CS 3630!

Lecture 5:
A Trash Sorting Robot: Perception, Planning, and Learning

Stay Informed About Robotics@GT

Sign up to the IRIM students e-mail list by visiting:

https://lists.gatech.edu/sympa/subscribe/gtrobotics-students

Quiz and Questionnaire

- First quiz will be released on Thursday, Sept 8, after class:
  - Quiz opens in Canvas at 5:00.
  - Quiz closes Saturday midnight.

➢ All details for this will be announced via Piazza. If you don’t monitor Piazza for course announcements, now is the time to start doing so.

Lecture 4 Recap

Sensing

For our trash sorting robot, we'll consider three sensors:

- Conductivity: A binary sensor that outputs *True* or *False*, based on measurement of electrical conductivity.
- Camera w/detection algorithms: This sensor outputs *bottle*, *cardboard*, or *paper*, based on a detection algorithm (note: it cannot detect scrap metal or cans).
- Scale: Outputs a continuous value that denotes the measured *weight in kg* of the object.

These three kinds of measurements are each treated using distinct probabilistic models.

Conditional Probability

- This conditional probability is defined in terms of the joint probability of $x\ and\ y$:

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

Assuming $P(y) \neq 0$

- *We can rewrite this expression as:*

$$P(x, y) = P(x \mid y)\, P(y)$$

- If X and Y are independent, then

$$P(x, y) = P(x)P(y)$$

Continuous random variables

- Recall the cumulative distribution function (CDF):

$$F_X(\alpha) = P(X \leq \alpha)$$

- If F_X is continuous everywhere, then X is a ***continuous random variable***.

- If X is a continuous random variable with CDF $F_X(\alpha)$, then the ***probability density function (pdf)*** for X is given by

$$f_X(x) = \frac{d}{dx} F_X(x)$$

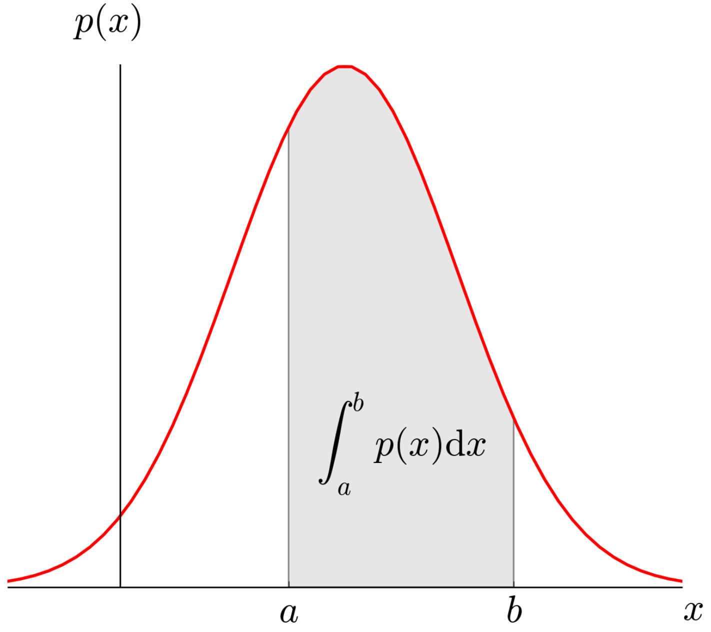
If we think of $F_X(\alpha)$ as probability mass for event $X \leq \alpha$, we can think of the derivative of mass as density.

Computing probabilities

For continuous random variables:

$$P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f_X(u) du$$

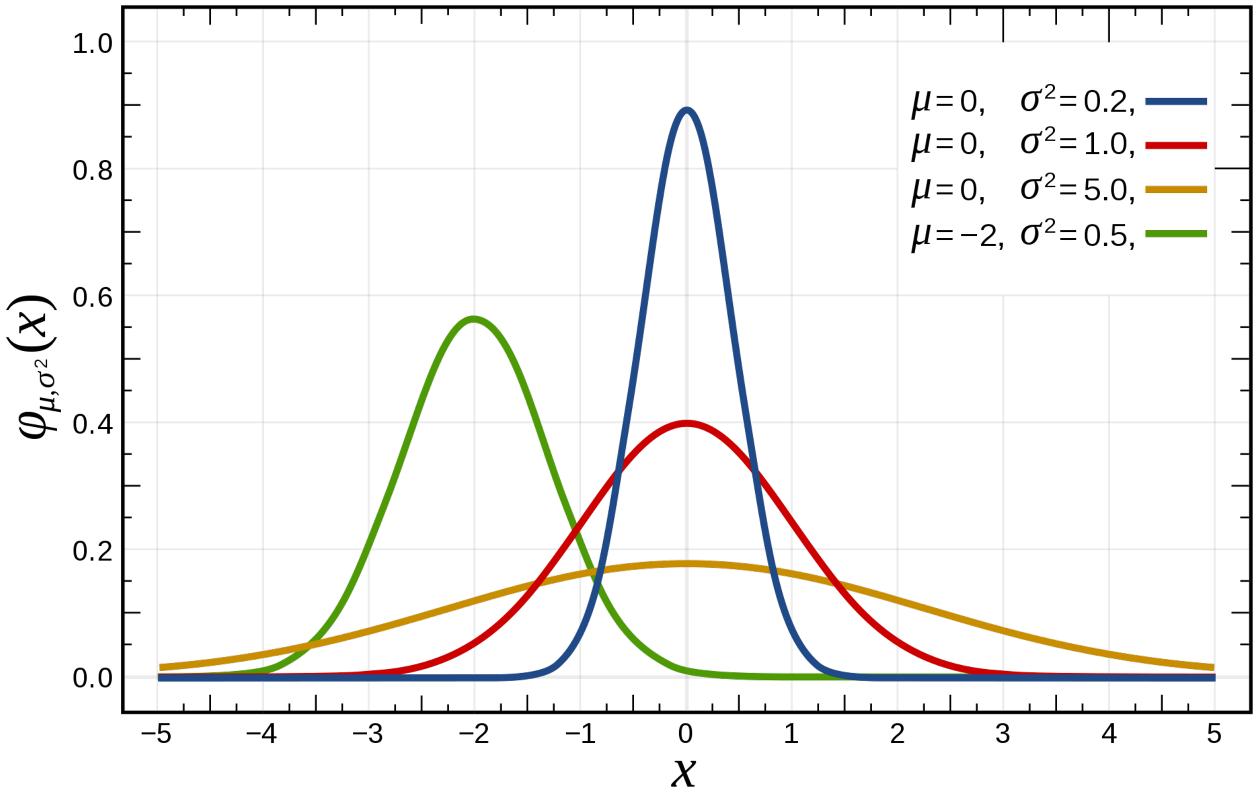
The probability that $\alpha \leq X \leq \beta$ is equal to the area under the pdf f_X between α and β.



