6.7900: Machine Learning
Lecture 2

Lecture start: Tues/Thurs 2:35pm
Who's speaking today? Prof. Tamara Broderick
Course website: gradml.mit.edu
Questions? Ask here or on piazza.com/mit/fall2024/67900/
Materials: Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

Last Time

I. Logistics & motivation
II. Setup: prediction
III. Decision rules, loss, risk
IV. Case where we know the data distribution

Today's Plan

I. Empirical risk minimization
II. Modeling
III. Maximum likelihood estimate (MLE)
IV. Pros & cons of the MLE

Recap

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!
  - We need to make an assumption about how our training data relates to our future data point(s)

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!
  - We need to make an assumption about how our training data relates to our future data point(s)
  - A very common assumption: $(X^{(n)}, Y^{(n)})$ are iid for all *n* (including both training data and future data)

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!
  - We need to make an assumption about how our training data relates to our future data point(s)
  - A very common assumption: $(X^{(n)}, Y^{(n)})$ are iid for all *n* (including both training data and future data)
    - George Box: "All models are wrong but some are useful"

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

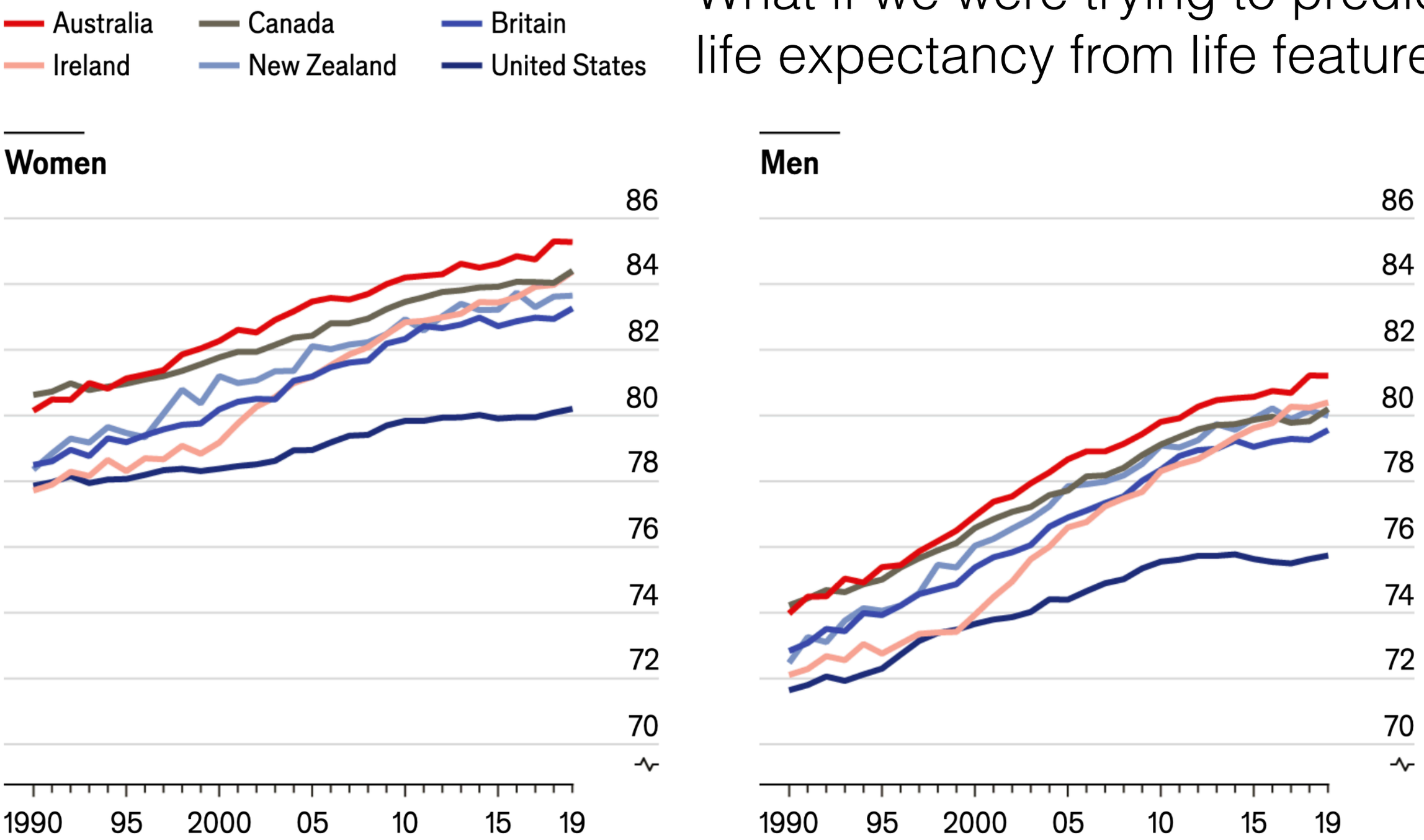
- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!
  - We need to make an assumption about how our training data relates to our future data point(s)
  - A very common assumption: $(X^{(n)}, Y^{(n)})$ are iid for all *n* (including both training data and future data)
    - George Box: "All models are wrong but some are useful"
    - Sometimes this assumption is fine, and sometimes it's not

Aside about the i.i.d. assumption

- What if we were trying to predict life expectancy from life features?

Aside about the i.i.d. assumption

- What if we were trying to predict life expectancy from life features?



Source: "Life expectancy and geographic variation in mortality: an observational comparison study of six high-income Anglophone countries", by R. Wilkie and J. Ho, *BMJ Open*, 2024

[The Economist, 2024/08/23]

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!
  - We need to make an assumption about how our training data relates to our future data point(s)
  - A very common assumption: $(X^{(n)}, Y^{(n)})$ are iid for all *n* (including both training data and future data)
    - George Box: "All models are wrong but some are useful"
    - Sometimes this assumption is fine, and sometimes it's not

Recap

features $x^{(n)} \in \mathcal{X}$, labels $y^{(n)} \in \mathcal{Y}$

- Training data $\mathcal{D} = \{\overbrace{(x^{(n)}, y^{(n)})}\}_{n=1}^{N}$
- We'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- Last time, we saw cases where, when we knew the distribution of (*X*, *Y*), choosing *h* was straightforward
- **Problem:** we don't usually know the distribution of (*X*, *Y*)
- **Idea**: Use our training data!
  - We need to make an assumption about how our training data relates to our future data point(s)
  - A very common assumption: $(X^{(n)}, Y^{(n)})$ are iid for all *n* (including both training data and future data)
    - George Box: "All models are wrong but some are useful"
    - Sometimes this assumption is fine, and sometimes it's not
- We make a lot of assumptions in machine learning. To do anything practical, you've got to make assumptions. But it's best practice to state your assumptions as clearly as possible

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

  - I.e. instead of minimizing risk (of a new point), we instead minimize **empirical risk** over the training data

risk $\mathbb{E}[L(Y, h(X))] \approx \frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$ empirical risk

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

  - I.e. instead of minimizing risk (of a new point), we instead minimize **empirical risk** over the training data

risk $\quad \mathbb{E}[L(Y, h(X))] \approx \frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$ empirical risk

- Why might this be a good approximation?

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

  - I.e. instead of minimizing risk (of a new point), we instead minimize **empirical risk** over the training data

$$\underbrace{\mathbb{E}[L(Y, h(X))]}_{\text{risk}} \approx \underbrace{\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))}_{\text{empirical risk}}$$

- Why might this be a good approximation?
  - **Law of large numbers**: Let $Z_1, Z_2, \ldots$ be iid random variables. Assume all necessary expectations exist. Then, with probability 1, $\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} Z_n = \mathbb{E}[Z_1]$

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

  - I.e. instead of minimizing risk (of a new point), we instead minimize **empirical risk** over the training data

$$\underbrace{\mathbb{E}[L(Y, h(X))]}_{\text{risk}} \approx \underbrace{\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))}_{\text{empirical risk}}$$

- Why might this be a good approximation?
  - **Law of large numbers**: Let $Z_1, Z_2, \ldots$ be iid random variables. Assume all necessary expectations exist. Then, with probability 1, $\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} Z_n = \mathbb{E}[Z_1]$
  - In our case:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} L(Y^{(n)}, h(X^{(n)})) = \mathbb{E}[L(Y, h(X))]$$

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

  - I.e. instead of minimizing risk (of a new point), we instead minimize **empirical risk** over the training data

$$\underbrace{\mathbb{E}[L(Y, h(X))]}_{\text{risk}} \approx \underbrace{\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))}_{\text{empirical risk}}$$

- Why might this be a good approximation?
  - **Law of large numbers**: Let $Z_1, Z_2, \ldots$ be iid random variables. Assume all necessary expectations exist. Then, with probability 1, $\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} Z_n = \mathbb{E}[Z_1]$
  - In our case:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} L(Y^{(n)}, h(X^{(n)})) = \mathbb{E}[L(Y, h(X))]$$

  - Let $Z \sim \text{Uniform}[0, 1]$; then $\mathbb{P}(Z \neq 0.9) = 1$

Empirical risk

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Idea**: in the risk, substitute the **empirical distribution** of the training data for "true" distribution of the data

$$\hat{p}(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x^{(n)}, y^{(n)})}(x, y)$$

  - I.e. instead of minimizing risk (of a new point), we instead minimize **empirical risk** over the training data

$$\text{risk} \quad \mathbb{E}[L(Y, h(X))] \approx \frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) \quad \text{empirical risk}$$

- Why might this be a good approximation?
  - **Law of large numbers**: Let $Z_1, Z_2, \ldots$ be iid random variables. Assume all necessary expectations exist. Then, with probability 1, $\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} Z_n = \mathbb{E}[Z_1]$
  - In our case:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} L(Y^{(n)}, h(X^{(n)})) = \mathbb{E}[L(Y, h(X))]$$

  - Let $Z \sim \text{Uniform}[0, 1]$; then $\mathbb{P}(Z \neq 0.9) = 1$
    - "Surely" or "always" vs "with prob 1" or "almost surely"

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk?

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk? Is this a good decision rule?

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk? Is this a good decision rule?
- **Generalization**: want rules to perform well on new data points that aren't exactly the same as those in the training set

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk? Is this a good decision rule?
- **Generalization**: want rules to perform well on new data points that aren't exactly the same as those in the training set
- **Overfitting**: good performance on training data but poor generalization

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk? Is this a good decision rule?
- **Generalization**: want rules to perform well on new data points that aren't exactly the same as those in the training set
- **Overfitting**: good performance on training data but poor generalization

note: difference between general concept and precise term of art

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk? Is this a good decision rule?
- **Generalization**: want rules to perform well on new data points that aren't exactly the same as those in the training set
- **Overfitting**: good performance on training data but poor generalization

note: difference between general concept and precise term of art

"Machine learning students overfit to overfitting" 2022

Empirical risk minimization

- So we'd like to choose a decision rule $h : \mathcal{X} \to \mathcal{Y}$ that minimizes risk on a new (as-yet-unseen) point: $\mathbb{E}[L(Y, h(X))]$
- **Empirical risk minimization**: Choose a decision rule *h* to minimize empirical risk over the training data

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

- Example: Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule:
  - $h(x) = 1$ if the timestamp of x matches the timestamp of any spam email in the training set exactly, else 0
  - What is the empirical risk? Is this a good decision rule?
- **Generalization**: want rules to perform well on new data points that aren't exactly the same as those in the training set
- **Overfitting**: good performance on training data but poor generalization

note: difference between general concept and precise term of art

"Machine learning students overfit to overfitting" 2022

- Either restrict *h* or approximate (X, Y) distribution differently

Modeling and maximum likelihood

- We want to approximate the distribution of the data

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example:

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta)$$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta)$$

why?

