CS 540 Introduction to Artificial Intelligence

PCA, Statistics & Math Review

University of Wisconsin-Madison

Fall ‘22

Announcements

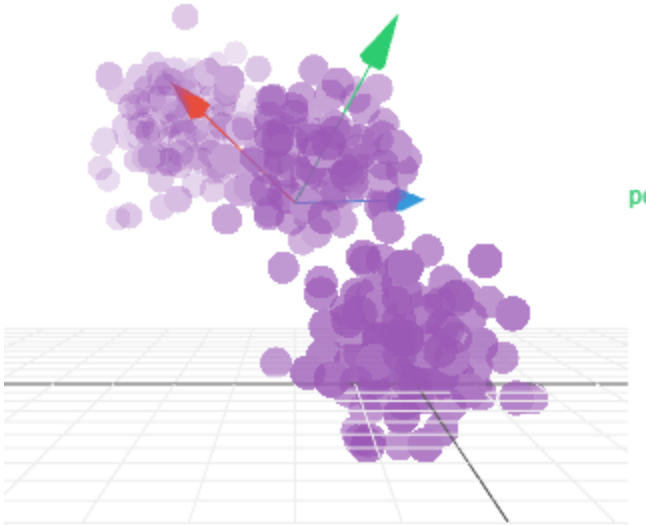
- **Homeworks**:
  - HW1, HW2 due Thursday. Midterm dates coming
- Class roadmap:

Thursday, Sept. 15	Linear Algebra and PCA
Tuesday, Sept. 20	PCA, Stats, Math Review
Thursday, Sept. 22	Introduction to Logic
Tuesday, Sept. 27	Natural Language Processing
Tuesday, Sept. 29	Machine Learning: Introduction

Fundamentals

PCA Setup

- **Inputs**
  - Data: $x_1, x_2, \ldots, x_n, \; x_i \in \mathbb{R}^d$
  - Can arrange into $X \in \mathbb{R}^{n \times d}$
  - **Centered**! $\frac{1}{n} \sum_{i=1}^{n} x_i = 0$
- **Outputs**
  - Principal components $v_1, v_2, \ldots, v_r \in \mathbb{R}^d$
  - Orthogonal!



Victor Powell

PCA Goals

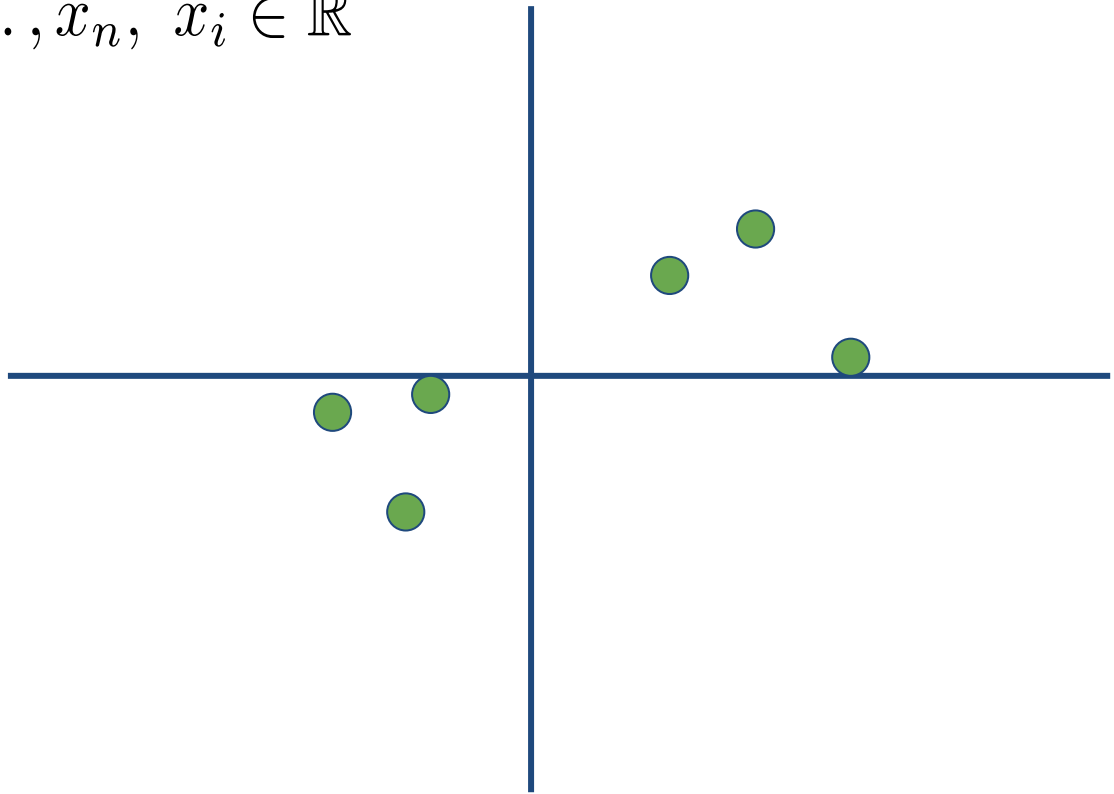
- Want directions/components (unit vectors) so that
  - Projecting data maximizes variance
  - What's projection?