The Gaussian distribution

- The Gaussian has two defining parameters.
- The mean, μ
  - Defines the "location" of the pdf.
  - The pdf is symmetric about the mean.
- The variance, σ^2
  - Defines the "spread" of the pdf.
  - Standard deviation is σ.
- The defining equation is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Conditional distributions

- Instead of thinking about five individual pdfs for the different objects, we can think of weight as a random variable characterized by conditional probability distributions:

Category	**Mean μ**	**Std dev σ**
Cardboard	20	10
Paper	5	5
Can	15	5
Scrap metal	150	100
Bottle	300	200

Category (C)	$f_{W\|C}(W\|C)$
Cardboard	$N(20, \sigma = 10)$
Paper	$N(5, \sigma = 5)$
Can	$N(15, \sigma = 5)$
Scrap metal	$N(150, \sigma = 100)$
Bottle	$N(300, \sigma = 200)$

$$f_{X|C}(x|C = Scrap\ Metal) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(x-150)^2}{2\ 100^2}}$$

Perception

Perception is the process of inferring the state of the world (and possibly of the robot itself) using sensor measurements and other contextual information.

In this chapter, we consider two approaches to perception that use conditional probability distributions:

- Maximum Likelihood Estimation
- MAP Estimation

We will also see how to combine measurements from multiple sensors (sometimes called sensor fusion) in a probabilistic framework.

Sensing vs perception

- Sensor models are *forward* models.
  - Given a description of the world and a model of the sensor,
  - ➢Determine the conditional probability

$$P(Observation\ O \mid State\ S)$$

- Perception is concerned with the *inverse* problem.
  - Given a set of observations and (possibly extra contextual information),
  - ➢Infer the probability map associated to the world state

$$P(State\ S \mid Observation\ O, Context)$$

  - Context could include previous sensor readings, knowledge about the robot's actions, etc.

Bayes theorem

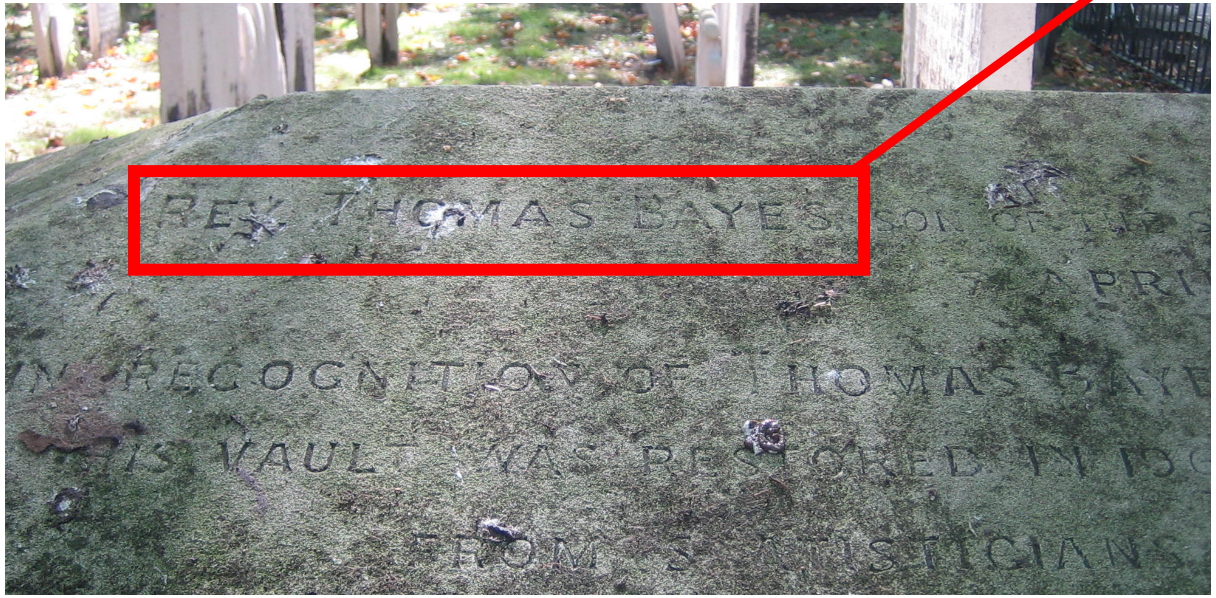
We want to compute:

$$P(State\ S\ |\ Observation\ O, Context)$$

But we are given

$$P(Observation\ O|\ State\ S)$$

Bayes derived his famous inversion equation for just this purpose.