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}} (1-\theta)^{1-y^{(n)}}$$

why?

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}} (1-\theta)^{1-y^{(n)}}$$

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log \theta + (1 - y^{(n)}) \log(1-\theta)]$$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{1-y^{(n)}}$$

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log \theta + (1 - y^{(n)}) \log(1-\theta)]$$

$$\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \theta^{-1} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-1} \sum_{n=1}^{N} (1 - y^{(n)})$$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{1-y^{(n)}}$$

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log \theta + (1 - y^{(n)}) \log(1-\theta)]$$

$$\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \theta^{-1} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-1} \sum_{n=1}^{N} (1 - y^{(n)})$$

$$\frac{d^2 \log p(\mathcal{D}|\theta)}{d\theta^2} = -\theta^{-2} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-2} \sum_{n=1}^{N} (1 - y^{(n)})$$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{1-y^{(n)}}$$

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log \theta + (1 - y^{(n)}) \log(1-\theta)]$$

$$\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \theta^{-1} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-1} \sum_{n=1}^{N} (1 - y^{(n)}) \overset{\text{set}}{=} 0$$

$$\frac{d^2 \log p(\mathcal{D}|\theta)}{d\theta^2} = -\theta^{-2} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-2} \sum_{n=1}^{N} (1 - y^{(n)})$$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

- We want to approximate the distribution of the data
- For the moment, let's consider data with no features
- Assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by **parameter** $\theta \in \Theta$
  - **Parametric model**: the parameter is finite-dimensional
  - The distribution might have density or pmf $p(y|\theta)$
  - Especially as a function of θ, $p(y|\theta)$ called the **likelihood**
- Idea: choose some $\hat{\theta}$, and use $p(y|\hat{\theta})$ as the approximation
- Idea: choose $\hat{\theta}$ to maximize **likelihood of the training data**
- Example: $y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{1-y^{(n)}}$$

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log \theta + (1 - y^{(n)}) \log(1-\theta)]$$

$$\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \theta^{-1} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-1} \sum_{n=1}^{N} (1 - y^{(n)}) \overset{\text{set}}{=} 0$$

$$\frac{d^2 \log p(\mathcal{D}|\theta)}{d\theta^2} = -\theta^{-2} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-2} \sum_{n=1}^{N} (1 - y^{(n)})$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

see bonus slide at end for careful handling of edge cases

Modeling and maximum likelihood

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y},\ \hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y},\ \hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y},\ \hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models

might get stuck in local optimum

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
  - Note: use log likelihood for numerical reasons

might get stuck in local optimum

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
- In many cases of interest, optimizers are fast

might get stuck in local optimum

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice

might get stuck in local optimum

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$

might get stuck in local optimum

cf. σ^2

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$ cf. σ^2

might get stuck in local optimum

$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta)$

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta)$$

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|f^{-1}(f(\hat{\theta}))) > p(\mathcal{D}|f^{-1}(f(\theta)))$$

might get stuck in local optimum

cf. σ^2

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$

might get stuck in local optimum

cf. σ^2

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta)$$

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|f^{-1}(f(\hat{\theta}))) > p(\mathcal{D}|f^{-1}(f(\theta)))$$

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$

might get stuck in local optimum

cf. σ^2

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta)$$
$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|f^{-1}(f(\hat{\theta}))) > p(\mathcal{D}|f^{-1}(f(\theta)))$$
$$\forall \eta \neq f(\hat{\theta}), \tilde{p}(\mathcal{D}|f(\hat{\theta})) > \tilde{p}(\mathcal{D}|\eta)$$

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta)$$
$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|f^{-1}(f(\hat{\theta}))) > p(\mathcal{D}|f^{-1}(f(\theta)))$$
$$\forall \eta \neq f(\hat{\theta}), \tilde{p}(\mathcal{D}|f(\hat{\theta})) > \tilde{p}(\mathcal{D}|\eta)$$

might get stuck in local optimum

cf. σ^2

note: changing parameter, not random variable

Modeling and maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
  - Exer: $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$, $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$
  - Predictive approximation $p(y|\hat{\mu}, \hat{\sigma}^2)$ avoids the degeneracy we saw in the empirical distribution
- Advantages of maximizing the likelihood:
  - If likelihood is differentiable w.r.t. the parameter, easy to use modern gradient-based optimizers & complex models
    - Note: use log likelihood for numerical reasons
  - In many cases of interest, optimizers are fast
  - Part of many successful uses of ML in practice
  - **Maximum likelihood estimate (MLE)** is invariant to reparametrization using a bijective function $\eta = f(\theta)$

might get stuck in local optimum

cf. σ^2

note: changing parameter, not random variable

$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta)$$
$$\forall \theta \neq \hat{\theta}, p(\mathcal{D}|f^{-1}(f(\hat{\theta}))) > p(\mathcal{D}|f^{-1}(f(\theta)))$$
$$\forall \eta \neq f(\hat{\theta}), \tilde{p}(\mathcal{D}|f(\hat{\theta})) > \tilde{p}(\mathcal{D}|\eta) \qquad \Rightarrow \hat{\eta} = f(\hat{\theta})$$

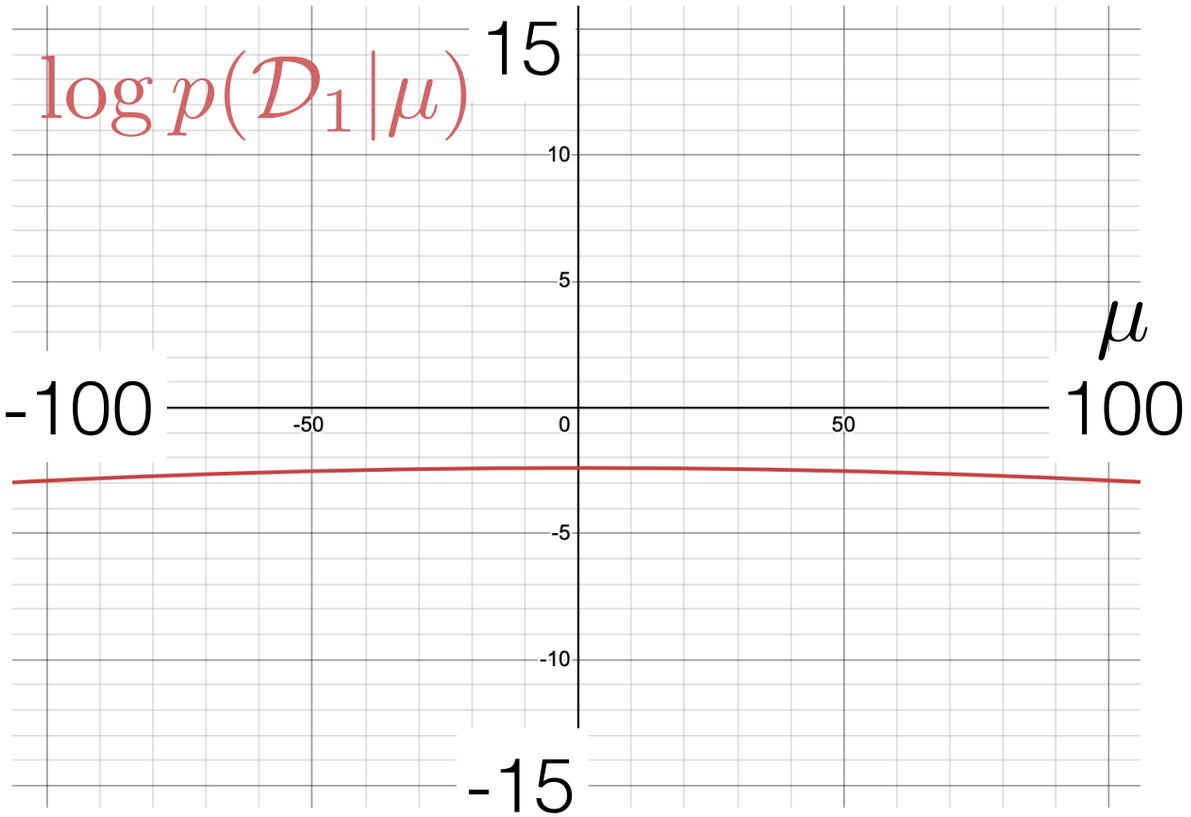
Potential issues with maximum likelihood

Potential issues with maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2

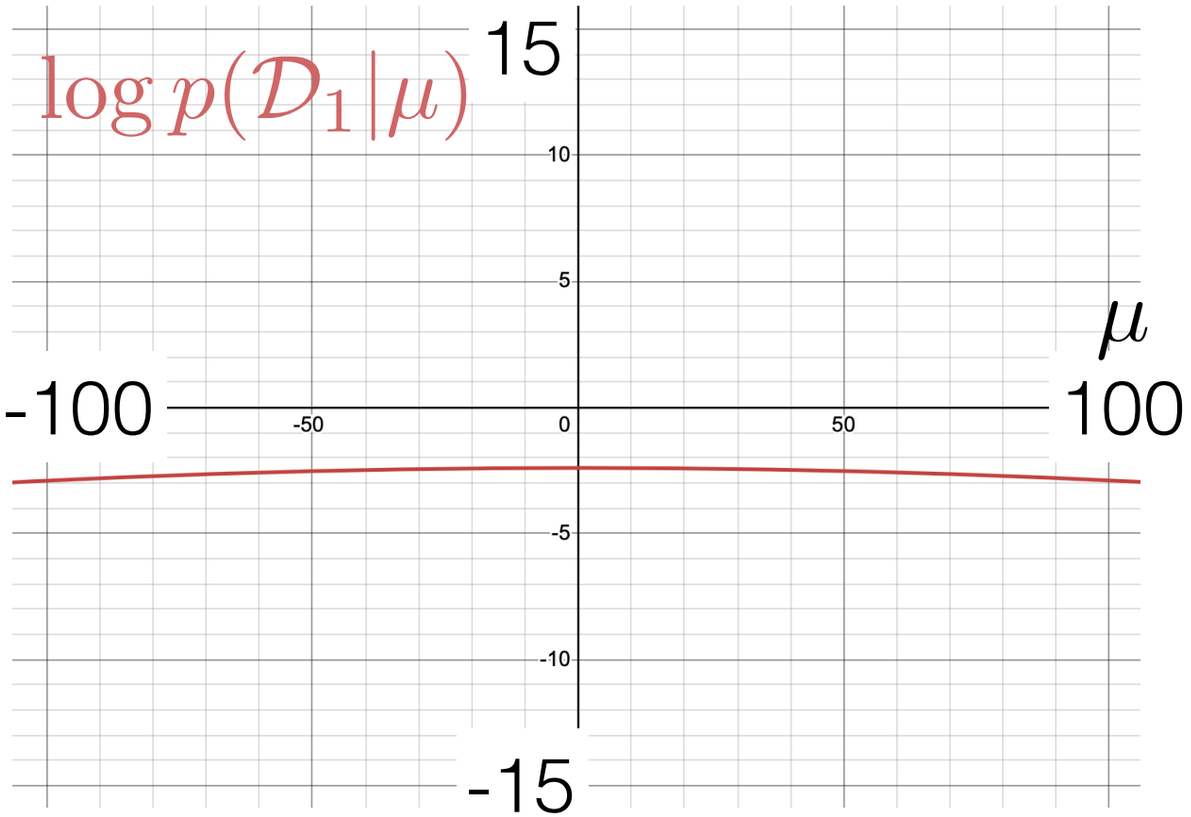
Potential issues with maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2



