+00	-1.1e+00	3.0e-01	4.8e-01	1.9e+00	-1.7e-01	-2.0e+00	-1.0e+00	-1.1e+00
X5	9.0e-02	-1.4e+00	2.0e-01	-7.1e-02	-2.5e-01	-5.6e-01	7.1e-01	-6.2e-01	1.1e+00	2.3e+00	-2.5e-02	-1.8e+00	1.5e+00	-4.0e-02	-8.2e-01	5.2e-02	-7.7e-01	4.9e-01	1.3e+00	-1.3e+00
X6	-2.7e-01	-1.4e+00	-2.6e-01	1.5e+00	1.7e+00	1.2e+00	5.6e-01	-1.3e+00	-2.4e-03	-4.1e-01	-1.7e-01	-7.0e-01	-2.4e-01	-2.3e+00	3.7e-01	4.5e-01	-5.8e-01	7.0e-01	-3.6e-01	1.4e+00
X7	-1.4e+00	1.3e+00	4.0e-01	-1.2e+00	1.8e+00	8.6e-01	-1.5e+00	5.6e-01	-1.7e-02	-2.0e-01	-2.2e-01	-2.5e-02	2.9e-01	-7.3e-01	8.8e-01	4.9e-01	-1.3e+00	7.4e-01	1.0e+00	-1.7e+00
X8	1.7e-01	-6.8e-01	1.7e-01	-3.0e-01	7.4e-01	1.1e+00	-7.7e-01	-9.5e-01	5.7e-01	1.1e+00	-9.6e-01	1.0e-01	-1.5e+00	3.4e-01	4.9e-01	-1.6e+00	2.6e+00	7.3e-01	-1.9e-01	-1.1e+00
X9	4.9e-01	1.3e-01	-1.5e+00	-8.7e-01	1.5e+00	-4.5e-01	1.1e+00	5.8e-02	6.7e-01	1.1e+00	-1.0e+00	1.2e-01	-1.0e+00	8.7e-01	1.5e+00	-7.4e-01	-1.4e+00	-1.8e+00	4.2e-01	8.7e-01
X10	-1.2e+00	7.2e-01	1.3e+00	1.0e+00	-1.3e+00	-1.5e+00	2.0e-01	4.5e-02	7.4e-01	6.4e-01	-1.0e+00	8.7e-01	-1.3e+00	-1.8e+00	6.3e-01	-8.6e-02	2.3e-01	-2.5e-01	1.7e+00	4.5e-01
X11	-6.3e-01	-1.2e+00	1.1e-02	3.5e-01	4.0e-01	5.9e-02	1.6e+00	4.7e-01	1.8e+00	-2.0e+00	8.3e-01	1.4e+00	-7.5e-02	2.1e-01	-4.8e-01	-1.5e+00	-5.4e-01	1.9e-01	4.8e-01	-1.4e+00
X12	1.8e-01	-1.5e+00	-1.2e+00	-9.9e-01	-4.1e-01	-4.9e-01	-2.2e+00	-2.7e-01	1.4e+00	3.4e-01	2.1e+00	9.9e-01	-3.4e-01	-4.9e-01	5.4e-01	8.9e-01	1.5e-01	-1.2e-01	5.9e-01	7.9e-01
X13	2.9e-01	-7.6e-01	1.4e+00	1.1e-02	-7.9e-01	4.5e-01	-1.2e+00	-1.7e+00	4.5e-01	-1.0e+00	-1.5e+00	4.1e-01	4.8e-01	1.9e+00	6.4e-01	-1.6e-01	-1.4e+00	8.0e-01	6.5e-01	1.0e+00



Data is different than last page, but still has:

$$\mathbb{E}(X_1) = 0, \ \mathbb{E}(X_2) = 0, \ \ldots, \ \mathbb{E}(X_{13}) = 0$$

$$\mathbb{E}(X_1 \cdot X_2) = 0, \ \ldots, \ \mathbb{E}(X_{12} \cdot X_{13}) = 0$$

$$\mathbb{E}(X_1 \cdot X_1) = 1, \ \ldots, \ \mathbb{E}(X_{13} \cdot X_{13}) = 1$$