Bayes is probably buried here
(Bunhill Fields Cemetery, London).

Bayes Theorem

We know that conjunction is commutative:

$$P(S, O) = P(O, S)$$

Using the definition of conditional probability:

$$P(S|O)P(O) = P(S, O) = P(O, S) = P(O|S)P(S)$$

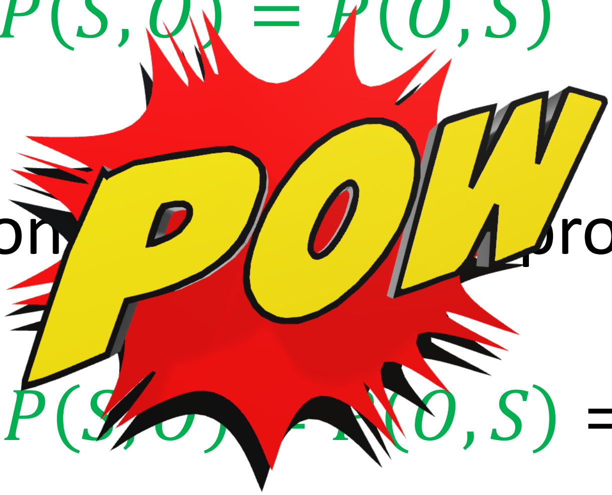
$$P(S|O)P(O) = P(O|S)P(S)$$

$$P(S|O) = \frac{P(O|S)P(S)}{P(O)}$$

Bayes Theorem

We know that conjunction is commutative:

$$P(S,O) = P(O,S)$$



Using the definition ... probability:

$$P(S|O)P(O) = P(S,O) = P(O,S) = P$$

$$|O)P(O) = P(O|S)P(S)$$

$$P(S|O) = \frac{P(O|S)P(S)}{P(O)}$$

Example

We roll one die, and an observer tells us things about the outcome.
We want to know if $X = 4$.

- Before we know anything, we believe $P(X = 4) = \frac{1}{6}$. PRIOR
- Now, suppose the observer tells us that X is even. EVIDENCE

$$P(X = 4 \mid X \text{ even}) = \frac{P(X \text{ even} \mid X = 4)P(X=4)}{P(X \text{ even})} = \frac{1 \times \frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$ Bayes

- We could also use Bayes to infer $P(X = \text{even} \mid X = 4)$:

$$P(X \text{ even} \mid X = 4) = \frac{P(X=4 \mid X \text{ even})P(X \text{ even})}{P(X=4)} = \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{6}} = 1$$ Somewhat less interesting

Interpreting Bayes theorem

- The individual terms on the right-hand side have intuitive interpretations
- We observe o, and we want to update our belief about S based on this observation.
- In this case,
  - We can think of o as evidence and $P(o)$ is the probability of observing this particular piece of evidence.
  - The function $\mathcal{L}(\mathrm{S};\mathrm{o}) \propto P(o|S)$ is called the likelihood of the state S given o.
  - The probability $P(S)$ is the prior probability for S.

$$P(S|o) = \frac{\text{likelihood } \cdot \text{prior}}{\text{evidence}} = \frac{P(o|S)P(S)}{P(0)}$$

$$\boxed{P(S|o) \propto \mathcal{L}(S;o)P(S) = \text{likelihood } \cdot \text{prior}}$$

About likelihoods…

Why do we call the conditional probability $\mathcal{L}(\mathrm{S};\mathrm{o}) \propto p(o|S)$ a *likelihood*, but we call $p(S|o)$ the *posterior*??

- We define the likelihood $\mathcal{L}(S;o)$ to be a function of S, ***not*** a function of o, i.e., *the likelihood is a function of the condition, not the observed event*:

$$\mathcal{L}(S;o) \propto p(o|S)$$

- Note: $\mathcal{L}(S;o)$ is ***not*** a probability.
- In particular, $\sum_s \mathcal{L}(S=s;o) \neq 1$

Example

- For our conductivity sensor, we defined the conditional probabilities $p(O|C)$ for each category C.
- The rows in this table represent conditional probabilities of sensor readings given object category.
- The columns in this table represent the likelihood of each category for a given sensor measurement.

Category (C)	P(False|C)	P(True|C)	Conditional probabilities – they sum to one!
Cardboard	0.99	0.01	$p(O\|Cardboard)$
Paper	0.99	0.01	$p(O\|Paper)$
Cans	0.1	0.9	$p(O\|Cans)$
Scrap Metal	0.15	0.85	$p(O\|Metal)$
Bottle	0.95	0.05	$p(O\|Bottle)$
	$\mathcal{L}(C;False)$	$\mathcal{L}(C;True)$	

A function of observation O!

Likelihoods of categories – they ***do not*** sum to one!

A function of C, parameterized by the particular sensor reading, o!

Perception

We’ve seen a lot of probability theory in the last minutes. How can we use these results to make inferences about the state of the world?

- Maximum Likelihood Estimation – simply use the likelihood
- MAP (Maximum A Posteriori) Estimation – Maximize the posterior given the sensor reading.

We’ll look now at each of these.