$$\sum_{i=1}^{n} \langle x_i, v \rangle^2 = \|Xv\|^2$$

Let's look at an example!

Projection: An Example

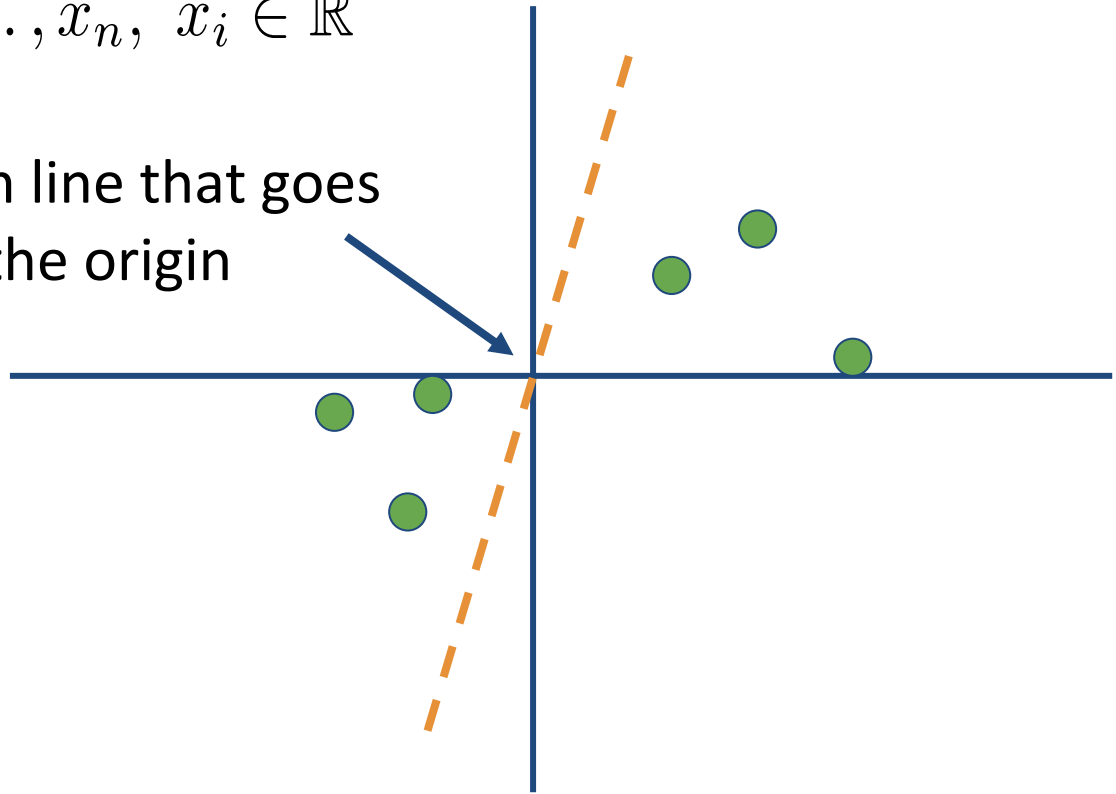
$x_1, x_2, \ldots, x_n, \; x_i \in \mathbb{R}^2$