Potential issues with maximum likelihood

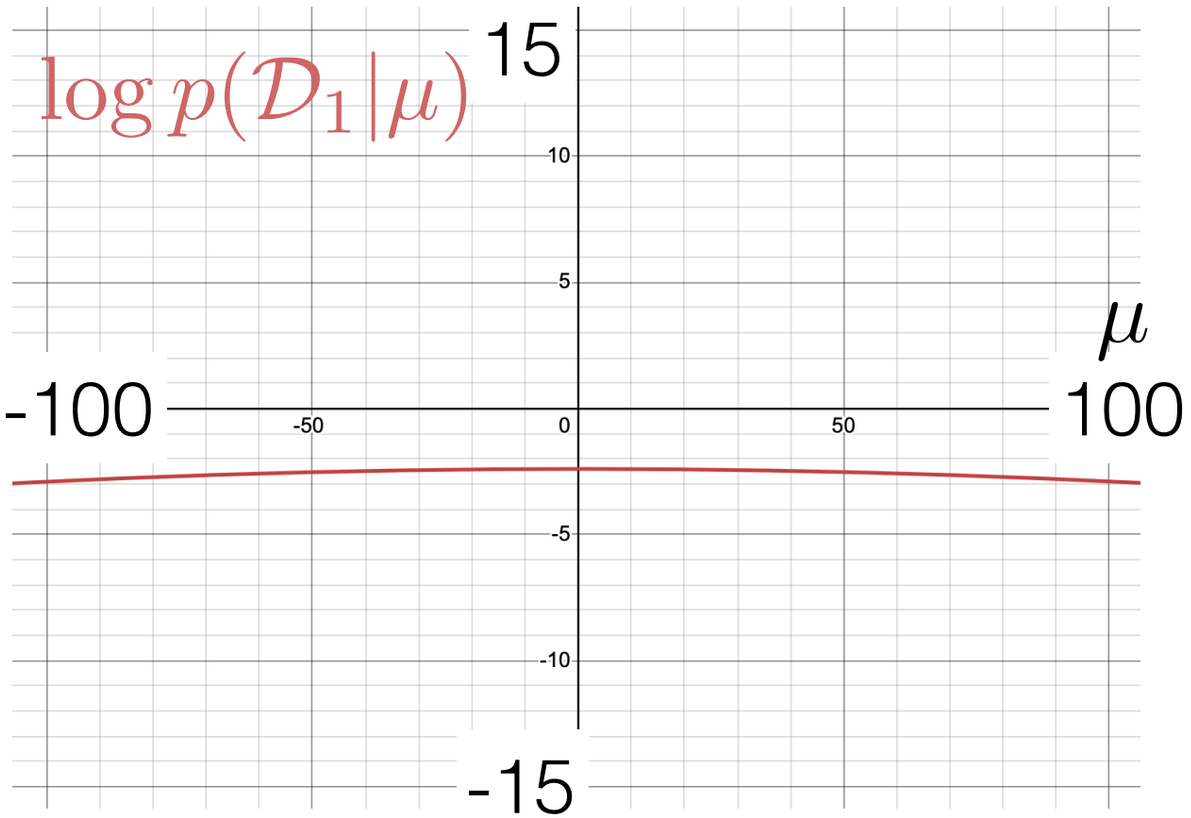
- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2



- One data point; MLE $\hat{\mu} = 0.5$

Potential issues with maximum likelihood

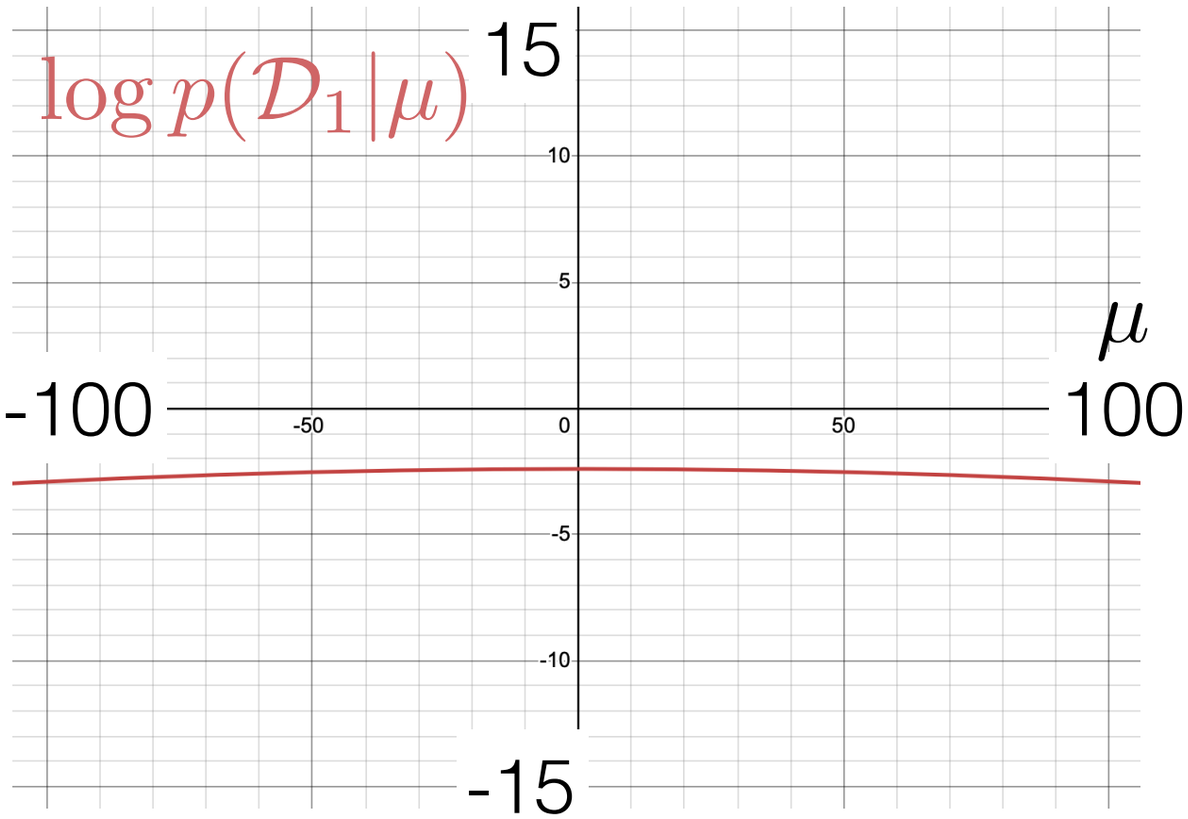
- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2



- One data point; MLE $\hat{\mu} = 0.5$
- $p(\mathcal{D}|\mu)$ within a factor of e of the likelihood at MLE for $\hat{\mu} \pm 100$

Potential issues with maximum likelihood

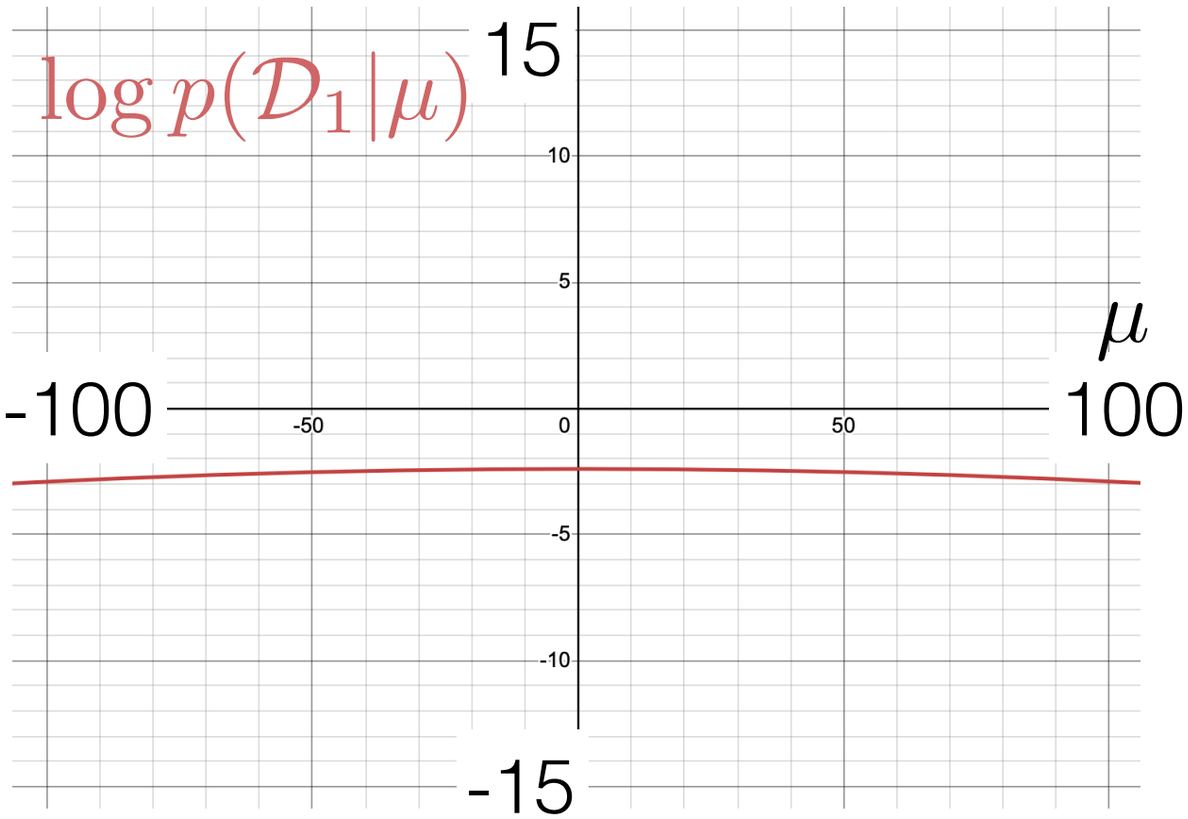
- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2



- One data point; MLE $\hat{\mu} = 0.5$
- $p(\mathcal{D}|\mu)$ within a factor of e of the likelihood at MLE for $\hat{\mu} \pm 100$
- Predictive is $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$