Maximum likelihood estimation

Recall Bayes law:

$$P(C|o) = \frac{P(o|C)P(C)}{P(o)}, \qquad o \text{ = sensor reading, } C \text{ = object category}$$

- Recall that $P(o)$ does not depend on the category of the object. It merely acts to *normalize* the posterior distribution.
- Suppose all categories are equally probably, i.e., $P(C) = \frac{1}{n}$ for each of our n Categories.
- We can now write Bayes law in a simple form:

$$P(C|o) \propto P(o|C) \propto L(C;o),$$

➢ *In this special case, maximizing the likelihood $L(C;o)$ is equivalent to maximizing the posterior probability $P(Category|observation)!$*

Maximum likelihood Estimation

We typically write the MLE problem as an optimization:

$$C^* = \arg\max_{\mathrm{C}} L(C; o)$$

in which the maximization is done w.r.t. the set

$$\mathrm{C} = \{\mathrm{Cardboard}, \mathrm{Paper}, \mathrm{Can}, \mathrm{Scrap\ Metal}, \mathrm{Bottle}\}$$

NOTE: For a given measurement, this maximization is super easy – only five values to examine.

Likelihood for continuous measurements

Recall that our weight sensor returns a continuous r.v. from a Gaussian distribution:

$$f_{W|C}(w|C) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(w-\mu)^2}{2\sigma^2}}$$

Category (C)	$f_{W|C}(W|C)$
Cardboard	$N(20, 10)$
Paper	$N(5, 5)$
Can	$N(15, 5)$
Scrap metal	$N(150, 100)$
Bottle	$N(300, 200)$

$N(\mu, \sigma^2)$ denotes the Gaussian distribution with mean and std deviation μ and σ

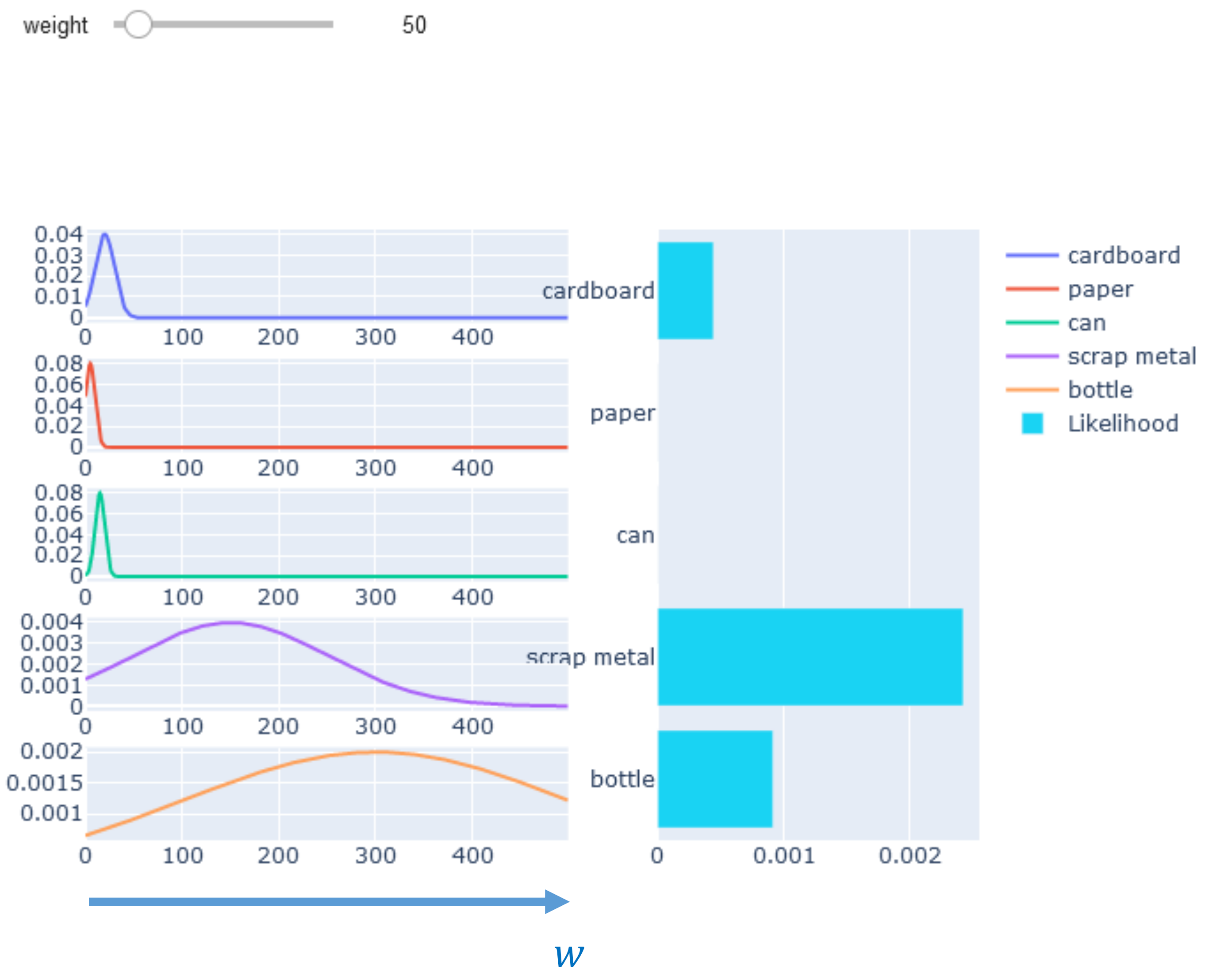
The likelihood function for category c is given by:

$$L(c; w) = \frac{1}{\sigma_c\sqrt{2\pi}} e^{-\frac{(w-\mu_c)^2}{2\sigma_c^2}}$$

For example,

$$L(Scrap\ Metal; w) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(w-150)^2}{2\ 100^2}}$$

Example



- In Section 2.4, you will find code to compute the likelihoods for all five categories, given a value for weight.
- You can play with this using the slider for weight.

For this example, we have chosen $w = 50$.

- On the left are the five conditional probabilities for the categories
- On the right are the likelihood values for $w = 50$.

