Predictions

COMS 4771
1. Simple prediction problems
Prediction problem #1

- A coin is tossed.
Prediction problem #1

- A coin is tossed.

- **Our task**: predict the outcome (either “heads” or “tails”).
Prediction problem #1

- A coin is tossed.

- **Our task**: predict the outcome (either “heads” or “tails”).

How should we predict?
Prediction problem #1

- A coin is tossed.

- **Our task**: predict the outcome (either “heads” or “tails”).

How should we predict?

1. **Physical model**
Prediction problem #1

- A coin is tossed.

![Coin toss]

- **Our task**: predict the outcome (either “heads” or “tails”).

**How should we predict?**

1. **Physical model**

...
Prediction problem #1

- A coin is tossed.

- **Our task**: predict the outcome (either “heads” or “tails”).

How should we predict?

1. **Physical model**
2. **Statistical model**
Prediction problem #1

- A coin is tossed.

- **Our task**: predict the outcome (either “heads” or “tails”).

How should we predict?

1. **Physical model**
   - ...

2. **Statistical model**
   - Assume outcome is *random*:
     - “heads” with probability $p$, “tails” with probability $1 - p$. 
Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict "heads".
- If $p < 1/2$, then predict "tails".
- If $p = 1/2$, doesn’t matter. But, for concreteness, predict "tails".

Using this strategy, what is the probability that you predict incorrectly? Is it possible to any better?
Prediction strategy for problem #1

Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.

If we encode “heads” = 1 and “tails” = 0, we say outcome is a Bernoulli random variable $Y \sim \text{Bern}(p)$. 
Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.
- If $p < 1/2$, then predict “tails”.

Using this strategy, what is the probability that you predict incorrectly? Is it possible to do any better?
Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.
- If $p < 1/2$, then predict “tails”.
- If $p = 1/2$, doesn’t matter. But, for concreteness, predict “tails”.
Prediction strategy for problem #1

Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.
- If $p < 1/2$, then predict “tails”.
- If $p = 1/2$, doesn’t matter. But, for concreteness, predict “tails”.

Using this strategy, what is the probability that you predict incorrectly?
Prediction strategy for problem #1

Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.
- If $p < 1/2$, then predict “tails”.
- If $p = 1/2$, doesn’t matter. But, for concreteness, predict “tails”.

Using this strategy, what is the probability that you predict incorrectly?

Is it possible to do any better?
Suppose we know $p$. How should we predict?

- If $p > 1/2$, then predict “heads”.
- If $p < 1/2$, then predict “tails”.
- If $p = 1/2$, doesn’t matter. But, for concreteness, predict “tails”.

Using this strategy, what is the probability that you predict incorrectly?

Is it possible to any better?

If we encode “heads” = 1 and “tails” = 0, we say outcome is a Bernoulli random variable $Y \sim \text{Bern}(p)$. 
Prediction problem #2

A ball is dropped in a Galton board.¹

¹You can see one at the New York Hall of Science!
Prediction problem #2

▶ A ball is dropped in a Galton board.\(^1\)

▶ **Our task:** predict the (horizontal) position of the ball when it lands.

(Assume we have agreed on a coordinate system.)

\(^1\)You can see one at the New York Hall of Science!
A ball is dropped in a Galton board.\footnote{You can see one at the New York Hall of Science!}

**Our task:** predict the (horizontal) position of the ball when it lands. (Assume we have agreed on a coordinate system.)

Quality of prediction $\hat{y}$ assessed by *loss function*. We'll use *squared loss* $(\hat{y} - y)^2$. 
Statistical model: outcome is $Y \sim N(\mu, \sigma^2)$, a normal distribution.
Model for problem #2

**Statistical model:** outcome is $Y \sim N(\mu, \sigma^2)$, a *normal distribution*.

- Parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$. 

![normal distribution graph]
Statistical model: outcome is \( Y \sim N(\mu, \sigma^2) \), a normal distribution.

- Parameters \( \mu \in \mathbb{R}, \sigma^2 > 0 \).
- Probability density function (pdf) for \( Y \) is

\[
\phi_{\mu, \sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(y - \mu)^2}{2\sigma^2} \right), \quad y \in \mathbb{R}.
\]
Statistical model: outcome is \( Y \sim N(\mu, \sigma^2) \), a *normal distribution*.

- Parameters \( \mu \in \mathbb{R}, \sigma^2 > 0 \).
- Probability density function (pdf) for \( Y \) is

\[
\phi_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad y \in \mathbb{R}.
\]

- Central moments of \( Y \): \( \mathbb{E}(Y) = \mu, \text{var}(Y) = \sigma^2, \ldots \)
Suppose we know $(\mu, \sigma^2)$. How should we predict?
Suppose we know \((\mu, \sigma^2)\). How should we predict?

If we predict \(\hat{y}\), what is the expected loss (a.k.a. risk) \(\mathcal{R}(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]\)?
Suppose we know \((\mu, \sigma^2)\). How should we predict?

If we predict \(\hat{y}\), what is the *expected loss* (a.k.a. *risk*) \(\mathcal{R}(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]\)?

For any \(\hat{y} \in \mathbb{R}\),

\[
\mathcal{R}(\hat{y}) = \mathbb{E}[(\hat{y} - Y)^2]
\]
Suppose we know \((\mu, \sigma^2)\). How should we predict?

If we predict \(\hat{y}\), what is the expected loss (a.k.a. risk) \(\mathcal{R}(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]\)?

For any \(\hat{y} \in \mathbb{R}\),

\[
\mathcal{R}(\hat{y}) = \mathbb{E}[(\hat{y} - Y)^2] \\
= \mathbb{E}[(\hat{y} - \mu + \mu - Y)^2]
\]
Suppose we know \((\mu, \sigma^2)\). How should we predict?

If we predict \(\hat{y}\), what is the *expected loss* (a.k.a. *risk*) \(\mathcal{R}(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]\)?

For any \(\hat{y} \in \mathbb{R}\),

\[
\mathcal{R}(\hat{y}) = \mathbb{E}[(\hat{y} - Y)^2] \\
= \mathbb{E}[(\hat{y} - \mu + \mu - Y)^2] \\
= (\hat{y} - \mu)^2 + 2(\hat{y} - \mu)\mathbb{E}[(\mu - Y)] + \mathbb{E}[(\mu - Y)^2]
\]
Suppose we know $(\mu, \sigma^2)$. How should we predict?

If we predict $\hat{y}$, what is the expected loss (a.k.a. risk) $\mathcal{R}(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]$?