Two inequalities: Markov and Chebychev

Markov: Suppose $X : \Omega \to \mathbb{R}$ is a random variable, and $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then for any $a > 0$

$$\mathbf{P}\{X \geq a\} \leq \frac{\mathbb{E}(X)}{a}$$

Proof: Define a random variable $I(\omega) := \begin{cases} a & \text{if } X(\omega) \geq a \\ 0 & \text{if } X(\omega) < a \end{cases}$

Check that $\mathbb{E}(I) = a \cdot \mathbf{P}\{X \geq a\}$. Since $a \geq 0$ and $X(\omega) \geq 0$ for all ω, it follows that $I(\omega) \leq X(\omega)$ for all ω

Therefore $a \cdot \mathbf{P}\{X \geq a\} = \mathbb{E}(I) \leq \mathbb{E}(X)$. Divide by the positive-valued a to get result.

Two inequalities: Markov and Chebychev

Chebyshev: Suppose $X : \Omega \to \mathbb{R}$ is a scalar-valued random variable with mean μ_X, variance Σ_{XX} (and standard-deviation σ_X). Then for any $\beta > 0$

$$\mathbf{P}\left\{|X - \mu_X| \geq \beta \cdot \sigma_X\right\} \leq \frac{1}{\beta^2}$$

Proof: Define a random variable $Y := (X - \mu_X)^2$. Apply Markov's inequality to Y, with $a = \beta^2 \sigma_X^2$. This gives

$$(\mathbf{P}\{Y \geq a\} =) \quad \mathbf{P}\{(X - \mu_X)^2 \geq \beta^2 \sigma_X^2\} \leq \frac{\mathbb{E}\,(X - \mu_X)^2}{\beta^2 \sigma_X^2} \quad (= \frac{\mathbb{E}(Y)}{a})$$

$$\mathbf{P}\{|X - \mu_X| \geq \beta \cdot \sigma_X\} \leq \frac{\sigma_X^2}{\beta^2 \sigma_X^2} = \frac{1}{\beta^2}$$

Compare this to what you might already know about Gaussian random variables

Probability: vector-valued Random Variables

$X : \Omega \to \mathbb{R}^n, Y : \Omega \to \mathbb{R}^m$, vector-valued random variables

$$\begin{aligned}\mu_X &:= \mathbb{E}(X) \\ &= \textstyle\sum_k p_k X(\omega_k) \\ &\in \mathbb{R}^n\end{aligned}$$

mean of X

The same as in the scalar case, just account for vector-values

$$\begin{aligned}\Sigma_{XX} &:= \mathbb{E}\left[(X - \mu_X)(X - \mu_X)^T\right] \\ &= \textstyle\sum_k p_k (X(\omega_k) - \mu_X)(X(\omega_k) - \mu_X)^T \\ &\in \mathbb{R}^{n \times n}\end{aligned}$$

variance of X

to simplify notation, write Σ_{XX} as Σ_X

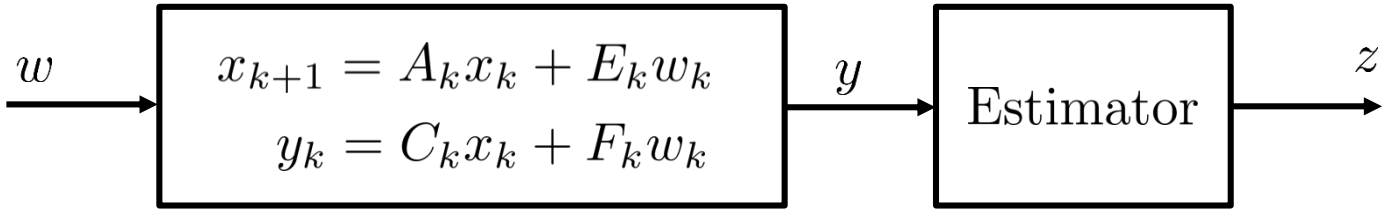
$$\begin{aligned}\Sigma_{XY} &:= \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)^T\right] \\ &= \textstyle\sum_k p_k (X(\omega_k) - \mu_X)(Y(\omega_k) - \mu_Y)^T \\ &\in \mathbb{R}^{n \times m}\end{aligned}$$