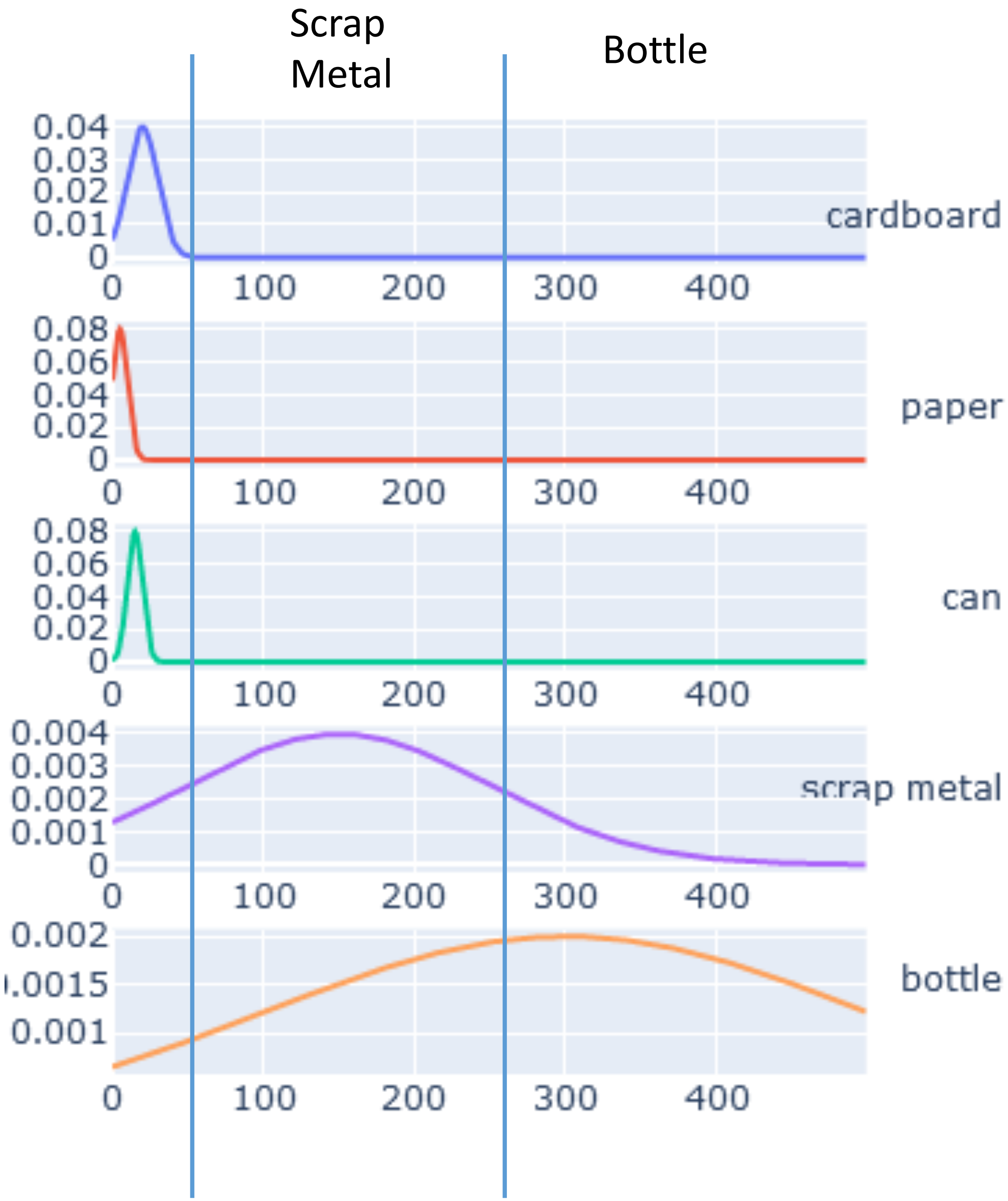
In this example, the maximum likelihood estimate is Scrap Metal.

Example (cont)

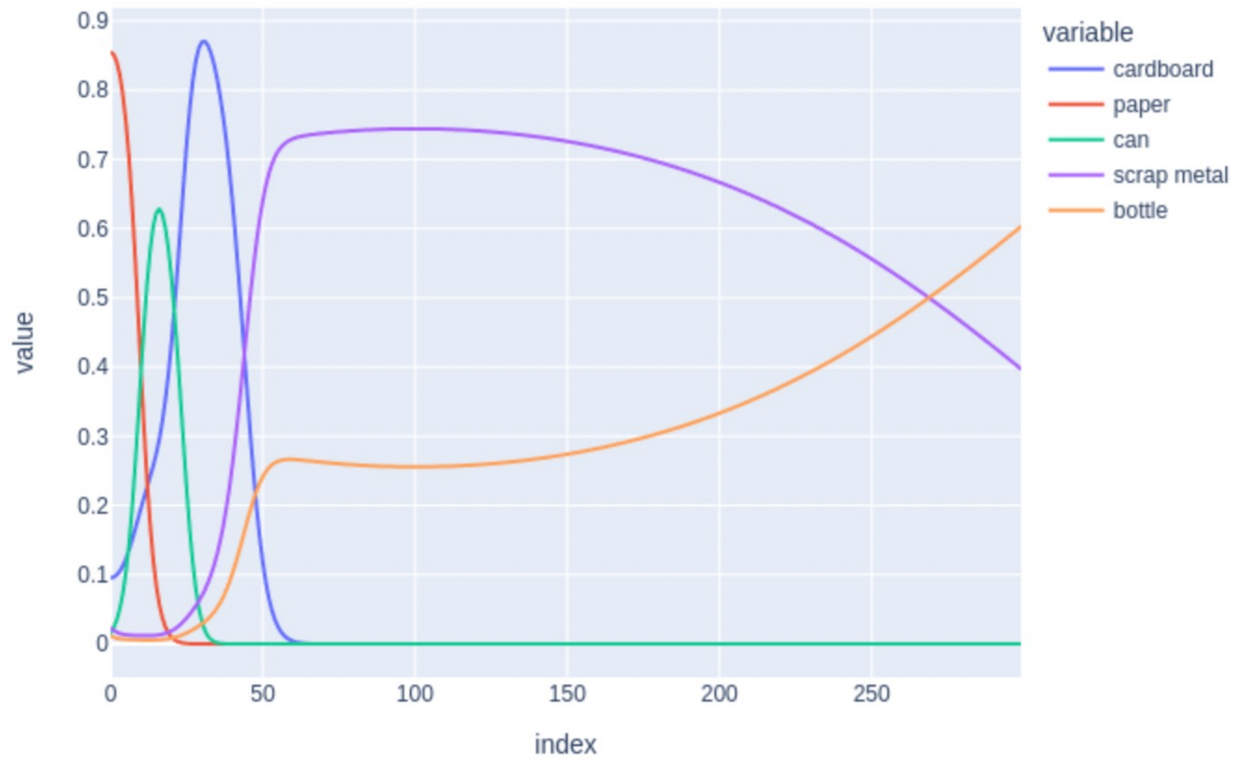
- As the weight increases, the maximum likelihood category changes from Paper to Can to Cardboard to Scrap Metal to Bottle.
- For example, Scrap Metal wins out for a long interval between approx. 45g and 270g
- Bottle becomes the MLE above 270g.