For any $\hat{y} \in \mathbb{R}$,

$$\mathcal{R}(\hat{y}) = \mathbb{E}[(\hat{y} - Y)^2] = \mathbb{E}[(\hat{y} - \mu + \mu - Y)^2] = (\hat{y} - \mu)^2 + 2(\hat{y} - \mu)\mathbb{E}[(\mu - Y)] + \mathbb{E}[(\mu - Y)^2] = (\hat{y} - \mu)^2 + \mathbb{E}[(\mu - Y)^2]$$
Suppose we know $(\mu, \sigma^2)$. How should we predict?

If we predict $\hat{y}$, what is the expected loss (a.k.a. risk) $R(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]$?

For any $\hat{y} \in \mathbb{R}$,

\[
R(\hat{y}) = \mathbb{E}[(\hat{y} - Y)^2] \\
= \mathbb{E}[(\hat{y} - \mu + \mu - Y)^2] \\
= (\hat{y} - \mu)^2 + 2(\hat{y} - \mu)\mathbb{E}[(\mu - Y)] + \mathbb{E}[(\mu - Y)^2] \\
= (\hat{y} - \mu)^2 + \mathbb{E}[(\mu - Y)^2] \\
= (\hat{y} - \mu)^2 + \text{var}(Y).
\]

So what is the best prediction?
Suppose we know \((\mu, \sigma^2)\). How should we predict?

If we predict \(\hat{y}\), what is the expected loss (a.k.a. risk) \(R(\hat{y}) := \mathbb{E}[(\hat{y} - Y)^2]\)?

For any \(\hat{y} \in \mathbb{R}\),

\[
R(\hat{y}) = \mathbb{E}[(\hat{y} - Y)^2] \\
= \mathbb{E}[(\hat{y} - \mu + \mu - Y)^2] \\
= (\hat{y} - \mu)^2 + 2(\hat{y} - \mu)\mathbb{E}[(\mu - Y)] + \mathbb{E}[(\mu - Y)^2] \\
= (\hat{y} - \mu)^2 + \mathbb{E}[(\mu - Y)^2] \\
= (\hat{y} - \mu)^2 + \text{var}(Y).
\]

So what is the best prediction?

Actually, this does not require \(Y\) to be normally distributed; it is a consequence of using squared loss.
1. **Prediction problem #1**: binary outcome $Y \sim \text{Bern}(p)$.
   
   - Loss function: *zero-one loss*
     $$\ell(\hat{y}, y) = 1 \{\hat{y} \neq y\} = \begin{cases} 
     0 & \text{if } \hat{y} = y; \\
     1 & \text{if } \hat{y} \neq y.
     \end{cases}$$

   - Optimal prediction:
     $$\hat{y}^* = 1 \{p > 1/2\} = \begin{cases} 
     1 & \text{if } p > 1/2; \\
     0 & \text{otherwise.}
     \end{cases}$$

2. **Prediction problem #2**: real-valued outcome $Y \sim \text{N}(\mu, \sigma^2)$.
   
   - Loss function: *squared loss*
     $$\ell(\hat{y}, y) = (\hat{y} - y)^2.$$ 

   - Optimal prediction:
     $$\hat{y}^* = \mu.$$
2. From data to predictions
What if we don’t know model parameters?

Often, we don’t know model parameters (e.g., $p$, $(\mu, \sigma^2)$), . . .
What if we don’t know model parameters?

Often, we don’t know model parameters (e.g., $p$, $(\mu, \sigma^2)$), . . .

but we see related observations (data) before we need to make prediction (e.g., previous balls dropped in Galton board).
What if we don’t know model parameters?

Often, we don’t know model parameters (e.g., $p$, $(\mu, \sigma^2)$), . . . but we see related observations (data) before we need to make prediction (e.g., previous balls dropped in Galton board).

**Plug-in principle:**

1. Estimate unknowns based on data.
2. Plug these estimates into formula.
What if we don’t know model parameters?

Often, we don’t know model parameters (e.g., $p$, $(\mu, \sigma^2)$), . . . but we see related observations (data) before we need to make prediction (e.g., previous balls dropped in Galton board).

Plug-in principle:

1. Estimate unknowns based on data.
2. Plug these estimates into formula.

But how is data related to outcome?
What if we don’t know model parameters?

Often, we don’t know model parameters (e.g., $p$, $(\mu, \sigma^2)$), . . .

but we see related observations (data) before we need to make prediction
(e.g., previous balls dropped in Galton board).

Plug-in principle:

1. Estimate unknowns based on data.
2. Plug these estimates into formula.

But how is data related to outcome?

IID model: Observations & outcome are independent & identically distributed (iid) random variables.
IID model: observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from Bern$(p)$, but we don’t know $p$. 

IID model for problem #1
IID model for problem #1

**IID model**: observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $\text{Bern}(p)$, but we don’t know $p$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.
IID model: observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $\text{Bern}(p)$, but we don’t know $p$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

   For problem #1, need an estimator $\hat{p}$ for $p$.
   Sometimes, we’ll explicitly write dependence on data, as in

   $$\hat{p} = \hat{p}(Y_1, \ldots, Y_n).$$
IID model for problem #1

**IID model**: observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $\text{Bern}(p)$, but we don’t know $p$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

   For problem #1, need an *estimator* $\hat{p}$ for $p$.

   Sometimes, we’ll explicitly write dependence on data, as in

   $$\hat{p} = \hat{p}(Y_1, \ldots, Y_n).$$

2. Plug estimate $\hat{p}$ into formula for optimal prediction.

   For problem #1, this is

   $$\hat{Y} := \begin{cases} 
   1 & \text{if } \hat{p} > 1/2; \\
   0 & \text{otherwise}.
   \end{cases}$$
IID model for problem #1

**IID model**: observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from Bern($p$), but we don’t know $p$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

   For problem #1, need an *estimator* $\hat{p}$ for $p$.
   Sometimes, we’ll explicitly write dependence on data, as in
   
   $$\hat{p} = \hat{p}(Y_1, \ldots, Y_n).$$

2. Plug estimate $\hat{p}$ into formula for optimal prediction.
   For problem #1, this is
   
   $$\hat{Y} := \begin{cases} 
   1 & \text{if } \hat{p} > 1/2; \\
   0 & \text{otherwise.}
   \end{cases}$$

What is a good estimator for $p$?
IID model for problem #1

**IID model:** observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $\text{Bern}(p)$, but we don’t know $p$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

   For problem #1, need an estimator $\hat{p}$ for $p$.
   Sometimes, we’ll explicitly write dependence on data, as in
   
   $$\hat{p} = \hat{p}(Y_1, \ldots, Y_n).$$

2. Plug estimate $\hat{p}$ into formula for optimal prediction.

   For problem #1, this is
   
   $$\hat{Y} := \begin{cases} 
   1 & \text{if } \hat{p} > 1/2; \\
   0 & \text{otherwise}. 
   \end{cases}$$

What is a good estimator for $p$? Let’s ask a statistician . . .
Maximum likelihood estimation

*Parametric statistical model:*

\[ \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \], a collection of probability distributions for observed data.
**Parametric statistical model:**

\[ \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \], a collection of probability distributions for observed data.