Projection: An Example

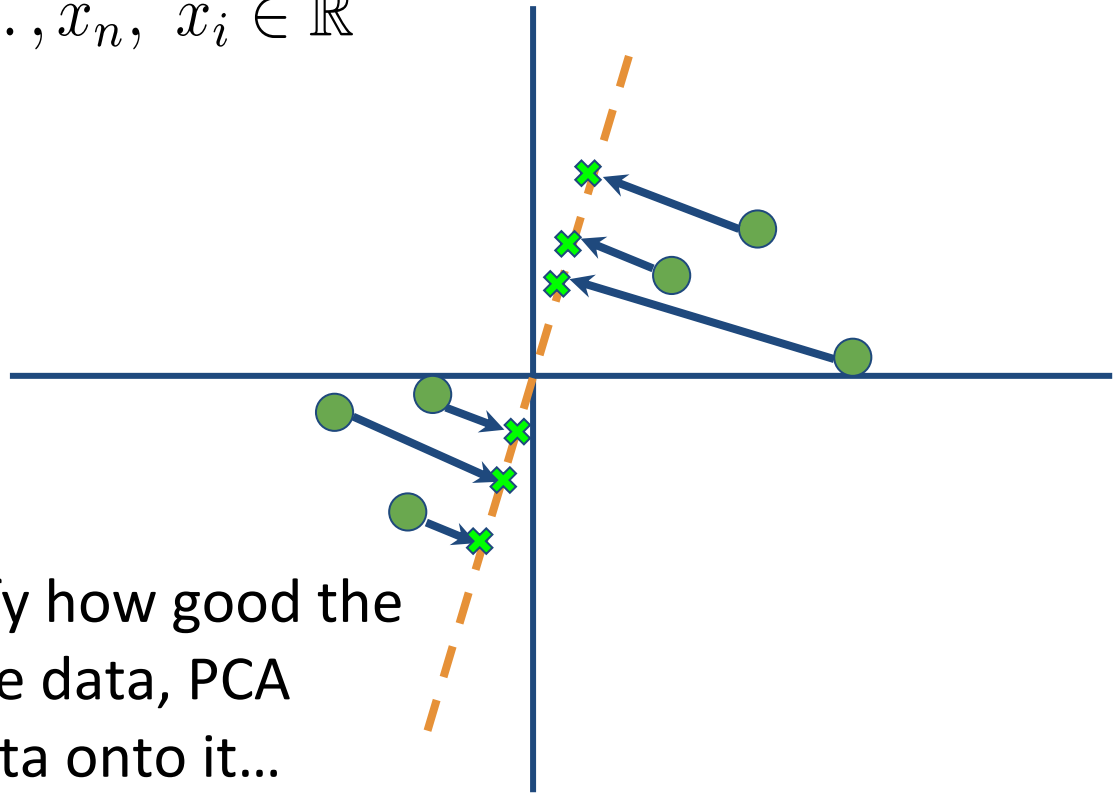
$x_1, x_2, \ldots, x_n, \; x_i \in \mathbb{R}^2$

