Lecture 24: Wrap up discussion of estimators, Markov chains

- Quiz out a week from Monday (Dec 6)
- HW 8 out tonight! Due Friday, Dec 3
- Office hours over the upcoming 8 days will change.
MLE Recipe

1. **Input** Given \( n \) iid samples \( x_1, \ldots, x_n \) from parametric model with parameter \( \theta \).

2. **Likelihood** Define your likelihood \( \mathcal{L}(x_1, \ldots, x_n | \theta) \).
   - For discrete \( \mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \Pr(x_i; \theta) \)
   - For continuous \( \mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i; \theta) \)

3. **Log** Compute \( \ln \mathcal{L}(x_1, \ldots, x_n | \theta) \)

4. **Differentiate** Compute \( \frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) \)

5. **Solve for** \( \hat{\theta} \) by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won’t ask you to do that in CSE 312.
\( n \) samples \( x_1, \ldots, x_n \in \mathbb{R} \) from Gaussian \( \mathcal{N}(\mu, \sigma^2) \). Most likely \( \mu \) and \( \sigma^2 \)?
Two-parameter optimization

Normal outcomes $x_1, \ldots, x_n$

**Goal:** estimate $\theta_1 = \mu = \text{expectation}$ and $\theta_2 = \sigma^2 = \text{variance}$

$$L(x_1, \ldots, x_n | \theta_1, \theta_2) = \left( \frac{1}{\sqrt{2\pi\theta_2}} \right)^n \prod_{i=1}^{n} e^{-\frac{(x_i - \theta_1)^2}{2\theta_2}}$$

$$\ln L(x_1, \ldots, x_n | \theta_1, \theta_2) =$$

$$= -n \frac{\ln(2\pi \theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$
Two-parameter estimation

\[ \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -n \left( \frac{\ln(2\pi \theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2} \right) \]

We need to find a solution \( \hat{\theta}_1, \hat{\theta}_2 \) to

\[ \begin{align*}
\frac{\partial}{\partial \theta_1} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) &= 0 \\
\frac{\partial}{\partial \theta_2} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) &= 0
\end{align*} \]
MLE estimates for mean and variance.

Normal outcomes $x_1, \ldots, x_n$

MLE estimator for expectation

\[ \hat{\theta}_\mu = \frac{\sum^n_i x_i}{n} \]

MLE estimator for variance

\[ \hat{\theta}_{\sigma^2} = \frac{1}{n} \sum^n_{i=1} (x_i - \hat{\theta}_\mu)^2 \]
MLE Recipe

1. **Input** Given \( n \) iid samples \( x_1, \ldots, x_n \) from parametric model with multiple parameters \( \theta = (\theta_1, \theta_2, \ldots, \theta_k) \)

2. **Likelihood** Define your likelihood function \( \mathcal{L}(x_1, \ldots, x_n \mid \theta) \).
   - For discrete \( \mathcal{L}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \Pr(x_i \mid \theta) \)
   - For continuous \( \mathcal{L}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \)

3. **Log** Compute \( \ln \mathcal{L}(x_1, \ldots, x_n \mid \theta) \)

4. **Differentiate** Compute \( \frac{\partial}{\partial \theta_i} \ln \mathcal{L}(x_1, \ldots, x_n \mid \theta) \) for each \( i \)

5. **Solve for \( \hat{\theta} \)** by setting derivatives to 0 and solving system of equations.

Generally, you need to verify that you’ve found a maximum, but we won’t ask you to do that in CSE 312.
Agenda

• Properties of estimators
• Markov chains
When is an estimator good?

Definition. An estimator of parameter $\theta$ is an unbiased estimator if

$$\mathbb{E}(\hat{\theta}_n) = \theta.$$
Example – Consistency

Normal outcomes $x_1, \ldots, x_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$  
Assume: $\sigma^2 > 0$

Unbiased

$$\hat{\theta}_\mu = \frac{\sum_i^n x_i}{n}$$

$$E(\hat{\theta}_\mu) = E\left(\frac{\sum_i^n x_i}{n}\right) = \frac{1}{n} \sum E(x_i) = \mu.$$  

Biased!

$$\hat{\theta}_{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\theta}_\mu)^2$$

$$E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\theta}_{\sigma^2})^2\right) = \sigma^2.$$
**Consistent Estimators & MLE**

**Definition.** An estimator is **unbiased** if \( \mathbb{E}(\hat{\theta}_n) = \theta \) for all \( n \geq 1 \).

**Definition.** An estimator is **consistent** if \( \lim_{n \to \infty} \mathbb{E}(\hat{\theta}_n) = \theta \).

**Theorem.** MLE estimators are consistent. (But not necessarily unbiased)

\[ f = e^{-x^2} \]
\( \hat{\Theta}_{\sigma^2} \) is biased, but consistent.

\[
E(\hat{\Theta}_{\sigma^2}) = \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \hat{\Theta}_1)^2] = \frac{1}{n} \sum_{i=1}^{n} E\left[(X_i - \frac{1}{n} \sum_{j=1}^{n} X_j)^2\right]
\]

\[
= (1 - \frac{1}{n}) \sigma^2 = \frac{n-1}{n} \sigma^2
\]

\( \hat{\Theta}_{\sigma^2} \) converges to \( \sigma^2 \), as \( n \to \infty \).