- **\( \Theta \):** parameter space.

Maximum likelihood estimator (MLE):

Let \( \hat{\theta} \) be the \( \theta \in \Theta \) of highest likelihood given observed data.
Maximum likelihood estimation

*Parametric statistical model:*
\[ \mathcal{P} = \{ P_{\theta} : \theta \in \Theta \} \], a collection of probability distributions for observed data.

- \( \Theta \): parameter space.
- \( \theta \in \Theta \): a particular parameter (or parameter vector).
Maximum likelihood estimation

**Parametric statistical model:**
\[ \mathcal{P} = \{ P_{\theta} : \theta \in \Theta \} \], a collection of probability distributions for observed data.

- \( \Theta \): parameter space.
- \( \theta \in \Theta \): a particular parameter (or parameter vector).
- \( P_{\theta} \): a particular probability distribution for observed data.

Likelihood of \( \theta \in \Theta \) given observed data:

- For discrete \( X \sim P_\theta \) with probability mass function \( p_\theta \):
  \[ L(\theta) := p_\theta(x) \].

- For continuous \( X \sim P_\theta \) with probability density function \( f_\theta \):
  \[ L(\theta) := f_\theta(x) \].

**Maximum likelihood estimator (MLE):**

Let \( \hat{\theta} \) be the \( \theta \in \Theta \) of highest likelihood given observed data.
Maximum likelihood estimation

*Parametric statistical model:* 
\[ \mathcal{P} = \{P_\theta : \theta \in \Theta\} \] , a collection of probability distributions for observed data.

- \( \Theta \): parameter space.
- \( \theta \in \Theta \): a particular parameter (or parameter vector).
- \( P_\theta \): a particular probability distribution for observed data.

*Likelihood* of \( \theta \in \Theta \) given observed data \( x \):
For discrete \( X \sim P_\theta \) with probability mass function \( p_\theta \),

\[ L(\theta) := p_\theta(x). \]

For continuous \( X \sim P_\theta \) with probability density function \( f_\theta \),

\[ L(\theta) := f_\theta(x). \]
Maximum likelihood estimation

**Parametric statistical model:**
\[ \mathcal{P} = \{ P_\theta : \theta \in \Theta \} , \text{ a collection of probability distributions for observed data.} \]
- \( \Theta \): parameter space.
- \( \theta \in \Theta \): a particular parameter (or parameter vector).
- \( P_\theta \): a particular probability distribution for observed data.

**Likelihood** of \( \theta \in \Theta \) given observed data \( x \):
For discrete \( X \sim P_\theta \) with probability mass function \( p_\theta \),
\[ L(\theta) := p_\theta(x). \]

For continuous \( X \sim P_\theta \) with probability density function \( f_\theta \),
\[ L(\theta) := f_\theta(x). \]

**Maximum likelihood estimator (MLE):**
Let \( \hat{\theta} \) be the \( \theta \in \Theta \) of highest likelihood given observed data.
MLE example for problem #1

\[ \mathcal{P} = \text{distributions on } n \text{ observations treated as iid } \text{Bern}(p) \text{ random variables.} \]
MLE example for problem #1

\( \mathcal{P} = \text{distributions on } n \text{ observations treated as iid } \text{Bern}(p) \text{ random variables.} \)

\[ \Theta = \{ p : 0 \leq p \leq 1 \}. \]
MLE example for problem #1

\[ \mathcal{P} = \text{distributions on } n \text{ observations treated as iid Bern}(p) \text{ random variables.} \]

- \[ \Theta = \{ p : 0 \leq p \leq 1 \} . \]
- Likelihood of \( p \) given data \((Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)\):

\[
\mathcal{L}(p) = \prod_{i=1}^{n} p^{y_i} (1 - p)^{1-y_i}.
\]
MLE example for problem #1

\( \mathcal{P} = \) distributions on \( n \) observations treated as iid \( \text{Bern}(p) \) random variables.

- \( \Theta = \{p : 0 \leq p \leq 1\} \).
- Likelihood of \( p \) given data \( (Y_1, \ldots, Y_n) = (y_1, \ldots, y_n) \):

\[
\mathcal{L}(p) = \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i}.
\]

- Often easier to determine maximizer of log-likelihood:

\[
\ln \mathcal{L}(p) = \sum_{i=1}^{n} y_i \ln p + (1 - y_i) \ln(1 - p).
\]
MLE example for problem #1

\( \mathcal{P} = \text{distributions on} \ n \ \text{observations treated as iid Bern}(p) \ \text{random variables.} \)

\( \Theta = \{p : 0 \leq p \leq 1\}. \)

\( \text{Likelihood of} \ p \ \text{given data} \ (Y_1, \ldots, Y_n) = (y_1, \ldots, y_n): \)

\[ L(p) = \prod_{i=1}^{n} p^{y_i} (1 - p)^{1-y_i}. \]

\( \text{Often easier to determine maximizer of log-likelihood:} \)

\[ \ln L(p) = \sum_{i=1}^{n} y_i \ln p + (1 - y_i) \ln(1 - p). \]

\( \text{Using calculus, we find that the maximizing value of} \ p \ \text{is} \)

\[ \hat{p}(y_1, \ldots, y_n) := \frac{1}{n} \sum_{i=1}^{n} y_i. \]

(See reading assignment for details.)
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid $\text{Bern}(p)$ random variables.
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid Bern$(p)$ random variables.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{p}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid Bern($p$) random variables.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

   $\hat{p}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := 1\{\hat{p}(Y_1, \ldots, Y_n) > 1/2\}.$
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid Bern($p$) random variables.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$
\hat{p}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.
$$

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := 1\{\hat{p}(Y_1, \ldots, Y_n) > 1/2\}$.

3. Outcome is $Y$, and mistake is made if $\hat{Y} \neq Y$. 

Theorem. Prediction $\hat{Y}$ satisfies

$$
P(\hat{Y} \neq Y) \leq \min\{p, 1-p\} + |2p - 1| \cdot e^{-n \cdot \text{RE}(1/2, p)}
$$

where

$$
\text{RE}(q, p) := q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}.
$$
How good is this approach?

Again, consider \( Y_1, \ldots, Y_n, Y \) iid \( \text{Bern}(p) \) random variables.