The transition points are known as *decision boundaries.*

These represent the locations in measurement space where our ML estimator changes its estimate.



MLE solution as a function of weight:



On the graph above you should see that, at as the weight increases, the maximum likelihood category changes respectively from `paper` to `can` to `cardboard`, then `scrap metal` wins out for a long interval between approx. $45g$ and $270g$, after which finally `bottle` becomes the MLE above $270g$. The transition points are known as *decision boundaries*, and represent the locations in measurement space where our ML estimator changes its estimate.

MAP estimation

- The MAP estimate is the category that maximizes the posterior probability of the category, C, given the observation, o, i.e., $C^* = \arg\max_{C} P(C|o)$
- Recall that Bayes gives the posterior as $P(C|o) = \frac{P(o|C)P(C)}{P(o)} \propto L(C;o)P(C)$
- Hence, maximizing the posterior is

$$\arg\max_{c \in C} P(c|o) = \arg\max_{c \in C} L(c;o)P(c)$$

- and the maximum a posteriori (MAP) estimate is

$$c^* = \arg\max_{c \in C} L(c;o)P(c)$$

Sensor fusion

- Suppose we have measurements from two sensors, say z_1 and z_2.
- How can we combine these measurements?
- We can still apply Bayes law:

$$P\left(C \middle| z_1, z_2\right) = \frac{P(z_1, z_2|C)P(C)}{P(z_1, z_2)} = \eta P(z_1, z_2|C)P(C)$$

- But what do we do with $P(z_1, z_2|C)$?

- We haven’t seen anything like conditional joint probabilities yet…

Conditional independence

- If we don't know the category, then measuring Z_1 might influence what we expect for Z_2.
- For example, if the object weight is small, we might expect that the object conductivity will be *False*, since Paper or Cardboard would be likely in this case.
- However, *if we knew the object category*, then observing Z_1 would not influence what we expect for Z_2.
- If we know the object is paper, it's weight will not change our expectation that conductivity will be *False*.
- This property is known as ***conditional independence***.

- We say that two random variables, say Z_1 and Z_2, are conditionally independent given C, if

$$P(Z_1, Z_2 | C) = P(Z_1 | C) P(Z_2 | C)$$

Sensor fusion

- It is straightforward to combine sensor measurements z_1 and z_2 if they are conditionally independent:

$$\begin{aligned} P\left(C \middle| z_1, z_2\right) &= \eta P(z_1|C)P(z_2|C)P(C) \\ &= L(C;z_1)L(C;z_2)P(C) \end{aligned}$$

- The posterior is proportional to the product of the likelihoods, weighted by the prior.

- The MAP estimate is now given by:

$$C^* = \arg\max_{\mathrm{C}} L(C;z_1)L(C;z_2)P(C)$$

➢ ***This idea can be extended to arbitrarily many sensor measurements.***

Planning

Planning is easy for the trash sorting robot.

- Any action can be executed at any time.
- Execution of actions has no effect on future actions.

➢ A “plan” is merely a single action, taken right now.

We’ll see four approaches:

- Maximize probability of making the right action using only prior information
- Minimizing worst-case cost using only prior information
- Minimizing expected cost using only prior information
- Incorporating sensor data

Relying on priors

- If we don't have any sensors available, the simplest decision-making strategy is to merely maximize the probability of choosing the right action.

Category	P(C)	Right Action
cardboard	0.20	Paper Bin
paper	0.30	Paper Bin
can	0.25	Metal Bin
scrap metal	0.20	Metal Bin
bottle	0.05	Glass Bin

Based on our priors:

- Placing trash in the paper bin would be the right action 50% of the time.
- Placing trash in the metal bin would be the right action 45% of the time.
- Placing trash in the glass bin would be the right action 5% of the time.

➢ ***Always place trash in the paper bin to maximize the probability of doing the right thing.***

- BUT… this approach doesn't take costs into account.
- Suppose putting paper in the metal bin could destroy trash sorting equipment.