A random line that goes through the origin



Projection: An Example

$x_1, x_2, \ldots, x_n, \; x_i \in \mathbb{R}^2$

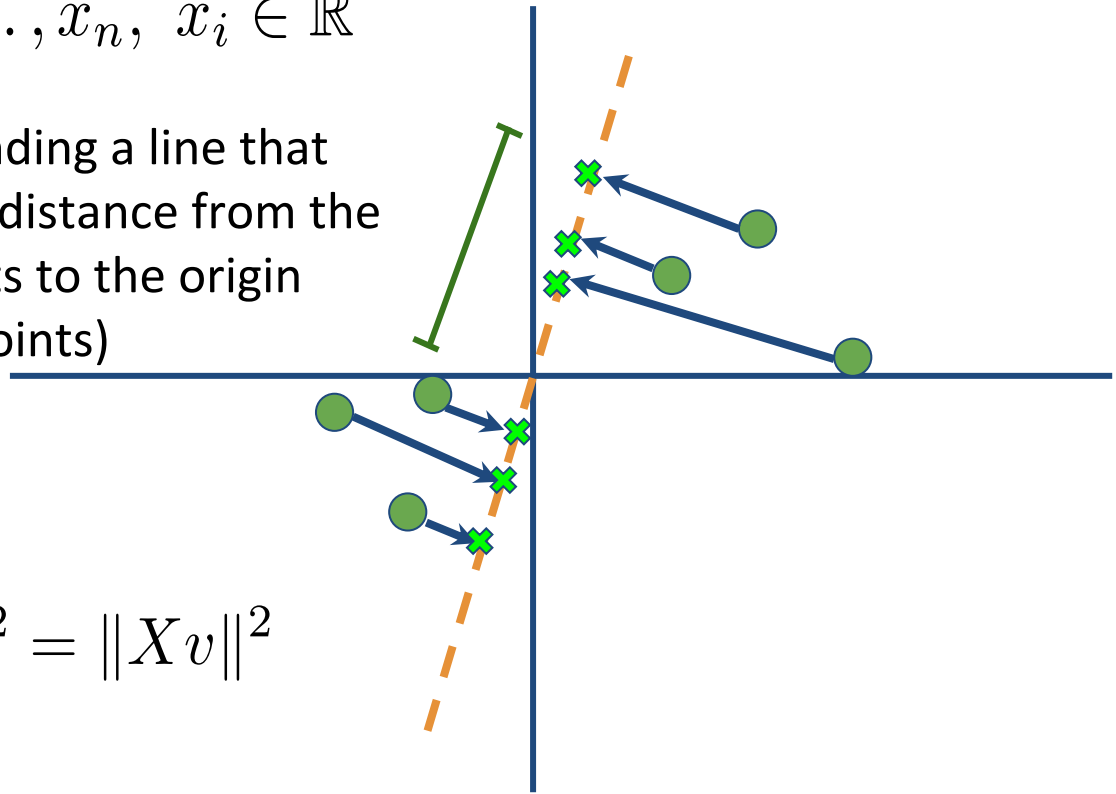


To quantify how good the line fits the data, PCA project data onto it…

Projection: An Example

$$x_1, x_2, \ldots, x_n, \; x_i \in \mathbb{R}^2$$

Goal of PCA: finding a line that **maximizes** the distance from the projected points to the origin (sum over all points)