1. We observe \( Y_1, \ldots, Y_n \), and then form estimate

\[
\hat{p}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.
\]

2. We predict \( \hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := 1\{\hat{p}(Y_1, \ldots, Y_n) > 1/2\} \).

3. Outcome is \( Y \), and mistake is made if \( \hat{Y} \neq Y \).

**Theorem.** Prediction \( \hat{Y} \) satisfies

\[
P(\hat{Y} \neq Y) \leq \min\{p, 1 - p\} + |2p - 1| \cdot e^{-n \cdot \text{RE}(\frac{1}{2}, p)}
\]

where \( \text{RE}(\frac{1}{2}, p) \) is the “relative entropy between \( \text{Bern}(\frac{1}{2}) \) and \( \text{Bern}(p) \)”.

Relative entropy between \( \text{Bern}(q) \) and \( \text{Bern}(p) \) is

\[
\text{RE}(q, p) := q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}.
\]
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid Bern($p$) random variables.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{p}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ 

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := 1\{\hat{p}(Y_1, \ldots, Y_n) > 1/2\}.$

3. Outcome is $Y$, and mistake is made if $\hat{Y} \neq Y$.

**Theorem.** Prediction $\hat{Y}$ satisfies

$$\mathbb{P}(\hat{Y} \neq Y) \leq \min\{p, 1-p\} + |2p - 1| \cdot e^{-n \cdot \text{RE}(\frac{1}{2}, p)}$$

where $\text{RE}(\frac{1}{2}, p)$ is the “relative entropy between Bern($\frac{1}{2}$) and Bern($p$)”.

*Relative entropy* between Bern($q$) and Bern($p$) is

$$\text{RE}(q, p) := q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}.$$
Relative entropy

Plot of $\text{RE}(\frac{1}{2}, p) = \frac{1}{2} \ln \frac{1}{4p(1-p)}$

- $\text{RE}(\frac{1}{2}, p) = 0$ if $p = \frac{1}{2}$.
- $\text{RE}(\frac{1}{2}, p) > 0$ if $p \neq \frac{1}{2}$. 
Relative entropy

Theorem. Prediction $\hat{Y}$ satisfies

$$\Pr(\hat{Y} \neq Y) \leq \min\{p, 1 - p\} + |2p - 1| \cdot e^{-n \cdot \text{RE}(\frac{1}{2}, p)}.$$ 

- $\text{RE}(\frac{1}{2}, p) = 0$ if $p = \frac{1}{2}$.
- $\text{RE}(\frac{1}{2}, p) > 0$ if $p \neq \frac{1}{2}$. 

Recall: Optimal prediction $\hat{y}^*$ satisfies

$$\Pr(\hat{y}^* \neq Y) = \min\{p, 1 - p\}.$$ 

For this problem, on average, using MLE is near-optimal when $n$ is large!
Relative entropy

\[
\text{Plot of } \text{RE}(\frac{1}{2}, p) = \frac{1}{2} \ln \frac{1}{4p(1-p)}
\]

**Theorem.** Prediction \( \hat{Y} \) satisfies

\[
\mathbb{P}(\hat{Y} \neq Y) \leq \min\{p, 1 - p\} + |2p - 1| \cdot e^{-n \cdot \text{RE}(\frac{1}{2}, p)}.
\]

Recall: Optimal prediction \( \hat{y}^* \) satisfies

\[
\mathbb{P}(\hat{y}^* \neq Y) = \min\{p, 1 - p\}.
\]

- \( \text{RE}(\frac{1}{2}, p) = 0 \) if \( p = \frac{1}{2} \).
- \( \text{RE}(\frac{1}{2}, p) > 0 \) if \( p \neq \frac{1}{2} \).
Relative entropy

Theorem. Prediction $\hat{Y}$ satisfies

$$\mathbb{P}(\hat{Y} \neq Y) \leq \min\{p, 1 - p\} + |2p - 1| \cdot e^{-n \cdot \text{RE}(\frac{1}{2}, p)}.$$ 

Recall: Optimal prediction $\hat{y}^*$ satisfies

$$\mathbb{P}(\hat{y}^* \neq Y) = \min\{p, 1 - p\}.$$ 

For this problem, on average, using MLE is near-optimal when $n$ is large!

- $\text{RE}(\frac{1}{2}, p) = 0$ if $p = \frac{1}{2}$.
- $\text{RE}(\frac{1}{2}, p) > 0$ if $p \neq \frac{1}{2}$. 

Plot of $\text{RE}(\frac{1}{2}, p) = \frac{1}{2} \ln \frac{1}{4p(1-p)}$.
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p \leq 1/2$, then

$$
P(\hat{Y} \neq Y)
$$

Similarly, if $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p > 1/2$, then

$$
P(\hat{Y} \neq Y) = (1 - p) + (2p - 1) \cdot P(\hat{Y} = 0)
$$
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p \leq 1/2$, then

$$
P(\hat{Y} \neq Y) = (1 - p) \cdot P(\hat{Y} = 1) + p \cdot P(\hat{Y} = 0)$$
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim \text{iid Bern}(p)$ for some $p \leq 1/2$, then

$$
\mathbb{P} ( \hat{Y} \neq Y ) = (1 - p) \cdot \mathbb{P} ( \hat{Y} = 1 ) + p \cdot \mathbb{P} ( \hat{Y} = 0 )
$$

$$
= (1 - p) \cdot \mathbb{P} ( \hat{Y} = 1 ) + p \cdot \{ 1 - \mathbb{P} ( \hat{Y} = 1 ) \}
$$
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p \leq 1/2$, then

\[
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \mathbb{P}(\hat{Y} = 0)
\]

\[
= (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \{1 - \mathbb{P}(\hat{Y} = 1)\}
\]

\[
= p + (1 - 2p) \cdot \mathbb{P}(\hat{Y} = 1)
\]
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p \leq 1/2$, then

$$
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \mathbb{P}(\hat{Y} = 0)
$$

$$
= (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \{1 - \mathbb{P}(\hat{Y} = 1)\}
$$

$$
= p + (1 - 2p) \cdot \mathbb{P}(\hat{Y} = 1)
$$

$$
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n > n/2).
$$
Proof of theorem

If \( Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p) \) for some \( p \leq 1/2 \), then

\[
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \mathbb{P}(\hat{Y} = 0)
\]

\[
= (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \{1 - \mathbb{P}(\hat{Y} = 1)\}
\]

\[
= p + (1 - 2p) \cdot \mathbb{P}(\hat{Y} = 1)
\]

\[
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n > n/2).
\]

