Lecture 24: Wrap up discussion of estimators, Markov chains

Slide Credit: Based on Stefano Tessaro’s slides for 312 19au
incorporating ideas from Ryan O’Donnell, Alex Tsun, Rachel Lin, Hunter Schafer & myself 😊
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 $\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \text{Pr}(x_i ; \theta)$
   - 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

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

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

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

Normal outcomes \( x_1, \ldots, x_n \)

MLE estimator for expectation

\[
\hat{\theta}_\mu = \frac{\sum_{i=1}^{n} x_i}{n}
\]

MLE estimator for variance

\[
\hat{\theta}_{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_\mu)^2
\]
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 \( \mathcal{L}(x_1, \ldots, x_n | \theta) \).
   - 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 \( \frac{\partial}{\partial \theta_i} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) \) 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
Definition. An estimator of parameter $\theta$ is an **unbiased estimator**

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

Normal outcomes \( x_1, \ldots, x_n \) iid according to \( \mathcal{N}(\mu, \sigma^2) \) \hspace{1cm} Assume: \( \sigma^2 > 0 \)

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

\[
\hat{\theta}_{\sigma^2} = \frac{1}{n} \sum^n_{i=1} (X_i - \hat{\theta}_\mu)^2 \quad \text{Biased!}
\]
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)
\( \hat{\theta}_{\sigma^2} \) is biased, but consistent.

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

linearity

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

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

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

\[
\hat{\theta}_{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\theta}_\mu)^2
\]

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 \(D_3\) \(\rightarrow\) \(\ldots\)

**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

1. Work
2. Surf
3. Email

Diagram:
- From Work to Surf with probability 0.6
- From Surf to Work with probability 0.6
- From Work to Email with probability 0.3
- From Surf to Email with probability 0.5
- From Email to Surf with probability 0.3
A day in my life

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?

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

<table>
<thead>
<tr>
<th>( t )</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_w(t) = \Pr(X(t) = \text{work}) )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( q_s(t) = \Pr(X(t) = \text{surf}) )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( q_e(t) = \Pr(X(t) = \text{email}) )</td>
<td></td>
<td></td>
<td></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

\[ q_{w}^{(t-1)} = \Pr(X^{(t-1)} = \text{work}) \]
\[ q_{s}^{(t-1)} = \Pr(X^{(t-1)} = \text{surf}) \]
\[ q_{e}^{(t-1)} = \Pr(X^{(t-1)} = \text{email}) \]

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

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

\[ q^{(t)} = (q_w^{(t)}, q_S^{(t)}, q_E^{(t)}) \]
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

$\pi = (\pi_W, \pi_S, \pi_E)$

Solution to $\pi = \pi P$