➢ ***We can do better…***

Minimizing worst-case outcomes

In order to account for the cost of taking the wrong actions, we assigned a cost to each action for each category:

COST	cardboard	paper	can	scrap metal	bottle
glass bin	2	2	4	6	0
metal bin	1	1	0	0	2
paper bin	0	0	5	10	3
nop	1	1	1	1	1

A conservative approach to decision making is to choose an action that minimizes the worst-case costs.

From the table, we see that the worst-case costs are as follows:

- Glass bin: 6
- Metal bin: 2
- Paper bin: 10
- Nop: 1

If we want to minimize the worst-case cost, we simply choose Nop, whose cost never exceeds 1.

This approach is very conservative indeed. Now, rather than take any risk, the robot merely stands motionless, letting each piece of trash pass along to human operators.

Minimizing expected cost

- If the robot will operate for a prolonged period of time, we might prefer to minimize the average cost over a long time horizon.
- We've seen how to do this using the concept of expectation.
- Let $cost(a, c)$ denote the cost of applying action a to an object of category c.

$$E[cost(a, C)] = \sum_{c \in \Omega} cost(a, c) P(C = c)$$

COST	Card board	paper	can	scrap metal	bottle	Expected Cost
glass bin	2	2	4	6	0	3.2
metal bin	1	1	0	0	2	0.6
paper bin	0	0	5	10	3	3.4
nop	1	1	1	1	1	1.0
$P(\omega)$	0.20	0.30	0.25	0.20	0.05	

Simply compute the expected cost for applying each action under the prior distribution on categories, as we have seen in a previous lecture.

Now it's a simple matter to see that placing the object in the metal bin is the action that minimizes the expected cost.

Incorporating sensor data

To incorporate sensor data, we merely modify the expectation above so to use $P(C = c|O = o)$ instead of the prior $P(C = c)$. This is called the **conditional expectation**.

$$E[cost(a, C)|O = o] = \sum_{c \in \Omega} cost(a, c) P(C = c|O = o)$$

Choosing the best action can now be framed as a minimization problem:

$$a^* = \arg\min_a E[cost(a, C)|O = o]$$

Note that the observation $\mathrm{S} = \mathrm{o}$, is *given*, and the expectation is taken with respect to the random category C.

Multiple sensors

If we have multiple sensor readings, say $Z_1 = z_1, \dots Z_n = z_n$ we merely condition on the joint event:

$$E[cost(a, C) | Z_1 = z_1, \dots Z_n = z_n] = \sum_{c \in \Omega} cost(a, c) P(C = c | Z_1 = z_1, \dots Z_n = z_n)$$

Choosing the best action can again be framed as a minimization problem:

$$a^* = \arg\min_a E[cost(a, C) | Z_1 = z_1, \dots Z_n = z_n]$$

If the sensor data are conditionally independent given the category C, this computation can be factored, as we saw earlier.

Learning

In this chapter, all of the useful information is characterized using probability distributions.

We'll see how to use statistical methods to estimate parameters of probability distributions:

- General definitions for mean and variance (not just for the Gaussian case)
- Estimating the mean and variance
- Unbiased estimators

Learning probability distributions

If the real world can be characterized by probability distributions, the obvious question is

“H*ow do we know what is the right probability distribution?”*

We’ll answer this in two steps:

1. Develop a set of parameters that characterizes a probability distribution.
2. Develop methods to estimate those parameters from data.

The mean, μ

- For a discrete probability distribution with pmf p_X, the ***mean, μ,*** is defined as

$$\mu = E[X] = \sum_{i=1}^{n} x_i \, p_X(x_i)$$

- For a continuous distribution, the mean is defined as

$$\mu = E[X] = \int x f_X(x) dx$$

- For a Gaussian distribution, we have

$$\int x f_X(x) dx = \int x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

With a little help from a friend in an advanced calculus class.

➢ ***For a Gaussian distribution, the parameter μ, the mean, is equal to $E[X]$!***

Estimating the mean

- The mean is one of the two parameters of a Gaussian distribution.
- In fact, the mean is a valuable piece of information about every distribution we will encounter.

➢ It's worth spending some time developing a method to estimate μ.

You all know the usual estimator. For a data set $\{x_i\}_{i=1,N}$, the estimate $\hat{\mu}$ is given by

$$\hat{\mu} = \frac{1}{N}\sum x_i$$

Is this a good estimator?
How can we know if it's a good estimator?
What properties should a good estimator have?

Unbiased estimators

- Definition: The estimator $\hat{\mu}$ is said to be an **unbiased** estimator of the mean μ if $E[\hat{\mu}] = \mu$.
- On average, over many trials, $\hat{\mu}$ will be a good approximation of μ.
- Is our estimator unbiased? Let's see.

$$E[\hat{\mu}] = \mathrm{E}[\frac{1}{N}\sum X_i]$$

- Luckily, Expectation is linear!

$$\mathrm{E}[\sum \alpha_i X_i] = \sum \alpha_i E[X_i]$$

- Therefore:

$$\mathrm{E}\left[\frac{1}{N}\sum X_i\right] = \frac{1}{N}\sum E[X_i] = \frac{1}{N}\sum \mu = \frac{1}{N}N\mu = \mathrm{\mu}$$

There's a lot of good news on this slide!