Similarly, if \( Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p) \) for some \( p > 1/2 \), then

\[
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) + (2p - 1) \cdot \mathbb{P}(\hat{Y} = 0)
\]

\[
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n \leq n/2).
\]
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p \leq 1/2$, then

\[
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \mathbb{P}(\hat{Y} = 0) \\
= (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \{1 - \mathbb{P}(\hat{Y} = 1)\} \\
= p + (1 - 2p) \cdot \mathbb{P}(\hat{Y} = 1) \\
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n > n/2).
\]

Similarly, if $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p > 1/2$, then

\[
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) + (2p - 1) \cdot \mathbb{P}(\hat{Y} = 0) \\
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n \leq n/2).
\]

We know, by linearity of expectation,

\[
\mathbb{E}[Y_1 + \cdots + Y_n] = p \cdot n.
\]
Proof of theorem

If $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p \leq 1/2$, then

$$
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \mathbb{P}(\hat{Y} = 0)
= (1 - p) \cdot \mathbb{P}(\hat{Y} = 1) + p \cdot \{1 - \mathbb{P}(\hat{Y} = 1)\}
= p + (1 - 2p) \cdot \mathbb{P}(\hat{Y} = 1)
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n > n/2).
$$

Similarly, if $Y_1, \ldots, Y_n, Y \sim_{iid} \text{Bern}(p)$ for some $p > 1/2$, then

$$
\mathbb{P}(\hat{Y} \neq Y) = (1 - p) + (2p - 1) \cdot \mathbb{P}(\hat{Y} = 0)
= \min\{p, 1 - p\} + |2p - 1| \cdot \mathbb{P}(Y_1 + \cdots + Y_n \leq n/2).
$$

We know, by linearity of expectation,

$$
\mathbb{E}[Y_1 + \cdots + Y_n] = p \cdot n.
$$

What is the probability that $Y_1 + \cdots + Y_n$ deviates above (or below) its mean?
Assume $Y_1, \ldots, Y_n \sim_{	ext{iid}} \text{Bern}(p), p \leq 1/2$. What is $\mathbb{P}(Y_1 + \cdots + Y_n > n/2)$?
Assume $Y_1, \ldots, Y_n \sim_{iid} \text{Bern}(p), \ p \leq 1/2$. What is $\mathbb{P}(Y_1 + \cdots + Y_n > n/2)$?

For any $y = (y_1, \ldots, y_n) \in \{0, 1\}^n$, let $\text{#heads}(y) := \sum_{i=1}^{n} y_i$. Also let

$$\mathcal{E} := \{ y \in \{0, 1\}^n : \text{#heads}(y) > n/2 \}.$$ 

Intuitively, these are the outcomes that are “unlikely” under $\text{Bern}(p)$. 

Proof of theorem (continued)
Assume $Y_1, \ldots, Y_n \sim_{	ext{iid}} \text{Bern}(p)$, $p \leq 1/2$. What is $\mathbb{P}(Y_1 + \cdots + Y_n > n/2)$?

For any $y = (y_1, \ldots, y_n) \in \{0, 1\}^n$, let $\#\text{heads}(y) := \sum_{i=1}^n y_i$. Also let

$$E := \{y \in \{0, 1\}^n : \#\text{heads}(y) > n/2\}.$$

Intuitively, these are the outcomes that are “unlikely” under $\text{Bern}(p)$.

Let $f := \text{pmf for Bern}(p)$, and let $g := \text{pmf for Bern}(1/2)$. 

Proof of theorem (continued)
Proof of theorem (continued)

Assume $Y_1, \ldots, Y_n \sim_{iid} \text{Bern}(p)$, $p \leq 1/2$. What is $\Pr(Y_1 + \cdots + Y_n > n/2)$?

For any $\mathbf{y} = (y_1, \ldots, y_n) \in \{0, 1\}^n$, let $\#\text{heads}(\mathbf{y}) := \sum_{i=1}^{n} y_i$. Also let

$$E := \{\mathbf{y} \in \{0, 1\}^n : \#\text{heads}(\mathbf{y}) > n/2\}.$$

Intuitively, these are the outcomes that are “unlikely” under Bern$(p)$.

Let $f := \text{pmf for Bern}(p)$, and let $g := \text{pmf for Bern}(1/2)$. Then

$$\Pr(Y_1 + \cdots + Y_n > n/2) = \sum_{\mathbf{y} \in E} f(\mathbf{y}) = \sum_{\mathbf{y} \in E} g(\mathbf{y}) \cdot \frac{f(\mathbf{y})}{g(\mathbf{y})}.$$
Proof of theorem (continued)

Assume $Y_1, \ldots, Y_n \sim_{iid} \text{Bern}(p)$, $p \leq 1/2$. What is $\mathbb{P}(Y_1 + \cdots + Y_n > n/2)$?

For any $y = (y_1, \ldots, y_n) \in \{0, 1\}^n$, let $\#\text{heads}(y) := \sum_{i=1}^n y_i$. Also let

$$\mathcal{E} := \{y \in \{0, 1\}^n : \#\text{heads}(y) > n/2\}.$$

Intuitively, these are the outcomes that are “unlikely” under Bern$(p)$.

Let $f := \text{pmf for Bern}(p)$, and let $g := \text{pmf for Bern}(1/2)$.

Then

$$\mathbb{P}(Y_1 + \cdots + Y_n > n/2) = \sum_{y \in \mathcal{E}} f(y) = \sum_{y \in \mathcal{E}} g(y) \cdot \frac{f(y)}{g(y)}.$$

The ratio $f(y)/g(y)$ is the likelihood ratio that compares likelihood of Bern$(p)$ to likelihood of Bern$(1/2)$ given observation $y$. 
Assume $Y_1, \ldots, Y_n \sim_{\text{iid}} \text{Bern}(p)$, $p \leq 1/2$. What is $\mathbb{P}(Y_1 + \cdots + Y_n > n/2)$?

For any $y = (y_1, \ldots, y_n) \in \{0, 1\}^n$, let $\#\text{heads}(y) := \sum_{i=1}^n y_i$. Also let

$$\mathcal{E} := \{y \in \{0, 1\}^n : \#\text{heads}(y) > n/2\}.$$

Intuitively, these are the outcomes that are “unlikely” under $\text{Bern}(p)$.

Let $f := \text{pmf for } \text{Bern}(p)$, and let $g := \text{pmf for } \text{Bern}(1/2)$.

Then

$$\mathbb{P}(Y_1 + \cdots + Y_n > n/2) = \sum_{y \in \mathcal{E}} f(y) = \sum_{y \in \mathcal{E}} g(y) \cdot \frac{f(y)}{g(y)}.$$

The ratio $f(y)/g(y)$ is the likelihood ratio that compares likelihood of $\text{Bern}(p)$ to likelihood of $\text{Bern}(1/2)$ given observation $y$.