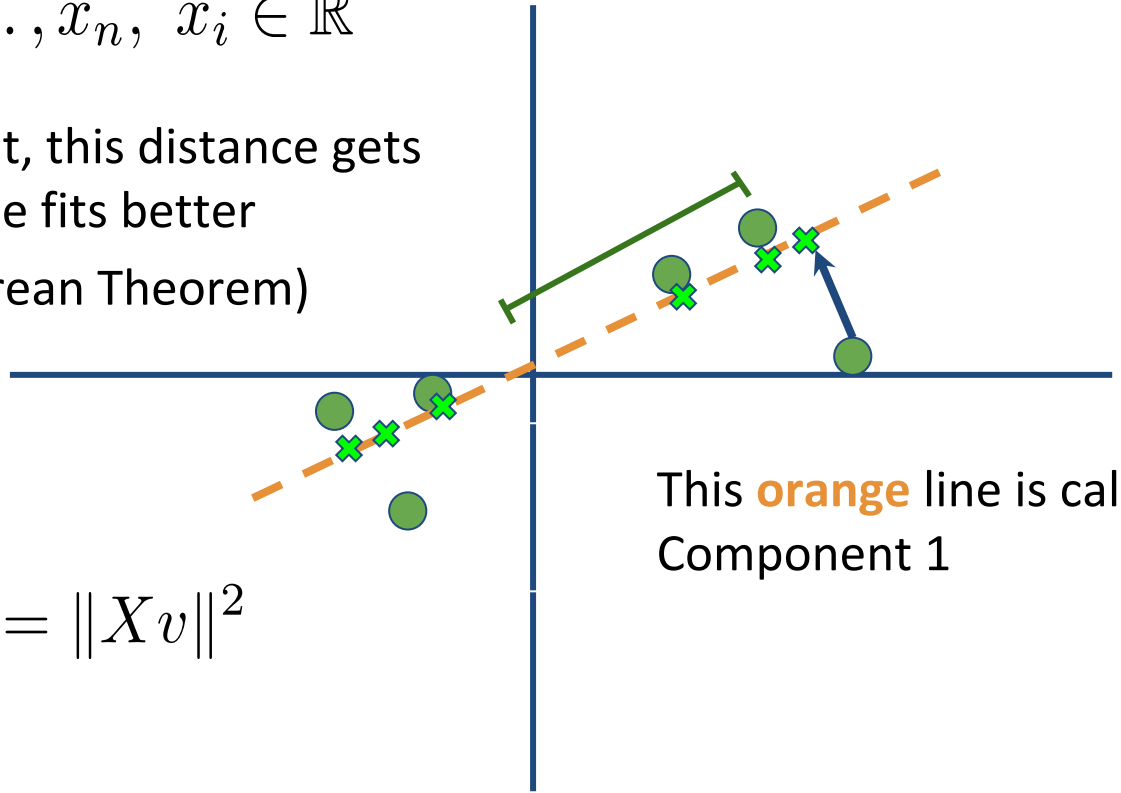
$$\sum_{i=1}^{n} \langle x_i, v \rangle^2 = \|Xv\|^2$$

Projection: An Example

$$x_1, x_2, \ldots, x_n, \; x_i \in \mathbb{R}^2$$

For a fixed point, this distance gets larger as the line fits better

(why? Pythagorean Theorem)



This **orange** line is called Principal Component 1

$$\sum_{i=1}^{n} \langle x_i, v \rangle^2 = \|Xv\|^2$$

PCA First Step

- First component,

$$v_1 = \arg\max_{\|v\|=1} \sum_{i=1}^{n} \langle v, x_i \rangle^2$$

- Same as getting

$$v_1 = \arg\max_{\|v\|=1} \|Xv\|^2$$

PCA Goals

- Want directions/components (unit vectors) so that
  - Projecting data maximizes variance

$$\sum_{i=1}^{n} \langle x_i, v \rangle = \|Xv\|^2$$

- Do this **recursively**
  - Get orthogonal directions

$$v_1, v_2, \ldots, v_r \in \mathbb{R}^d$$

PCA Recursion

- Once we have *k-1* components, next?

$$\hat{X}_k = X - \sum_{i=1}^{k-1} X v_i v_i^T$$

Deflation

- Then do the same thing

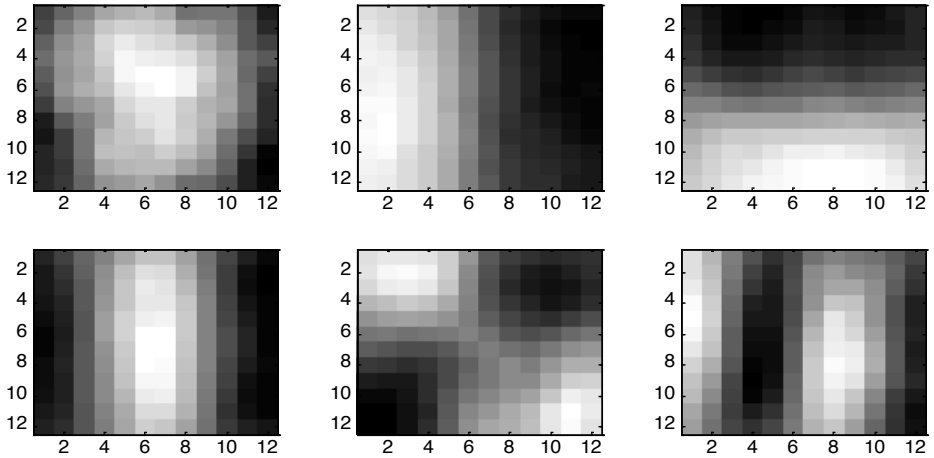
$$v_k = \arg \max_{\|v\|=1} \|\hat{X}_k w\|^2$$