Potential issues with maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2

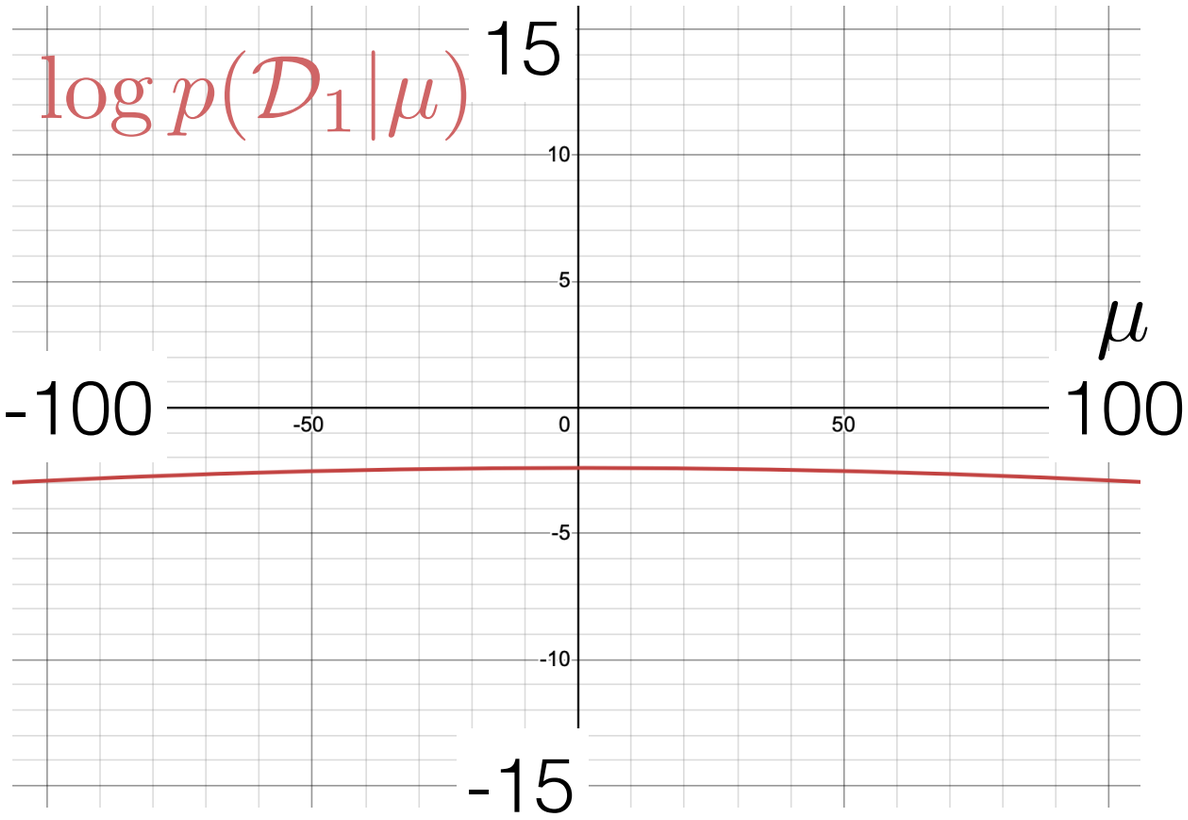


- One data point; MLE $\hat{\mu} = 0.5$
- $p(\mathcal{D}|\mu)$ within a factor of e of the likelihood at MLE for $\hat{\mu} \pm 100$
- Predictive is $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$

- 10^6 points; MLE $\hat{\mu} = 0.5$

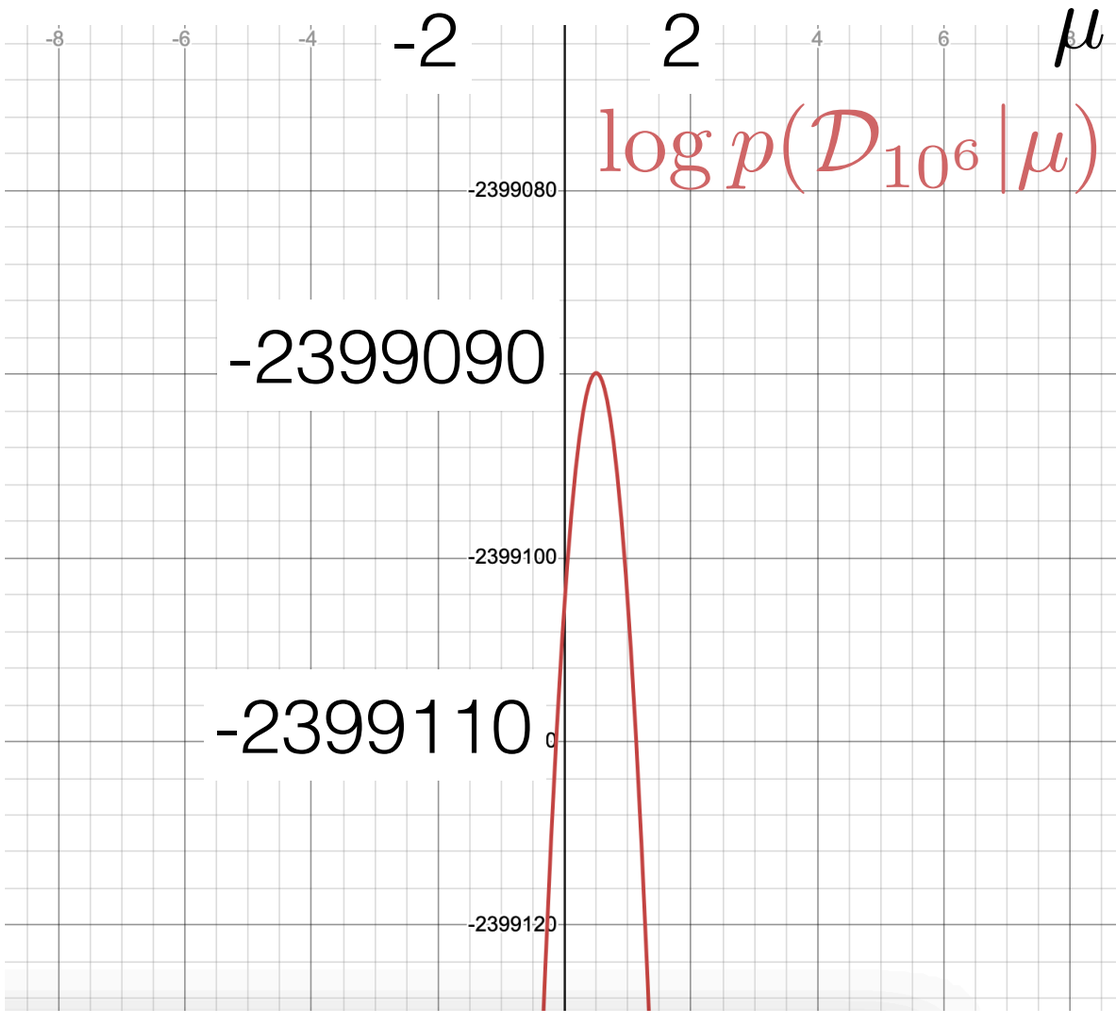
Potential issues with maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2



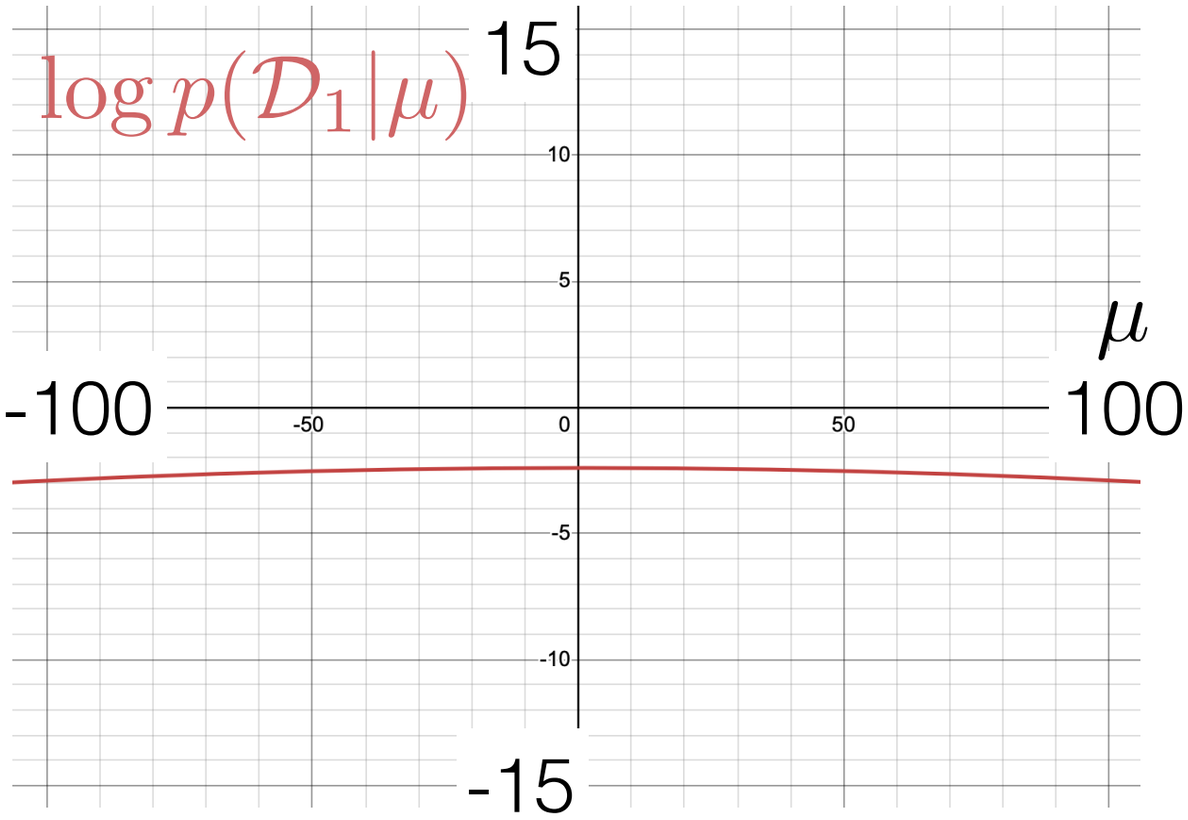
- One data point; MLE $\hat{\mu} = 0.5$
- $p(\mathcal{D}|\mu)$ within a factor of e of the likelihood at MLE for $\hat{\mu} \pm 100$
- Predictive is $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$

- 10^6 points; MLE $\hat{\mu} = 0.5$



Potential issues with maximum likelihood

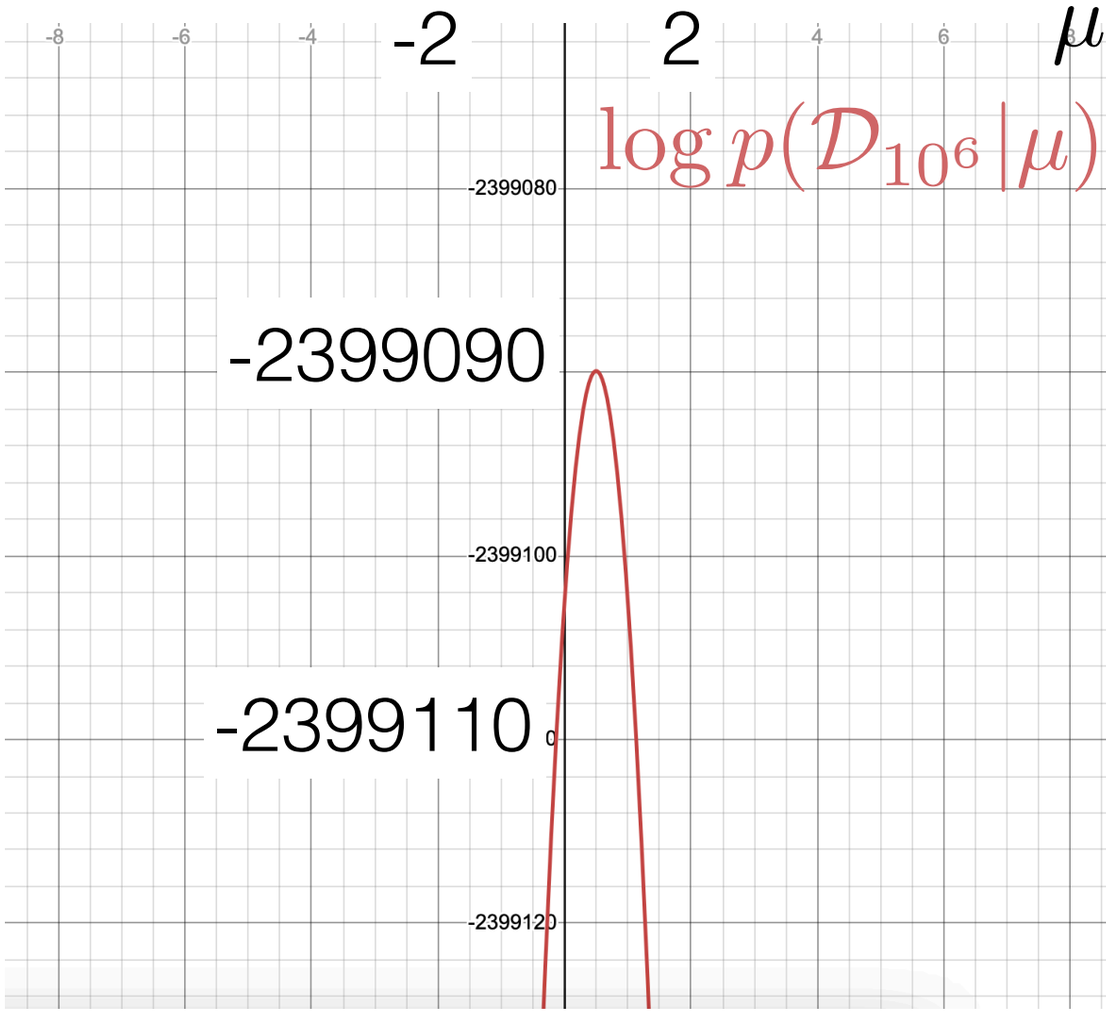
- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2