We’ll prove that for any $y \in \mathcal{E}$, the likelihood ratio is exponentially small in $n$.

(Proof for case where $p > 1/2$ is similar.)
Proof of theorem (finale)

Pick any $y$ with $h := \#\text{heads}(y) > n/2$. 

Likelihood ratio of $\text{Bern}(p)$ to $\text{Bern}(1/2)$ given observation $y$:

$$f(y) = p \cdot (1 - p) \leq (p \cdot (1 - p)) \leq e^{-n \cdot \text{RE}(1/2, p)}.$$
Proof of theorem (finale)

Pick any $y$ with $h := \#\text{heads}(y) > n/2$.

Likelihood ratio of $\text{Bern}(p)$ (where $p \leq 1/2$) to $\text{Bern}(1/2)$ given observation $y$:

$$\frac{f(y)}{g(y)}$$
Proof of theorem (finale)

Pick any \( y \) with \( h := \#\text{heads}(y) > n/2 \).

Likelihood ratio of \( \text{Bern}(p) \) (where \( p \leq 1/2 \)) to \( \text{Bern}(1/2) \) given observation \( y \):

\[
\frac{f(y)}{g(y)} = \frac{p^{\#\text{heads}(y)} \cdot (1 - p)^{\#\text{tails}(y)}}{(1/2)^{\#\text{heads}(y)} \cdot (1/2)^{\#\text{tails}(y)}}
\]
Proof of theorem (finale)

Pick any $y$ with $h := \#\text{heads}(y) > n/2$.

Likelihood ratio of $\text{Bern}(p)$ (where $p \leq 1/2$) to $\text{Bern}(1/2)$ given observation $y$:

$$\frac{f(y)}{g(y)} = \frac{p^{\#\text{heads}(y)} \cdot (1 - p)^{\#\text{tails}(y)}}{(1/2)^{\#\text{heads}(y)} \cdot (1/2)^{\#\text{tails}(y)}}$$

$$= \left( \frac{p}{1/2} \right)^h \cdot \left( \frac{1 - p}{1/2} \right)^{n-h}$$
Proof of theorem (finale)

Pick any \( y \) with \( h := \#\text{heads}(y) > n/2 \).

Likelihood ratio of \( \text{Bern}(p) \) (where \( p \leq 1/2 \)) to \( \text{Bern}(1/2) \) given observation \( y \):

\[
\frac{f(y)}{g(y)} = \frac{p^{\#\text{heads}(y)} \cdot (1 - p)^{\#\text{tails}(y)}}{(1/2)^{\#\text{heads}(y)} \cdot (1/2)^{\#\text{tails}(y)}}
\]

\[
= \left( \frac{p}{1/2} \right)^h \cdot \left( \frac{1 - p}{1/2} \right)^{n-h}
\]

\[
\leq \left( \frac{p}{1/2} \right)^{n/2} \cdot \left( \frac{1 - p}{1/2} \right)^{n/2}
\]
Proof of theorem (finale)

Pick any \( y \) with \( h := \#\text{heads}(y) > n/2 \).

Likelihood ratio of \( \text{Bern}(p) \) (where \( p \leq 1/2 \)) to \( \text{Bern}(1/2) \) given observation \( y \):

\[
\frac{f(y)}{g(y)} = \frac{p^{\#\text{heads}(y)} \cdot (1 - p)^{\#\text{tails}(y)}}{(1/2)^{\#\text{heads}(y)} \cdot (1/2)^{\#\text{tails}(y)}}
= \left( \frac{p}{1/2} \right)^h \cdot \left( \frac{1 - p}{1/2} \right)^{n-h}
\leq \left( \frac{p}{1/2} \right)^{n/2} \cdot \left( \frac{1 - p}{1/2} \right)^{n/2}
= (4p(1-p))^{n/2} = e^{-n \cdot \text{RE}(1/2, p)}.
\]
Proof of theorem (finale)

Pick any $y$ with $h := \#\text{heads}(y) > n/2$.

Likelihood ratio of $\text{Bern}(p)$ (where $p \leq 1/2$) to $\text{Bern}(1/2)$ given observation $y$:

$$\frac{f(y)}{g(y)} = \frac{p^{\#\text{heads}(y)} \cdot (1 - p)^{\#\text{tails}(y)}}{(1/2)^{\#\text{heads}(y)} \cdot (1/2)^{\#\text{tails}(y)}}$$

$$= \left( \frac{p}{1/2} \right)^h \cdot \left( \frac{1 - p}{1/2} \right)^{n-h}$$

$$\leq \left( \frac{p}{1/2} \right)^{n/2} \cdot \left( \frac{1 - p}{1/2} \right)^{n/2}$$

$$= (4p(1 - p))^{n/2} = e^{-n \cdot \text{RE}(1/2, p)}.$$

Therefore

$$\mathbb{P}(Y_1 + \cdots + Y_n > n/2) = \sum_{y \in \mathcal{E}} g(y) \cdot \frac{f(y)}{g(y)} \leq \sum_{y \in \mathcal{E}} g(y) \cdot e^{-n \cdot \text{RE}(1/2, p)}$$

$$\leq e^{-n \cdot \text{RE}(1/2, p)}.$$
**IID model for problem #2**

**IID model:** Observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $N(\mu, \sigma^2)$, but we don’t know $\mu$ or $\sigma^2$. 
IID model: Observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $N(\mu, \sigma^2)$, but we don’t know $\mu$ or $\sigma^2$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.
IID model: Observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $\mathcal{N}(\mu, \sigma^2)$, but we don’t know $\mu$ or $\sigma^2$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

For problem #2, need an estimator $\hat{\mu}$ for $\mu$. (Formula doesn’t involve $\sigma^2$.)

Sometimes, we’ll explicitly write dependence on data, as in

$$\hat{\mu} = \hat{\mu}(Y_1, \ldots, Y_n).$$
IID model: Observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $\mathcal{N}(\mu, \sigma^2)$, but we don’t know $\mu$ or $\sigma^2$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

   For problem #2, need an estimator $\hat{\mu}$ for $\mu$. (Formula doesn’t involve $\sigma^2$.) Sometimes, we’ll explicitly write dependence on data, as in

   $$\hat{\mu} = \hat{\mu}(Y_1, \ldots, Y_n).$$

2. Plug estimate $\hat{\mu}$ into formula for optimal prediction.

   For problem #2, this is

   $$\hat{Y} := \hat{\mu}.$$
IID model: Observations $Y_1, \ldots, Y_n$ and outcome $Y$ are iid from $N(\mu, \sigma^2)$, but we don't know $\mu$ or $\sigma^2$.