Application: Image Compression

- Start with image; divide into 12x12 patches
  - I.E., 144-D vector
  - **Original image:**

Application: Image Compression

- 6 most important components (as an image)



Application: Image Compression

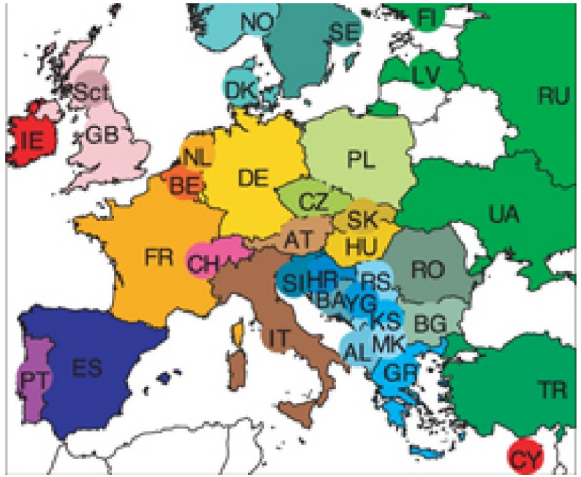
- Project to 6D,

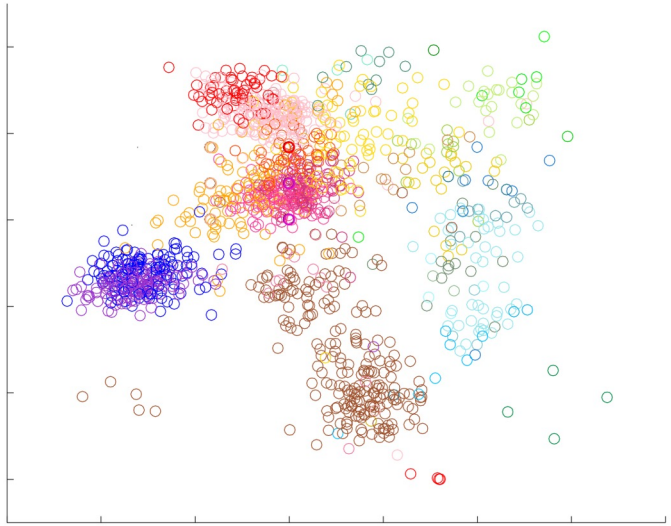
Compressed

Original

Application: Exploratory Data Analysis

- [**Novembre et al. ’08**]: Take top two singular vectors of people x SNP matrix (POPRES)





“Genes Mirror Geography in Europe”

Readings

- Vast literature on linear algebra.
- Local class: **Math 341**.

- **Suggested reading:**
  - Lecture notes on PCA by Roughgarden and Valiant
    https://web.stanford.edu/class/cs168/l/l7.pdf
  - 760 notes by Zhu
    https://pages.cs.wisc.edu/~jerryzhu/cs760/PCA.pdf

Break & Quiz

Q 1.1: Are these statements true or false?
(A) The first principal component is found by minimizing the variation of the projected points.
(B) The dimension of original data representation is always higher than the dimension of transformed representation of PCA.

A. True, True
B. True, False
C. False, True
D. False, False

Break & Quiz

Q 1.1: Are these statements true or false?
(A) The first principal component is found by minimizing the variation of the projected points.
(B) The dimension of original data representation is always higher than the dimension of transformed representation of PCA.

A. True, True
B. True, False
C. False, True
D. False, False

Review: Bayesian Inference

- Conditional Prob. & Bayes:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- *H*: some class we'd like to infer from evidence
  - Need to plug in prior, likelihood, etc.
  - How to estimate?