- One data point; MLE $\hat{\mu} = 0.5$
- $p(\mathcal{D}|\mu)$ within a factor of e of the likelihood at MLE for $\hat{\mu} \pm 100$
- Predictive is $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$

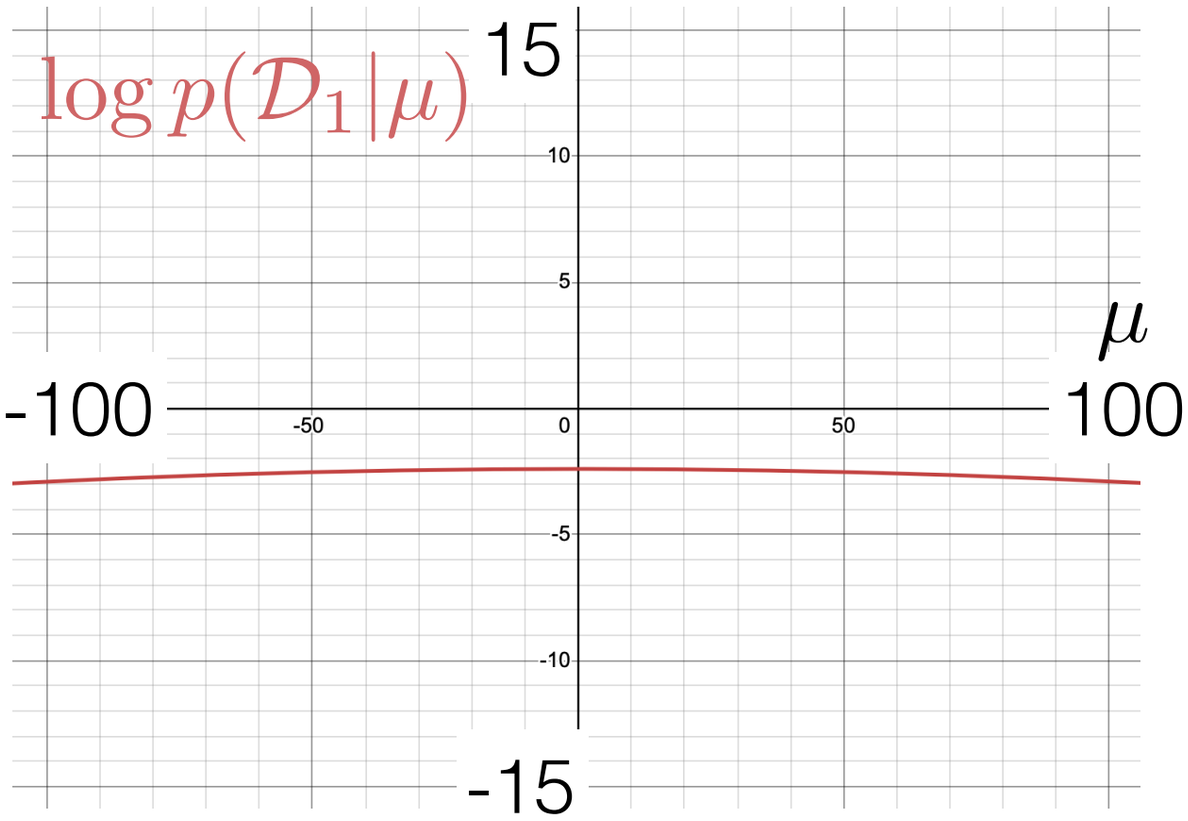
- 10^6 points; MLE $\hat{\mu} = 0.5$
- For μ values over 1 unit different from MLE, $p(\mathcal{D}|\mu)$ is less by a factor of over

$$e^{30} \approx 10^{13}$$



Potential issues with maximum likelihood

- Example: $y^{(n)} \in \mathbb{R}, y^{(n)} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2$ fixed to 100^2

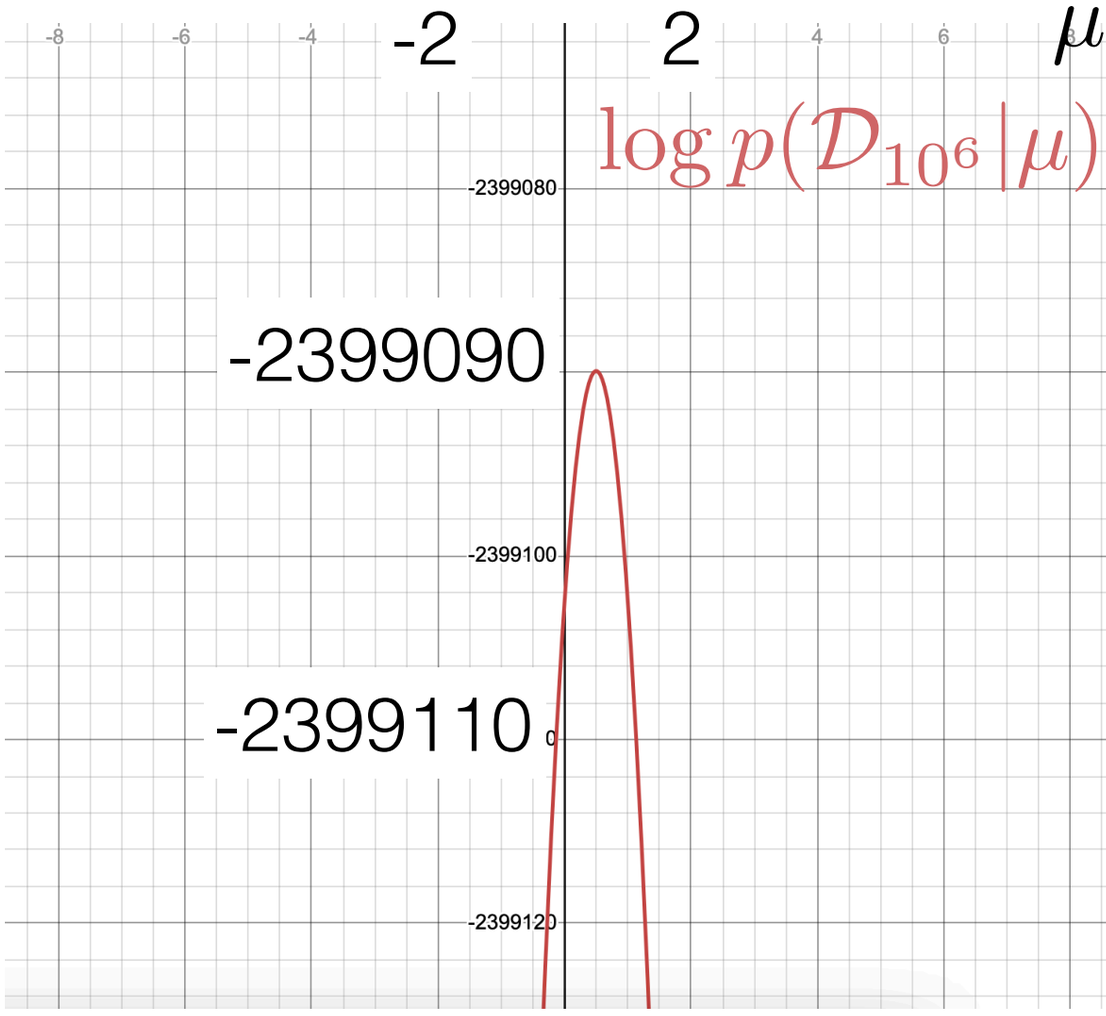


- One data point; MLE $\hat{\mu} = 0.5$
- $p(\mathcal{D}|\mu)$ within a factor of e of the likelihood at MLE for $\hat{\mu} \pm 100$
- Predictive is $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$

- 10^6 points; MLE $\hat{\mu} = 0.5$
- For μ values over 1 unit different from MLE, $p(\mathcal{D}|\mu)$ is less by a factor of over

$$e^{30} \approx 10^{13}$$

- Same predictive in both cases $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$



Potential issues with maximum likelihood

Potential issues with maximum likelihood

- Example: Back to Bernoulli

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

  - If we have a "small" amount of data and haven't seen any 1's yet, $p(1|\hat{\theta}) = \text{Bernoulli}(1|0) = 0$

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

  - If we have a "small" amount of data and haven't seen any 1's yet, $p(1|\hat{\theta}) = \text{Bernoulli}(1|0) = 0$
  - But that predictor seems unlikely to generalize well

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

  - If we have a "small" amount of data and haven't seen any 1's yet, $p(1|\hat{\theta}) = \text{Bernoulli}(1|0) = 0$
  - But that predictor seems unlikely to generalize well

- But don't we always have "big data" these days?

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

  - If we have a "small" amount of data and haven't seen any 1's yet, $p(1|\hat{\theta}) = \text{Bernoulli}(1|0) = 0$
  - But that predictor seems unlikely to generalize well

- But don't we always have "big data" these days?
  - Might have large volume with little info per parameter
  - E.g. in genomics, can have thousands of individuals (data points) and millions of genetic variants (parameters)

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

  - If we have a "small" amount of data and haven't seen any 1's yet, $p(1|\hat{\theta}) = \text{Bernoulli}(1|0) = 0$
  - But that predictor seems unlikely to generalize well

- But don't we always have "big data" these days?
  - Might have large volume with little info per parameter
  - E.g. in genomics, can have thousands of individuals (data points) and millions of genetic variants (parameters)

- We're comparing how likely the data are under various parameter settings (and just choosing a single extreme)

Potential issues with maximum likelihood

- Example: Back to Bernoulli

$$y^{(n)} \in \{0, 1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0, 1]$$

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

  - If we have a "small" amount of data and haven't seen any 1's yet, $p(1|\hat{\theta}) = \text{Bernoulli}(1|0) = 0$
  - But that predictor seems unlikely to generalize well

- But don't we always have "big data" these days?
  - Might have large volume with little info per parameter
  - E.g. in genomics, can have thousands of individuals (data points) and millions of genetic variants (parameters)

- We're comparing how likely the data are under various parameter settings (and just choosing a single extreme)
  - Perhaps what we really want is closer to the following: how likely are various parameter settings given the data?