\( \hat{\Theta}_{\sigma^2} \) is “consistent”

Sample variance – Unbiased!
Agenda

- Properties of estimators
- Markov chains
So far, a single-shot random process

Random Process → Outcome Distribution $D$
So far, a single-shot random process

Random Process $\rightarrow$ Outcome Distribution $D$

Many-step random process

Random Process 1 $\rightarrow$ Outcome Distribution $D_1$ $\rightarrow$ Random Process 2 $\rightarrow$ Outcome Distribution $D_2$ $\rightarrow$ Random Process 3 $\rightarrow$ Outcome Distribution $D_3$ $\rightarrow$ ...
So far, a single-shot random process

Random Process $\rightarrow$ Outcome Distribution $D$

Today:
see a very special type of DTSP
Called a Markov Chain

Many-step random process

Random Process 1 $\rightarrow$ Outcome Distribution $D_1$ $\rightarrow$ Random Process 2 $\rightarrow$ Outcome Distribution $D_2$ $\rightarrow$ Random Process 3 $\rightarrow$ Outcome Distribution $D_3$ $\rightarrow$ ...

Definition: A discrete-time stochastic process (DTSP) is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \ldots$ where $X^{(t)}$ is the value at time $t$. 
A day in my life
A day in my life

**t = 0**

![Diagram showing states and transitions](image)

This type of probabilistic finite automaton is called a **Markov Chain**

The next state depends only on the **current state** and not on the **history**

For ANY $t \geq 0$,

- if I was working at time $t$, then at $t+1$
  - with probability 0.4 I continue working
  - with probability 0.6, I switch to surfing, and
  - with probability 0, I switch to emailing

This is called **History Independent** (similar to memoryless)
A day in my life

Many interesting questions.

1. What is the probability that I work at time 1?
2. What is the probability that I work at time 2?

\[ X(t) \text{ state I’m in at time t (random variable)} \]

\[
q_w(t) = \text{Pr}(X(t) = \text{work})
\]
\[
q_s(t) = \text{Pr}(X(t) = \text{surf})
\]
\[
q_E(t) = \text{Pr}(X(t) = \text{email})
\]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(q_w)</td>
<td>1</td>
<td>0.4</td>
<td>(q_w \cdot 0.4 + q_s \cdot 0.1)</td>
</tr>
<tr>
<td>(q_s)</td>
<td>0</td>
<td>0.6</td>
<td>(q_s \cdot 0.6 + q_s \cdot 0.6)</td>
</tr>
<tr>
<td>(q_E)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
A day in my life

Many interesting questions

1. What is the probability that I work at time 1?
2. What is the probability that I work at time 2?
3. What is the probability that I work at time \( t=100 \)?
4. What is the probability that I’m working at some random time far in the future?
A day in my life

What is the probability I’m in each state at time t, as a function of the probability distribution over states at time t-1

\[ t \geq 1 \]

\[ X(t) \text{ state I’m in at time } t \text{ (random variable)} \]

\[ q_w(t) = 0.4^{(t-1)} \]

\[ q_s(t) = 0.1^{(t-1)} \]

\[ q_E(t) = 0.5^{(t-1)} \]

\[ q_w(t-1) = \Pr(X(t-1) = \text{work}) \]

\[ q_s(t-1) = \Pr(X(t-1) = \text{surf}) \]

\[ q_E(t-1) = \Pr(X(t-1) = \text{email}) \]
\[
(q_w^{(t)}, q_s^{(t)}, q_e^{(t)}) = (q_w^{(t-1)}, q_s^{(t-1)}, q_e^{(t-1)}) \begin{pmatrix}
.4 & .6 & 0 \\
.1 & .6 & .3 \\
.5 & 0 & .5 \\
\end{pmatrix}
\]

Transition Probability Matrix

\[
P = \begin{pmatrix}
.4 & .6 & 0 \\
.1 & .6 & .3 \\
.5 & 0 & .5 \\
\end{pmatrix}
\]

\[q^{(t)} = q^{(t-1)} P\]

\[q^{(t)} = (q_w^{(t)}, q_s^{(t)}, q_e^{(t)})\]
Apply $q^{(t)} = q^{(t-1)} P$ inductively.

$P = \begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}$

$\Rightarrow q^{(t)} = q^{(0)} P^t$
The t-step walk $P^t$  

Recall $q^{(t)} = q^{(0)} P^t$  

$$P = \begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}$$

What does this say about $q^{(t)}$?
Observation

If $q^{(t)} = q^{(t-1)}$ then it will never change again!

Called a “stationary distribution” and has a special name

$$\bm{\pi} = (\pi_W, \pi_S, \pi_E)$$

Solution to $\bm{\pi} = \bm{\pi} \bm{P}$