covariance of X and Y

sometimes notate with a comma as $\Sigma_{X,Y}$

Kalman Filter Setup

Linear dynamical system



- Discrete-time model
- (random variable) Internal state, $x_k \in \mathbb{R}^n$, not measured, but to-be-estimated
- (random variable) Driven by noise, $w_k \in \mathbb{R}^p$
- (random variable) Measured output, $y_k \in \mathbb{R}^m$, is noisy, linear combinations of x_k
- Known matrices, $\{A_k, E_k, C_k, F_k\}_{k=0}^{\infty}$ defining relationships between (x, y, w)
- Known statistical properties (means, variances, correlations) of $\{w_k\}_{k=0}^{\infty}$, and x_0
  - All other random variables, $y_0, y_1, \ldots,$ and $x_1, x_2, \ldots,$ are a consequence of the dynamic evolution

Build “optimal estimator”: yields “best” estimate of x_k, given $y_0, y_1, \ldots, y_{k-1}$

Properties of Variance

Random variable X, with variance $\Sigma_{XX} := \mathbb{E}(X - \mu_X)(X - \mu_X)^T$

Fact: (i, j)'th entry is covariance of X_i and X_j

Fact: $\Sigma_{XX} = \Sigma_{XX}^T \in \mathbb{R}^{n \times n}$

Fact: $\Sigma_{XX} \succeq 0$

Proof: Let $v \in \mathbb{R}^n$.

$$
\begin{aligned}
v^T \Sigma_{XX} v &= v^T \left(\mathbb{E}(X - \mu_X)(X - \mu_X)^T \right) v \\
&= \mathbb{E} v^T (X - \mu_X)(X - \mu_X)^T v \\
&= \mathbb{E} \left(v^T (X - \mu_X) \right)^2 \\
&= \sum_k p_k \left(v^T (X(\omega_k) - \mu_X) \right)^2 \\
&\geq 0
\end{aligned}
$$

It is generally positive-definite, unless the elements of X are linearly dependent – that is, there is a non-trivial linear combination that has zero variance (ie., does not vary over the sample space)

Properties of Variance (continued)

Random variable X, with variance $\Sigma_{XX} := \mathbb{E}(X - \mu_X)(X - \mu_X)^T$

Fact: Suppose every non-trivial linear combination of the entries of X has nonzero variance. Then $\Sigma_{XX} \succ 0$

Proof: Let $v \in \mathbb{R}^n$, $v \neq 0_n$.

$$\begin{aligned} v^T \Sigma_{XX} v &= v^T \left(\mathbb{E}(X - \mu_X)(X - \mu_X)^T \right) v \\ &= \mathbb{E} v^T (X - \mu_X)(X - \mu_X)^T v \\ &= \mathbb{E} \left(v^T (X - \mu_X) \right)^2 \\ &> 0 \end{aligned}$$

Matrix Ordering: impact for Variance

Vector-valued random variables X, Z, each of dimension n.

Fact: If $\Sigma_{XX} \succeq \Sigma_{ZZ}$, then for every $v \in \mathbb{R}^n$

$$\mathrm{Var}(v^T X) \geq \mathrm{Var}(v^T Z)$$

Remark: every linear combination of the X variables has more variance than the same linear combination of Z variables.