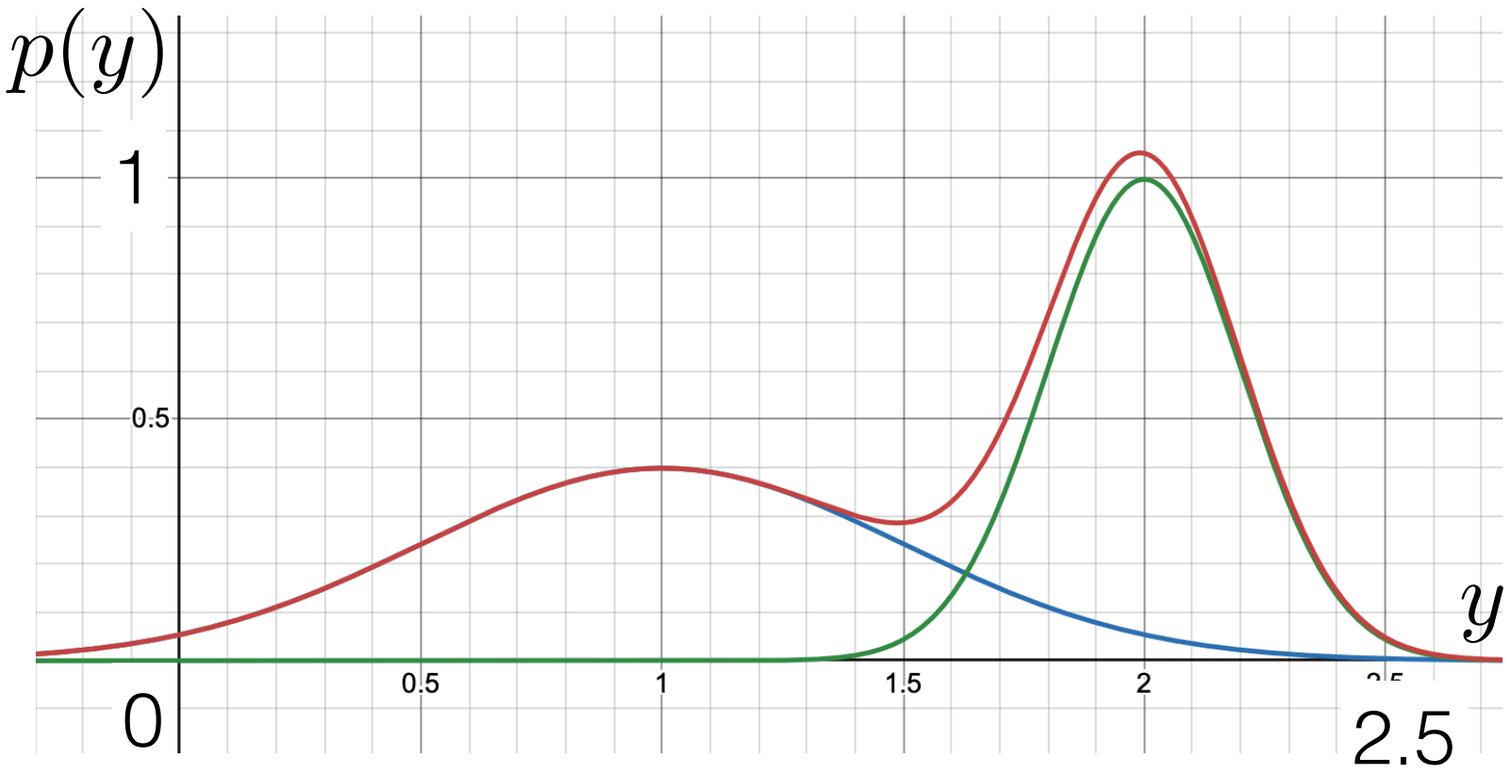
Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

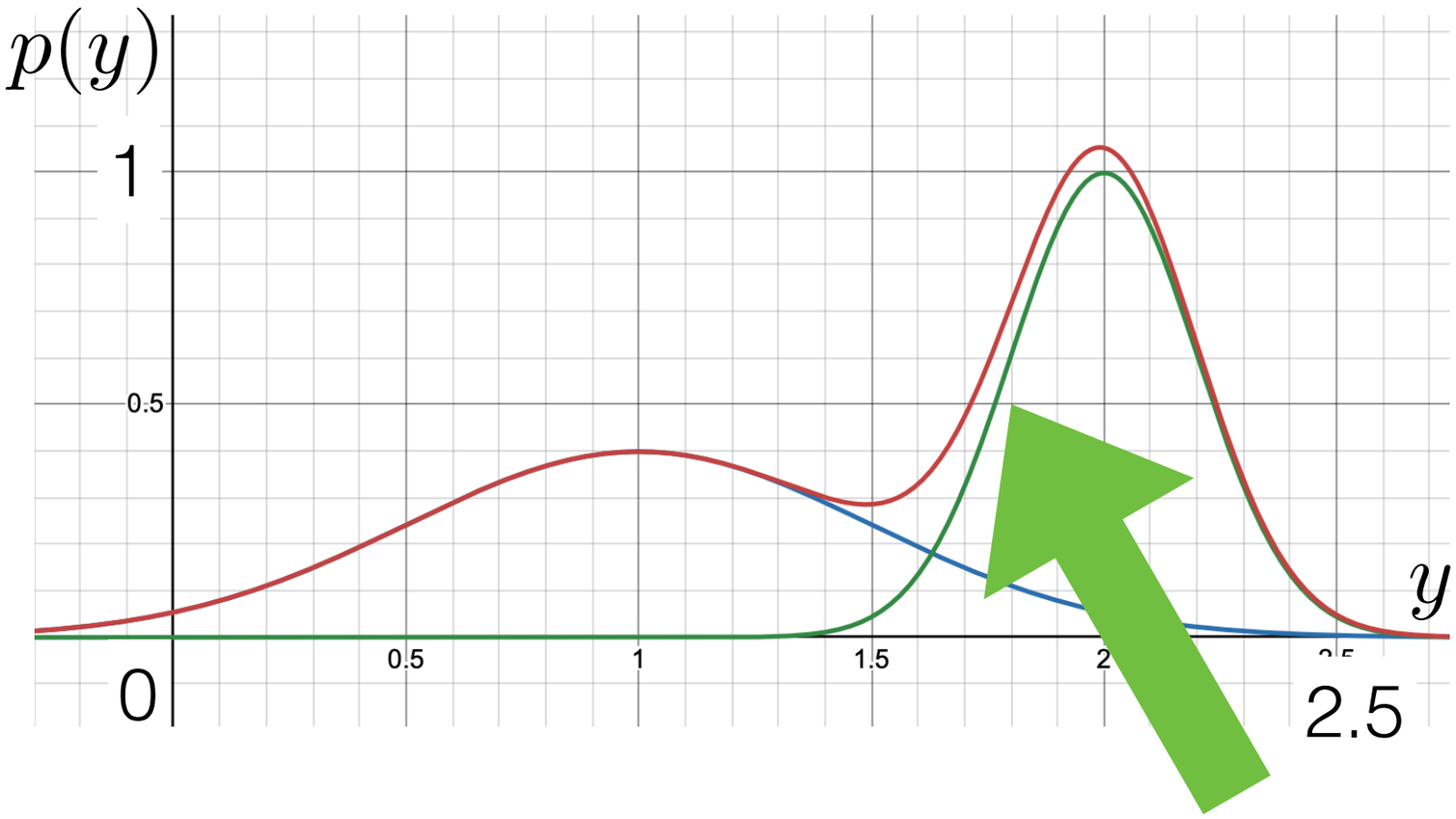
$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

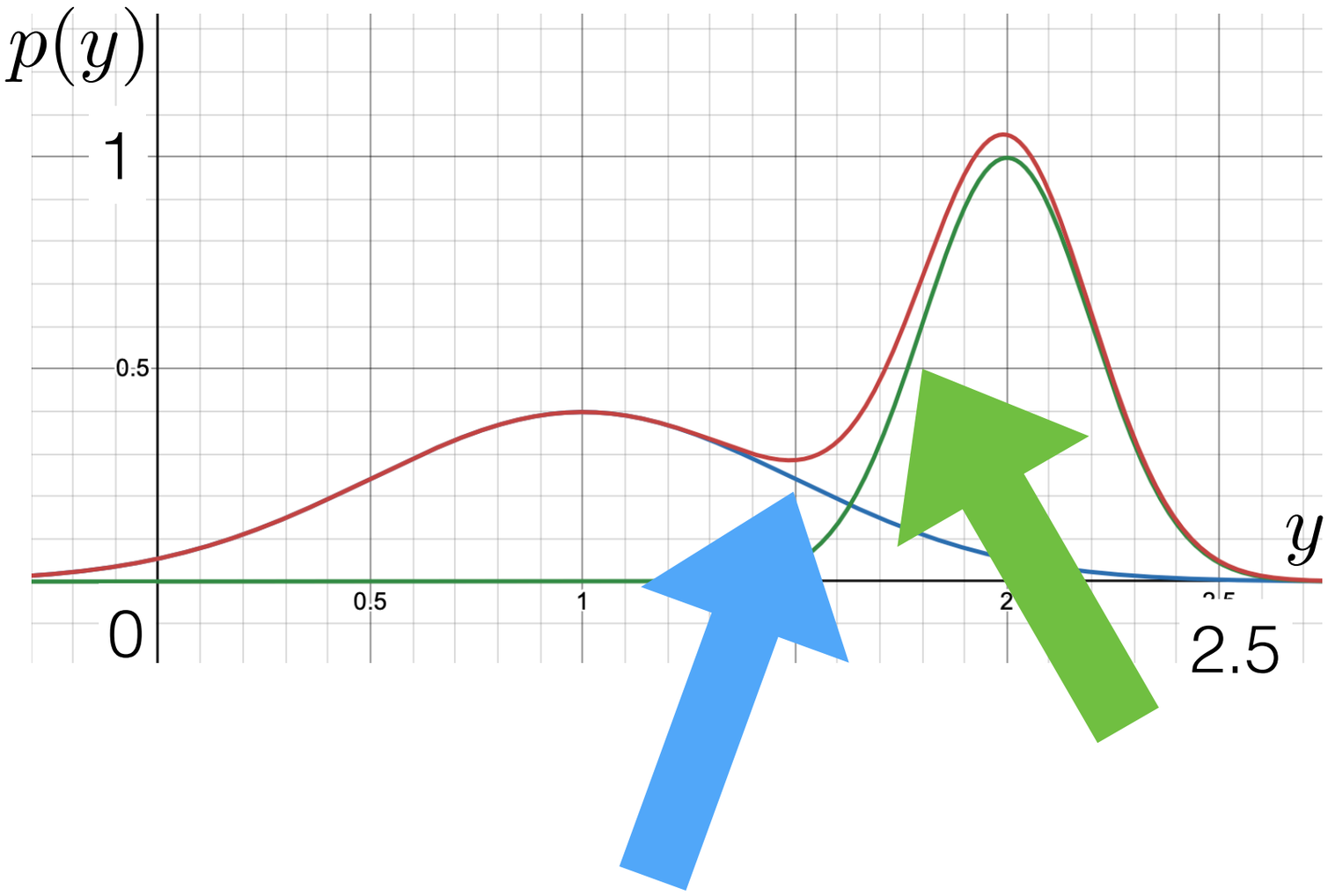
$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