➢ $\boldsymbol{\hat{\mu} = \frac{1}{N}\sum x_i}$ **is an unbiased estimator of the mean of a distribution!**

- *We never used any property of the specific distribution.*

➢ *This works for both continuous and discrete random variables (replace sums by integrals)!*

Expectation is linear (an aside)

Expectation is linear: $\boldsymbol{E[\sum \alpha_i X_i] = \sum \alpha_i E[X_i]}$

<u>Sketch of proof (for two rv's):</u>

$E[\alpha X + \beta Y] = \sum_i \sum_j (\alpha x_i + \beta y_i) p_{XY}(x_i, y_j)$ — Two random variables, X and Y, with joint pmf p_{XY}

$= \sum_i \sum_j \alpha x_i p_{XY}(x_i, y_j) + \sum_i \sum_j \beta y_j p_{XY}(x_i, y_j)$ — Apply distributivity

$= \alpha \sum_i x_i \sum_j p_{XY}(x_i, y_j) + \beta \sum_j y_j \sum_i p_{XY}(x_i, y_j)$ — Factor the sums

$= \alpha \sum_i x_i p_X(x_i) \ + \beta \sum_j y_j p_Y(y_j)$ — The marginal distribution p_X is given by $\sum_j p_{XY}(x_i, y_j)$, i.e., "integrate" out the y part of the distribution.

$= \alpha E[X] + \beta E[Y]$ — Apply the definition of expectation.

We can easily extend this to continuous r.v.'s by replacing summations with integrals, and pmf's by pdf's.

Variance

- Consider a random variable with mean μ.
- The ***variance,*** σ^2***,*** is defined as the expected value of the squared distance to the mean:

$$\sigma^2 = E[(X-\mu)^2]$$

- For a Gaussian distribution, we have

$$\int (x-\mu)^2 f_X(x)dx = \int (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

*With a **lot** of help from a friend in an advanced calculus class.*

➢ ***For a Gaussian distribution, it's not a coincidence that we use the term variance for the parameter*** σ^2

Estimating the variance

- The obvious way to estimate the variance is to merely calculate the average of the squared distance of the x_i from $\hat{\mu}$:

$$\widehat{\sigma_b}^2 = \frac{1}{N}\sum(x_i - \hat{\mu})^2$$

- Is this an unbiased estimate? (*Hint: Notice the subscript.*)

$$E[\hat{\sigma}_b{}^2] = E\left[\frac{1}{N}\sum(x_i - \hat{\mu})^2\right] = \frac{N-1}{N}\sigma^2 < \sigma^2$$

- This estimate is biased, but it's easy to fix:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum(x_i - \hat{\mu})^2$$

Biased estimate of variance (an aside)

Use every algebra trick you know…

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\,\right)^2$$

then S^2 is a biased estimator of σ^2, because

$$\operatorname{E}[S^2] = \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bigg((X_i-\mu) - (\overline{X}-\mu)\bigg)^2\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bigg((X_i-\mu)^2 - 2(\overline{X}-\mu)(X_i-\mu) + (\overline{X}-\mu)^2\bigg)\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}(\overline{X}-\mu)^2\sum_{i=1}^{n}1\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}(\overline{X}-\mu)^2\cdot n\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}-\mu)^2\right]$$

To continue, we note that by subtracting μ from both sides of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we get

$$\overline{X} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu).$$

Meaning, (by cross-multiplication) $n\cdot(\overline{X}-\mu) = \sum_{i=1}^{n}(X_i-\mu)$. Then, the previous becomes:

$$\operatorname{E}[S^2] = \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}-\mu)^2\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\cdot n\cdot(\overline{X}-\mu) + (\overline{X}-\mu)^2\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)^2 + (\overline{X}-\mu)^2\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2\right]$$

$$= \operatorname{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \operatorname{E}\left[(\overline{X}-\mu)^2\right]$$

$$= \sigma^2 - \operatorname{E}\left[(\overline{X}-\mu)^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2 < \sigma^2.$$

**See Wikipedia Bias of an estimator, or your favorite statistics book.*

Biased estimate of variance (an aside)

Use every algebra trick you know…

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\,\right)^2$$

then S^2 is a biased estimator of σ^2, because

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i \ldots$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big((X \ldots$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i \ldots$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i \ldots$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i \ldots$$

Expectation is linear.

The term $(x_i - \hat{\mu})^2$ is not linear.

And that's why we need all of this algebra….

To continue, we note that by subtracting μ from both sides of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we get

$$\overline{X} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu).$$

Meaning, (by cross-multiplication) $n \cdot (\overline{X} - \mu) = \sum_{i=1}^{n}(X_i - \mu)$. Then, the previous becomes:

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\overline{X} - \mu)^2\right]$$

$$\ldots + (\overline{X} - \mu)^2\Big]$$

$$= \sigma^2 - \mathrm{E}\left[(\overline{X} - \mu)^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2 < \sigma^2.$$

**See Wikipedia [Bias of an estimator], or your favorite statistics book.*

Learning a Gaussian distribution

- A Gaussian distribution is completely specified by its mean and variance, which is why we can write $N(\mu, \sigma^2)$. Once we know μ, σ^2, there is nothing more to know.

- In this case, $\hat{\mu}$ and $\hat{\sigma}^2$ are said to be ***sufficient statistics***.

- For a Gaussian distribution, there's simply nothing more to know, so estimating other quantities will not increase or knowledge about the underlying distribution.

$$\hat{\mu} = \frac{1}{N}\sum x_i \qquad \hat{\sigma}^2 = \frac{1}{N-1}\sum (x_i - \hat{\mu})^2$$

- In a typical statistics class, you'll spend some time studying various distributions, determining sufficient statistics for those distributions, deriving the corresponding unbiased estimators.

➢ ***Not in this class.***

Next Lecture: A Vacuum Cleaning Robot

- Simple state space: collection of rooms in a house
- Uncertainty in actions: Markov Decision Process (MDP)
- Uncertainty in sensing for a sequence of measurements: Hidden Markov Model (HMM)
- Planning using Value Iteration
- Reinforcement Learning (RL)