Proof: Let $v \in \mathbb{R}^n$.

$$\begin{aligned}
\mathrm{Var}(v^T X) &= \mathbb{E}\left(v^T(X-\mu_X)\right)^2 = \mathbb{E}v^T(X-\mu_X)(X-\mu_X)^T v \\
&= v^T\left[\mathbb{E}(X-\mu_X)(X-\mu_X)^T\right]v = v^T\Sigma_{XX}v \\
&\geq v^T\Sigma_{ZZ}v = v^T\left[\mathbb{E}(Z-\mu_Z)(Z-\mu_Z)^T\right]v \\
&= \mathbb{E}v^T(Z-\mu_Z)(Z-\mu_Z)^T v = \mathbb{E}\left(v^T(Z-\mu_Z)\right)^2 = \mathrm{Var}(v^T Z)
\end{aligned}$$

Estimation: linear, unbiased, minimum-variance estimator

Random variables *Z* (n-by-1) and *P* (m-by-1), with:

- known means

$$\mathbb{E}(Z) =: m_Z, \qquad \mathbb{E}(P) =: m_P$$

Notation warning: Should means be m or μ?

- known variances

$$\mathbb{E}(Z - m_Z)(Z - m_Z)^T =: \Sigma_Z, \qquad \mathbb{E}(P - m_P)(P - m_P)^T =: \Sigma_P$$

- known correlation.

$$\mathbb{E}(Z - m_Z)(P - m_P)^T =: \Sigma_{Z,P}$$

Find "best" linear approximation of *Z* in terms of *P*? Any matrix *K* and vector *b* defines a linear "estimator" of Z in terms of P,

$$\mathcal{L}(Z|P) := KP + b$$

K, *b* can depend on

$$m_P, m_Z, \Sigma_P, \Sigma_Z, \Sigma_{Z,P}$$

"minimize" some measure of $e := (KP + b) - Z$

require unbiased $\Leftrightarrow$ $\mu_e = 0$: $\Rightarrow$ $m_Z = Km_P + b$

so, once K is chosen, define $b := m_Z - Km_P$

Estimation: linear, unbiased, minimum-variance estimator

Specifically: find matrix K to minimize (in $\preceq$ sense) the variance of e

$$\begin{aligned} e &= (KP + b) - Z \\ &= (KP + m_Z - Km_P) - Z \qquad (\text{recall } b = m_Z - Km_P) \\ &= K(P - m_P) - (Z - m_Z) \end{aligned}$$

$$\Rightarrow \quad \mathbb{E}(ee^T) = K\Sigma_P K^T - K\Sigma_{P,Z} - \Sigma_{P,Z}^T K^T + \Sigma_Z$$

$f(K) := KSK^T - KT - T^T K^T + W$
$K^* := T^T S^{-1} \Rightarrow$
$W - T^T S^{-1} T = f(K^*) \preceq f(K)$

Use "matrix completion of squares" result, assuming $\Sigma_P \succ 0$

best value is $K^* = \Sigma_{Z,P}\Sigma_P^{-1}$

optimal estimator is: $\mathcal{L}(Z|P) = \Sigma_{Z,P}\Sigma_P^{-1}(P - m_P) + m_Z$

minimum error variance: $\Sigma_{Z-\mathcal{L}(Z|P)} = \Sigma_Z - \Sigma_{Z,P}\Sigma_P^{-1}\Sigma_{Z,P}^T$

Summary: linear minimum-variance estimator (fact #1)

Random variables Z (n-by-1) and P (m-by-1), with:

- known means $\mathbb{E}(Z) =: m_Z, \quad \mathbb{E}(P) =: m_P$
- known variances $\mathbb{E}(Z - m_Z)(Z - m_Z)^T =: \Sigma_Z, \quad \mathbb{E}(P - m_P)(P - m_P)^T =: \Sigma_P$
- known correlation $\mathbb{E}(Z - m_Z)(P - m_P)^T =: \Sigma_{Z,P}$

Define $\mathcal{L}(Z|P) := \Sigma_{Z,P}\Sigma_P^{-1}(P - m_P) + m_Z$

$\mathcal{L}(Z|P)$ is an affine function of the random variable P (and hence is a random variable)