Samples and Estimation

- Usually, we don't know the distribution (P)
  - Instead, we see a bunch of samples

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability *P(H)*
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$

Samples and Estimation

- Typical statistics problem: **estimate parameters** from samples
  - Estimate probability *P(H)*
  - Estimate the mean $E[X]$
  - Estimate parameters $P_\theta(X)$
- Example: Bernoulli with parameter *p*
  - Mean $E[X]$ is p

Examples: Sample Mean

- Bernoulli with parameter *p*
- See samples $x_1, x_2, \ldots, x_n$
  - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  - No different from counting heads

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $E[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $E[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by [0,1,1,2,2,0,1,2]. Empirically estimate $E[X^2]$

A. 9/8
B. 15/8
C. 1.5
D. There aren't enough samples to estimate $E[X^2]$

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating *P(X)* for some random variable *X* that takes on 100 values. You see 50 samples. How many of your *P(X=a)* estimates might be 0?

A. None.
B. Between 5 and 50, exclusive.
C. Between 50 and 100, inclusive.
D. Between 50 and 99, inclusive.

Estimating Multinomial Parameters

- *k*-sized die (special case: *k*=2 coin)
- Face *i* has probability p_i, for *i=1,...,k*
- In *n* rolls, we observe face *i* showing up n_i times

$$\sum_{i=1}^{k} n_i = n$$

- Estimate ($p_{1,\ldots,}\ p_k$) from this data ($n_{1,\ldots,}\ n_k$)

Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters $(\widehat{p_1}, \ldots, \widehat{p_k})$

$$\widehat{p_i} = \frac{n_i}{n}$$

- "frequency estimate"

Regularized Estimate

- Equivalent to a specific Maximum A Posterori (MAP) estimate, or smoothing
- Hyperparameter $\epsilon > 0$

$$\hat{p}_i = \frac{n_i + \epsilon}{n + k\epsilon}$$

- Avoids zero when *n* is small
- Biased, but has smaller variance

Estimating 1D Gaussian Parameters

- Gaussian distribution $N(\mu, \sigma^2)$
- Observe *n* data points from this distribution

$$x_1, \dots, x_n$$

- Estimate μ, σ^2 from this data

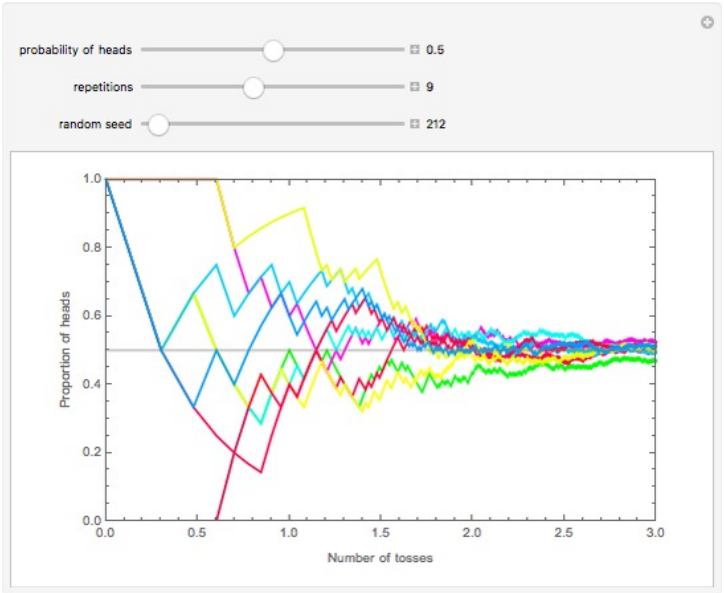
Estimating 1D Gaussian Parameters

- Mean estimate $\hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$
- Variance estimates
  - Unbiased $s^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$
  - MLE $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$

Estimation Theory

- How do we know that the sample mean is a good estimate of the true mean?
  - Law of large numbers
  - Central limit theorems
  - Concentration inequalities

$$P(|\mathbb{E}[X] - \hat{\mathbb{E}}[X]| \geq t) \leq \exp(-2nt^2)$$



Wolfram Demo