1. Use $Y_1, \ldots, Y_n$ to estimate unknowns.

   For problem #2, need an estimator $\hat{\mu}$ for $\mu$. (Formula doesn’t involve $\sigma^2$.) Sometimes, we’ll explicitly write dependence on data, as in
   $$\hat{\mu} = \hat{\mu}(Y_1, \ldots, Y_n).$$

2. Plug estimate $\hat{\mu}$ into formula for optimal prediction.

   For problem #2, this is
   $$\hat{Y} := \hat{\mu}.$$

What is a good estimator for $\mu$?
MLE example for problem #2

\[ \mathcal{P} = \text{distributions on } n \text{ observations treated as iid } N(\mu, \sigma^2) \text{ random variables.} \]
MLE example for problem #2

\[ \mathcal{P} = \text{distributions on } n \text{ observations treated as iid } \mathcal{N}(\mu, \sigma^2) \text{ random variables.} \]

\[ \Theta = \{ (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \}. \]
MLE example for problem #2

\[ \mathcal{P} = \text{distributions on } n \text{ observations treated as iid } N(\mu, \sigma^2) \text{ random variables.} \]

\[ \Theta = \{ (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \}. \]

\[ \text{Likelihood of } (\mu, \sigma^2) \text{ given } (Y_1, \ldots, Y_n) = (y_1, \ldots, y_n): \]

\[ L(\mu, \sigma^2) = \prod_{i=1}^{n} \phi_{\mu, \sigma^2}(y_i). \]
MLE example for problem #2

\[ P = \text{distributions on } n \text{ observations treated as iid } N(\mu, \sigma^2) \text{ random variables.} \]

- \( \Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\} \).
- Likelihood of \((\mu, \sigma^2)\) given \((Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)\):

\[
L(\mu, \sigma^2) = \prod_{i=1}^{n} \phi_{\mu, \sigma^2}(y_i).
\]

- Often easier to determine maximizer of log-likelihood:

\[
\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \ln \phi_{\mu, \sigma^2}(Y_i) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 + \frac{n}{2} \ln \frac{1}{2\pi\sigma^2}.
\]
MLE example for problem #2

\( \mathcal{P} = \text{distributions on } n \text{ observations treated as iid } N(\mu, \sigma^2) \text{ random variables.} \)

- \( \Theta = \{ (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \} \).
- Likelihood of \((\mu, \sigma^2)\) given \((Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)\):
  \[
  \mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \phi_{\mu, \sigma^2}(y_i).
  \]

- Often easier to determine maximizer of log-likelihood:
  \[
  \ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \ln \phi_{\mu, \sigma^2}(Y_i) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 + \frac{n}{2} \ln \frac{1}{2\pi\sigma^2}.
  \]

- Using calculus, we find that maximizing values of \(\mu\) and \(\sigma^2\) are
  \[
  \hat{\mu}(y_1, \ldots, y_n) := \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \hat{\sigma}^2(y_1, \ldots, y_n) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2.
  \]
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$. 
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

   $$\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ 

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := \hat{\mu}(Y_1, \ldots, Y_n)$. 

A simple computation shows that, in expectation (over $Y_1, \ldots, Y_n$ and $Y$),

$$\mathbb{E}[(\hat{y}(Y_1, \ldots, Y_n) - Y)^2] = \left(1 + \frac{1}{n}\right) \text{var}(Y).$$

Recall: optimal prediction $\hat{y}^*$ has $R(\hat{y}^*) = \mathbb{E}[(\hat{y}^* - Y)^2] = \text{var}(Y)$.

For this problem, on average, using MLE is near-optimal when $n$ is large!
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ 

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := \hat{\mu}(Y_1, \ldots, Y_n)$.

3. Outcome is $Y$, and squared loss is $(\hat{Y} - Y)^2$. 

A simple computation shows that, in expectation (over $Y_1, \ldots, Y_n$ and $Y$),

$$\mathbb{E}[(\hat{y}(Y_1, \ldots, Y_n) - Y)^2] = (1 + \frac{1}{n}) \text{var}(Y).$$ 

Recall: optimal prediction $\hat{y}^*$ has $\mathbb{R}(\hat{y}^*) = \mathbb{E}[(\hat{y}^* - Y)^2] = \text{var}(Y)$.

For this problem, on average, using MLE is near-optimal when $n$ is large!
How good is this approach?

Again, consider \( Y_1, \ldots, Y_n, Y \) iid random variables with \( \mu := \mathbb{E}(Y) \).

1. We observe \( Y_1, \ldots, Y_n \), and then form estimate

\[
\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.
\]

2. We predict \( \hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := \hat{\mu}(Y_1, \ldots, Y_n) \).

3. Outcome is \( Y \), and squared loss is \( (\hat{Y} - Y)^2 \).

A simple computation shows that, in expectation (over \( Y_1, \ldots, Y_n \) and \( Y \)),

\[
\mathbb{E}[(\hat{y}(Y_1, \ldots, Y_n) - Y)^2] = \left(1 + \frac{1}{n}\right) \text{var}(Y).
\]
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$  

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := \hat{\mu}(Y_1, \ldots, Y_n)$.

3. Outcome is $Y$, and squared loss is $(\hat{Y} - Y)^2$.

A simple computation shows that, in expectation (over $Y_1, \ldots, Y_n$ and $Y$),

$$\mathbb{E}[(\hat{y}(Y_1, \ldots, Y_n) - Y)^2] = \left(1 + \frac{1}{n}\right) \text{var}(Y).$$

Recall: optimal prediction $\hat{y}^*$ has $\mathcal{R}(\hat{y}^*) = \mathbb{E}[(\hat{y}^* - Y)^2] = \text{var}(Y)$. 

For this problem, on average, using MLE is near-optimal when $n$ is large!
How good is this approach?

Again, consider $Y_1, \ldots, Y_n, Y$ iid random variables with $\mu := \mathbb{E}(Y)$.

1. We observe $Y_1, \ldots, Y_n$, and then form estimate

$$\hat{\mu}(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ 

2. We predict $\hat{Y} = \hat{y}(Y_1, \ldots, Y_n) := \hat{\mu}(Y_1, \ldots, Y_n)$.

3. Outcome is $Y$, and squared loss is $(\hat{Y} - Y)^2$.

A simple computation shows that, in expectation (over $Y_1, \ldots, Y_n$ and $Y$),

$$\mathbb{E}[(\hat{y}(Y_1, \ldots, Y_n) - Y)^2] = \left(1 + \frac{1}{n}\right) \text{var}(Y).$$

Recall: optimal prediction $\hat{y}^*$ has $\mathcal{R}(\hat{y}^*) = \mathbb{E}[(\hat{y}^* - Y)^2] = \text{var}(Y)$.

For this problem, on average, using MLE is near-optimal when $n$ is large!
Why MLE?