$\mu_{\mathcal{L}(Z|P)} = m_Z$ (ie., the affine function is "unbiased")

$\Sigma_{Z-\mathcal{L}(Z|P)} = \Sigma_Z - \Sigma_{Z,P}\Sigma_P^{-1}\Sigma_{Z,P}^T$

Any other affine function $\mathcal{A}(P)$ with $\mu_{\mathcal{A}(P)} = m_Z$, has error variance $\Sigma_{Z-\mathcal{A}(P)} \succeq \Sigma_{Z-\mathcal{L}(Z|P)}$

The estimate is an affine function of P

- For any particular $\omega \in \Omega$, this estimator results in an estimation error, $e(\omega) := Z(\omega) - \mathcal{L}(Z|P(\omega))$. For some ω, the error might be very small; on other ω, the error might be large.
- The average of ee^T, taken over all $\omega \in \Omega$, and weighted by their probabilities represents a matrix-valued cost (ie., the variance of the estimation error).
- Over all unbiased affine estimators, the estimator derived here yields the minimum **average** error.

linear minimum-variance estimator: Linearity

Random variables Z_1, Z_2, P. Matrices A_1, A_2

$$\mathcal{L}(A_1 Z_1 + A_2 Z_2 | P) = A_1 \mathcal{L}(Z_1|P) + A_2 \mathcal{L}(Z_2|P)$$

$$Z := A_1 Z_1 + A_2 Z_2 \quad \Rightarrow \quad \mu_Z := A_1 \mu_{Z_1} + A_2 \mu_{Z_2}$$

$$\begin{aligned}
\Sigma_{Z,P} &= \mathbb{E}(Z - \mu_Z)(P - \mu_P)^T \\
&= \mathbb{E}(A_1 Z_1 + A_2 Z_2 - (A_1 \mu_{Z_1} + A_2 \mu_{Z_2}))(P - \mu_P)^T \\
&= \mathbb{E}(A_1 (Z_1 - \mu_{Z_1}))(P - \mu_P)^T + \mathbb{E}(A_2 (Z_2 - \mu_{Z_2}))(P - \mu_P)^T \\
&= A_1 \mathbb{E}(Z_1 - \mu_{Z_1})(P - \mu_P)^T + A_2 \mathbb{E}(Z_2 - \mu_{Z_2})(P - \mu_P)^T \\
&= A_1 \Sigma_{Z_1,P} + A_2 \Sigma_{Z_2,P}
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}(Z|P) &= \Sigma_{ZP} \Sigma_P^{-1} (P - m_P) + m_Z \\
&= (A_1 \Sigma_{Z_1,P} + A_2 \Sigma_{Z_2,P}) \Sigma_P^{-1} (P - m_P) + (A_1 \mu_{Z_1} + A_2 \mu_{Z_2}) \\
&= A_1 \mathcal{L}(Z_1|P) + A_2 \mathcal{L}(Z_2|P)
\end{aligned}$$

linear minimum-variance estimator: adding uncorrelated RV

Random variables X, W, P, with $\mu_W = 0, \Sigma_{W,P} = 0$. Matrix C

$$\mathcal{L}(CX + W|P) = C\mathcal{L}(X|P)$$

$$Z := CX + W \qquad \Rightarrow \quad \mu_Z := C\mu_X$$

$$\begin{aligned}\Sigma_{Z,P} &= \mathbb{E}(C(X - \mu_X) + W)(P - \mu_P)^T \\ &= C\Sigma_{X,P}\end{aligned}$$

$$\begin{aligned}\mathcal{L}(Z|P) &= \Sigma_{ZP}\Sigma_P^{-1}(P - m_P) + m_Z \\ &= C\Sigma_{X,P}\Sigma_P^{-1}(P - m_P) + C\mu_X \\ &= C\left(\Sigma_{X,P}\Sigma_P^{-1}(P - m_P) + \mu_X\right) \\ &= C\mathcal{L}(X|P)\end{aligned}$$