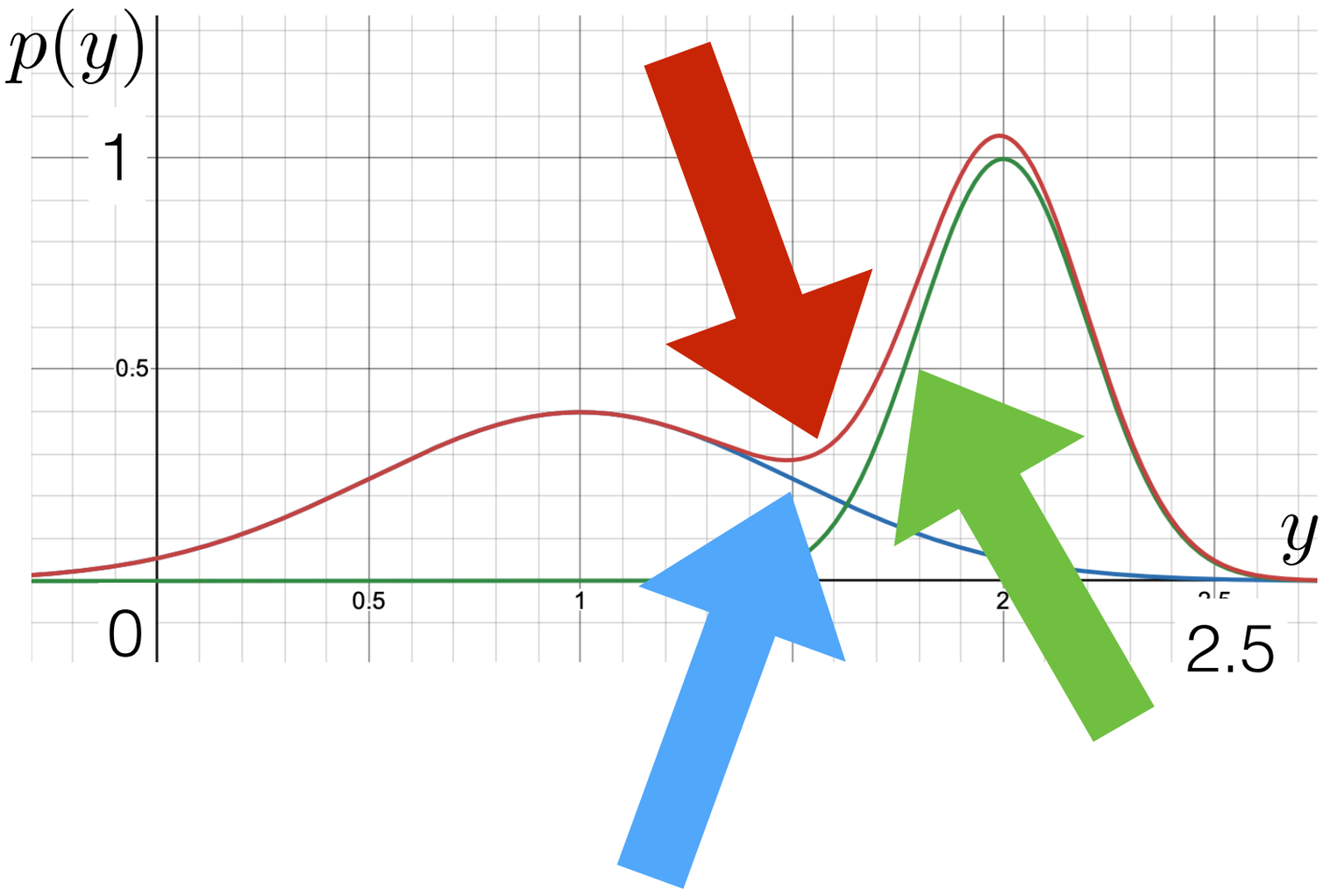
$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

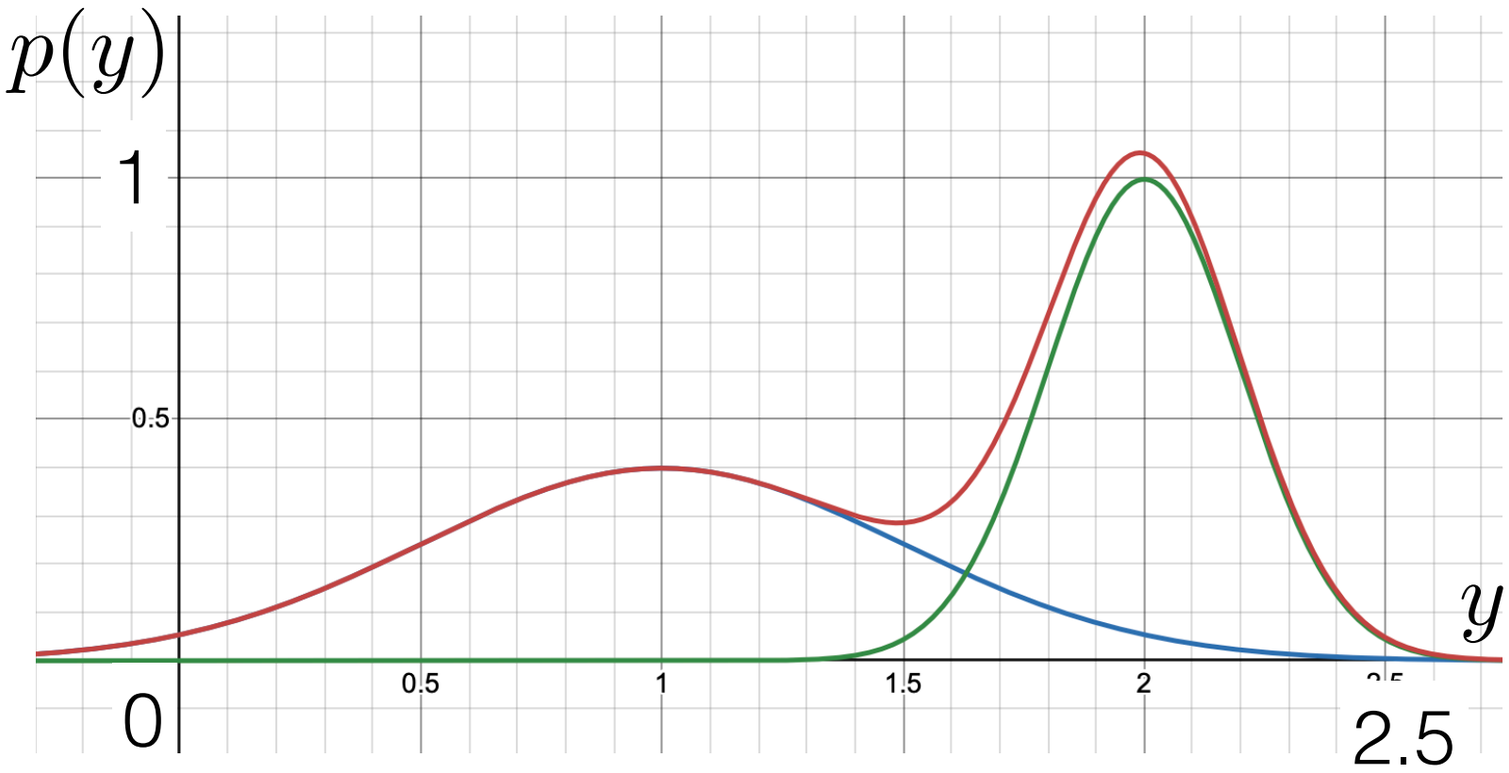
$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

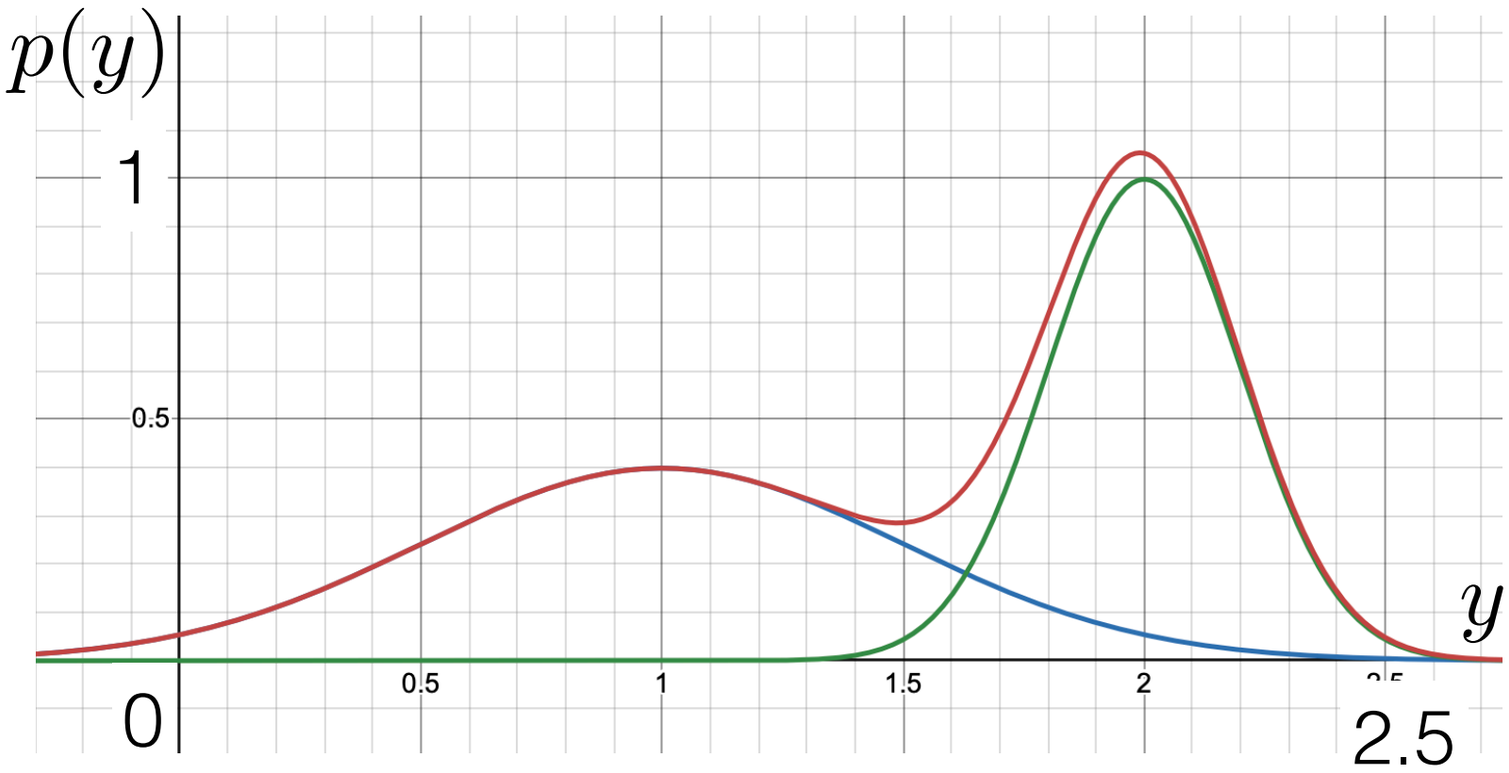
$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