In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.
In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.

- “Optimal” in a certain asymptotic sense—many decades of work in the 20th century were devoted to proving this.
In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.

▶ “Optimal” in a certain asymptotic sense—many decades of work in the 20th century were devoted to proving this.

▶ Relatively simple principle to follow—estimator is defined as the solution to a particular optimization problem.
In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.

- "Optimal" in a certain asymptotic sense—many decades of work in the 20th century were devoted to proving this.

- Relatively simple principle to follow—estimator is defined as the solution to a particular optimization problem.

(Caveat: sometimes the optimization problem doesn’t have a meaningful solution, and sometimes the problem is computationally intractable!)
In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.

- **“Optimal” in a certain asymptotic sense**—many decades of work in the 20th century were devoted to proving this.

- **Relatively simple principle to follow**—estimator is defined as the solution to a particular *optimization problem*.
  (Caveat: sometimes the optimization problem doesn’t have a meaningful solution, and sometimes the problem is computationally intractable!)

- **Formulas/algorithms already figured out for many statistical models!**
Why MLE?

In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.

- **“Optimal” in a certain asymptotic sense**—many decades of work in the 20th century were devoted to proving this.

- **Relatively simple principle to follow**—estimator is defined as the solution to a particular **optimization problem**.
  (Caveat: sometimes the optimization problem doesn’t have a meaningful solution, and sometimes the problem is computationally intractable!)

- **Formulas/algorithms already figured out for many statistical models!**

**Caveat:** Our primary goal isn’t parameter estimation; it’s prediction!

- These goals are not always aligned!
  (We’ll see a striking example of this later in the course.)
In the context of parameter estimation, MLE is regarded by some statisticians as the most natural first approach to try.

- "Optimal" in a certain asymptotic sense—many decades of work in the 20th century were devoted to proving this.

- Relatively simple principle to follow—estimator is defined as the solution to a particular optimization problem. (Caveat: sometimes the optimization problem doesn’t have a meaningful solution, and sometimes the problem is computationally intractable!)

- Formulas/algorithms already figured out for many statistical models!

**Caveat:** Our primary goal isn’t parameter estimation; it’s prediction!

- These goals are not always aligned! (We’ll see a striking example of this later in the course.)

**Next time:** prediction functions via MLE.
Other kinds of predictions

What are other kinds of predictions we may want to make?

▶ Multi-class (a.k.a. multi-category): \{1, ..., K\} (General categorical distribution)
▶ Counts: \(N\) (Poisson distribution)
▶ Durations: \(R\) (non-negative reals) (Exponential distribution)
▶ Probability distributions: \(\Delta_{K-1}\) (probability distributions over \{1, ..., K\}) (Dirichlet distribution)
▶ Sequences: \{1, ..., K\} \(N\) (Markov chains)
▶ Rankings: e.g., george \(\succ\) john \(\succ\) paul \(\succ\) ringo (Plackett-Luce distribution)
▶ And many others!
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \( \{1, \ldots, K\} \)
  (General categorical distribution)

- Counts: \( N \) (Poisson distribution)

- Durations: \( R^+ \) (non-negative reals, Exponential distribution)

- Probability distributions: \( \Delta_{K-1} \) (probability distributions over \( \{1, \ldots, K\} \)) (Dirichlet distribution)

- Sequences: \( \{1, \ldots, K\} \)^N (Markov chains)

- Rankings: e.g., \( \text{George} \succ \text{John} \succ \text{Paul} \succ \text{Ringo} \) (Plackett-Luce distribution)

- And many others!
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \( \{1, \ldots, K\} \)
  (General categorical distribution)

- Counts: \( \mathbb{N} \)
  (Poisson distribution)
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \{1, \ldots, K\}
  (General categorical distribution)

- Counts: \mathbb{N}
  (Poisson distribution)

- Durations: \mathbb{R}_+ (non-negative reals)
  (Exponential distribution)

- And many others!
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \{1, \ldots, K\}
  (General categorical distribution)

- Counts: \(\mathbb{N}\)
  (Poisson distribution)

- Durations: \(\mathbb{R}_+\) (non-negative reals)
  (Exponential distribution)

- Probability distributions: \(\Delta^{K-1}\) (probability distributions over \{1, \ldots, K\})
  (Dirichlet distribution)

- Sequences: \{1, \ldots, K\}^N (Markov chains)

- Rankings: e.g., George \(\succ\) John \(\succ\) Paul \(\succ\) Ringo (Plackett-Luce distribution)

- And many others!
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \{1, \ldots, K\}
  (General categorical distribution)

- Counts: \(\mathbb{N}\)
  (Poisson distribution)

- Durations: \(\mathbb{R}_+\) (non-negative reals)
  (Exponential distribution)

- Probability distributions: \(\Delta^{K-1}\) (probability distributions over \{1, \ldots, K\})
  (Dirichlet distribution)

- Sequences: \(\{1, \ldots, K\}^\mathbb{N}\)
  (Markov chains)

- Rankings: e.g., george \(\succ\) john \(\succ\) paul \(\succ\) ringo
  (Plackett-Luce distribution)

- And many others!
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \( \{1, \ldots, K\} \)
  (General categorical distribution)

- Counts: \( \mathbb{N} \)
  (Poisson distribution)

- Durations: \( \mathbb{R}_+ \) (non-negative reals)
  (Exponential distribution)

- Probability distributions: \( \Delta^{K-1} \) (probability distributions over \( \{1, \ldots, K\} \))
  (Dirichlet distribution)

- Sequences: \( \{1, \ldots, K\}^\mathbb{N} \)
  (Markov chains)

- Rankings: e.g., george \( \succ \) john \( \succ \) paul \( \succ \) ringo
  (Plackett-Luce distribution)

▶ And many others!
Other kinds of predictions

What are other kinds of predictions we may want to make?

- Multi-class (a.k.a. multi-category): \(\{1, \ldots, K\}\)
  (General categorical distribution)

- Counts: \(\mathbb{N}\)
  (Poisson distribution)

- Durations: \(\mathbb{R}_+\) (non-negative reals)
  (Exponential distribution)

- Probability distributions: \(\Delta^{K-1}\) (probability distributions over \(\{1, \ldots, K\}\))
  (Dirichlet distribution)

- Sequences: \(\{1, \ldots, K\}^\mathbb{N}\)
  (Markov chains)

- Rankings: e.g., george \(\succ\) john \(\succ\) paul \(\succ\) ringo
  (Plackett-Luce distribution)

- And many others!
1. Statistical models for simple prediction problems, and the optimal predictions in these models.

2. How to derive near-optimal predictions from data in iid models (for zero-one loss and squared loss).