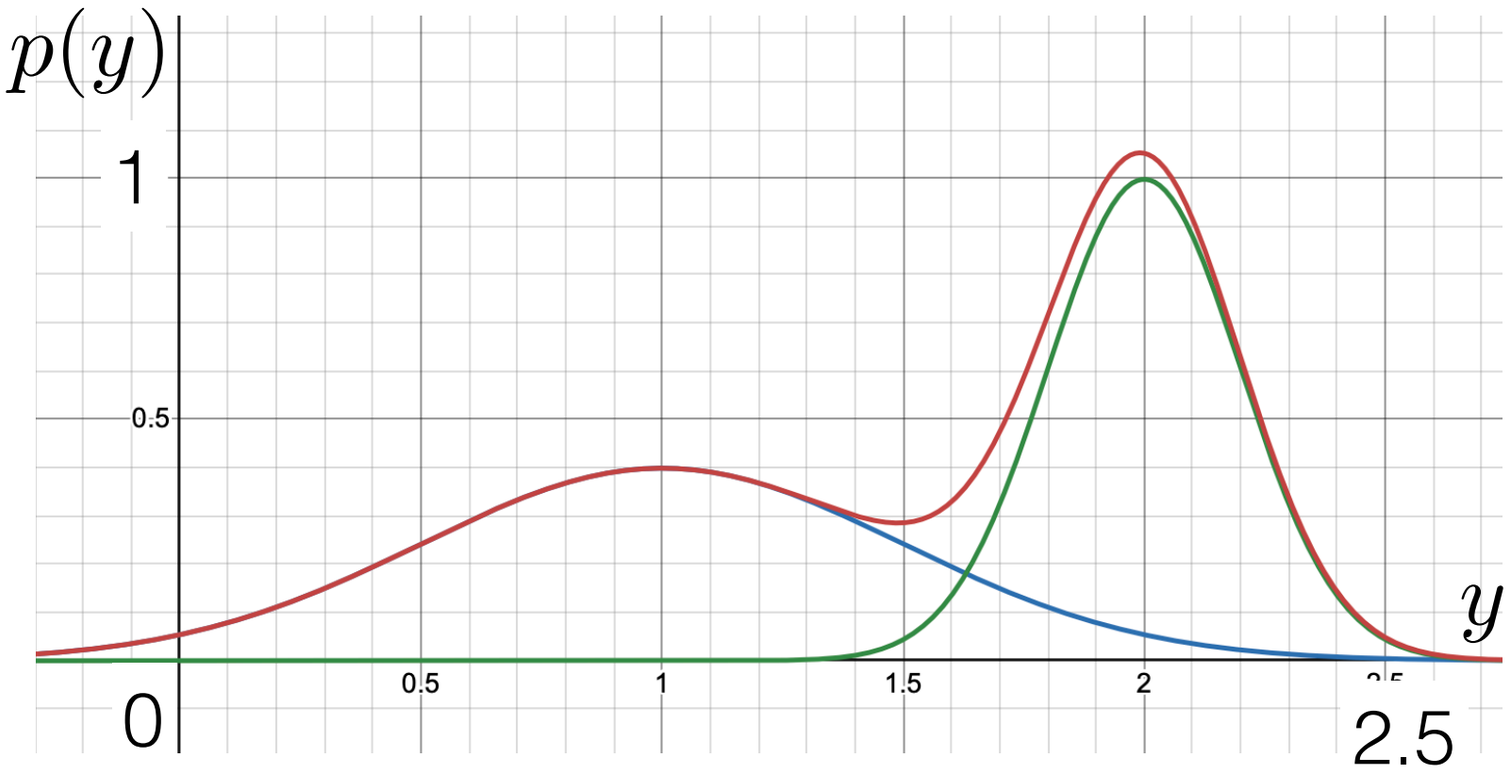


- We'd like to find the MLE for *N* iid data

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

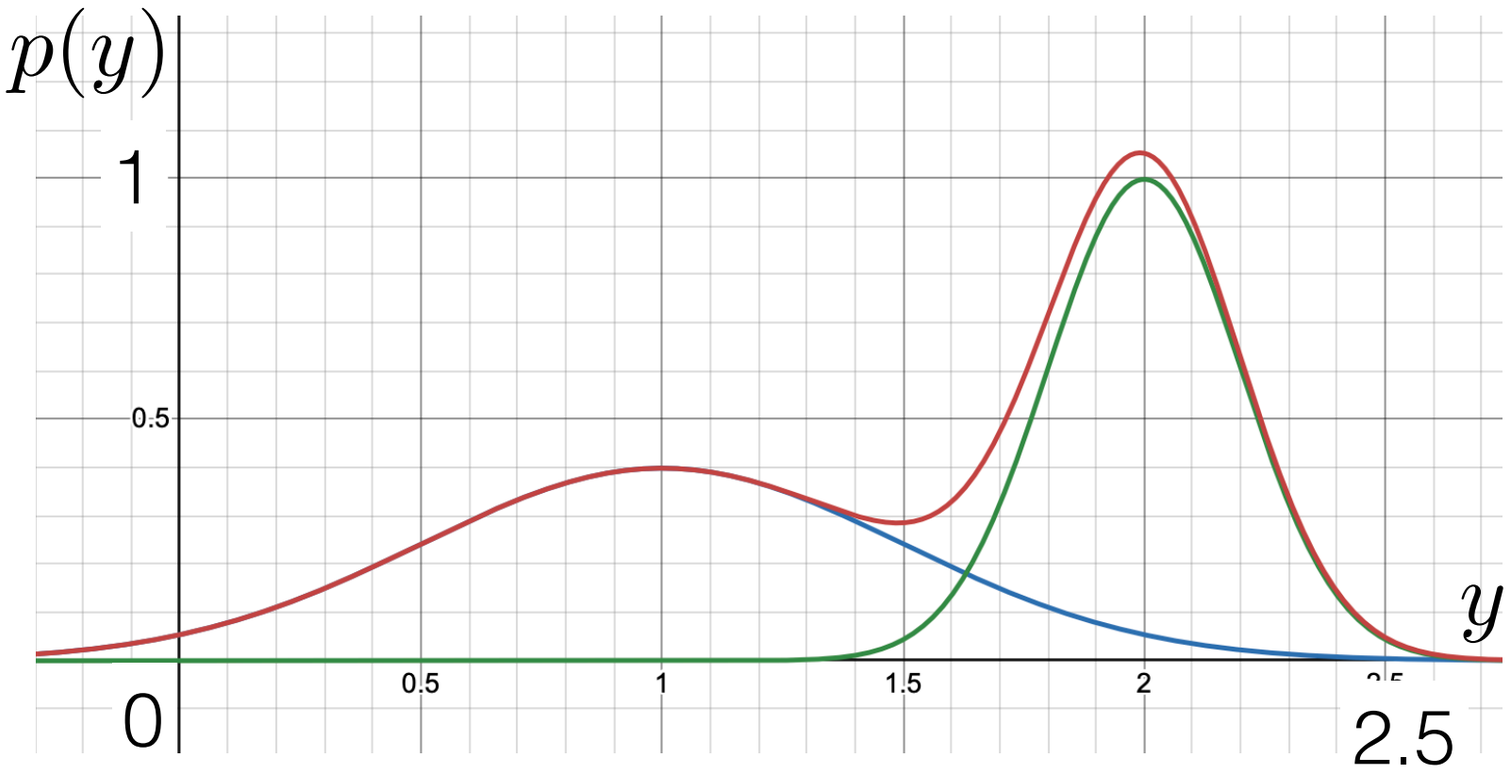


- We'd like to find the MLE for *N* iid data
- If we find the MLE, no other parameter setting should have higher likelihood

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

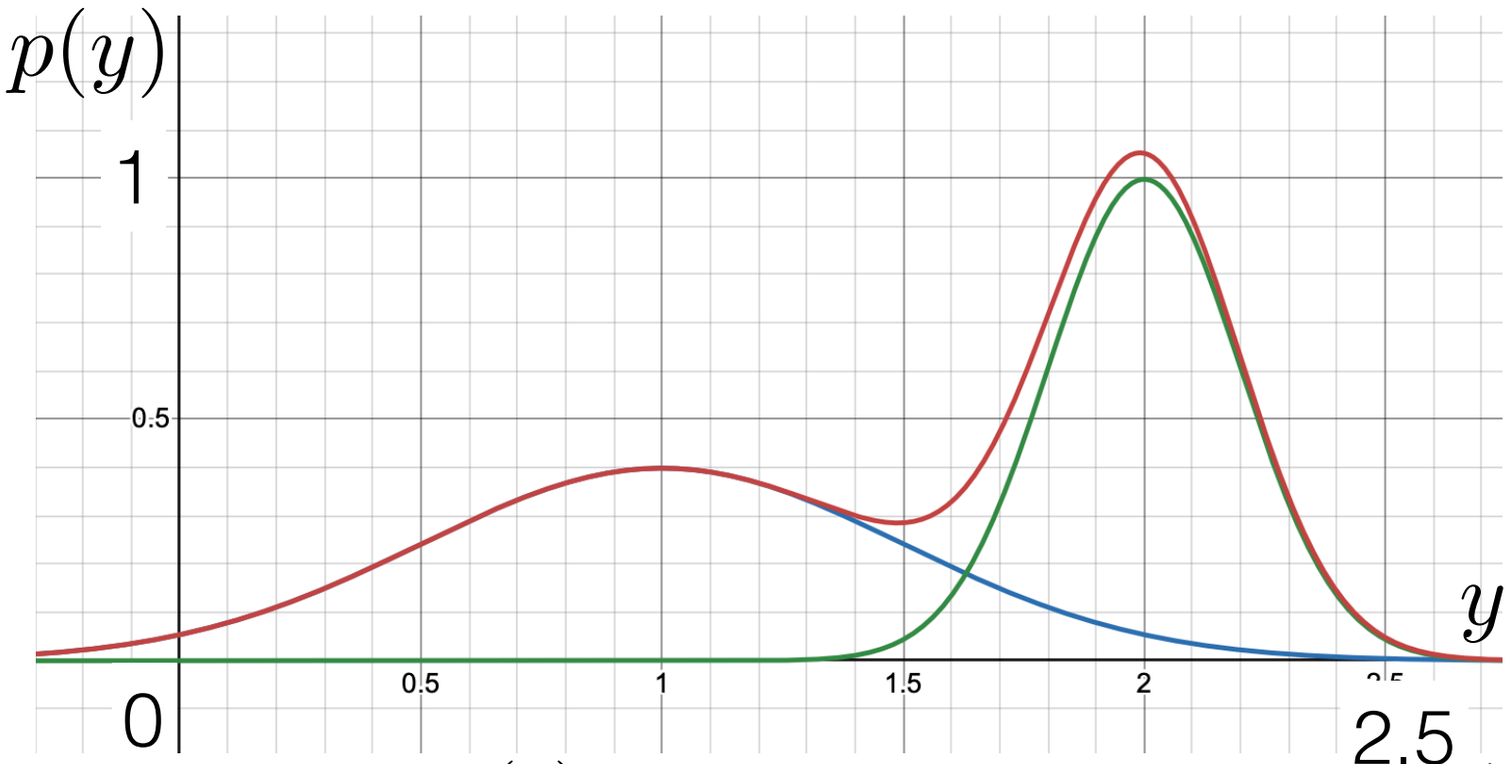


- We'd like to find the MLE for *N* iid data
- If we find the MLE, no other parameter setting should have higher likelihood
- Set $\mu_1^* = y^{(1)}$

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0, 1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



- We'd like to find the MLE for *N* iid data
- If we find the MLE, no other parameter setting should have higher likelihood
- Set $\mu_1^* = y^{(1)}$
- Then $p(y^{(1)}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}$ and

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0,1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

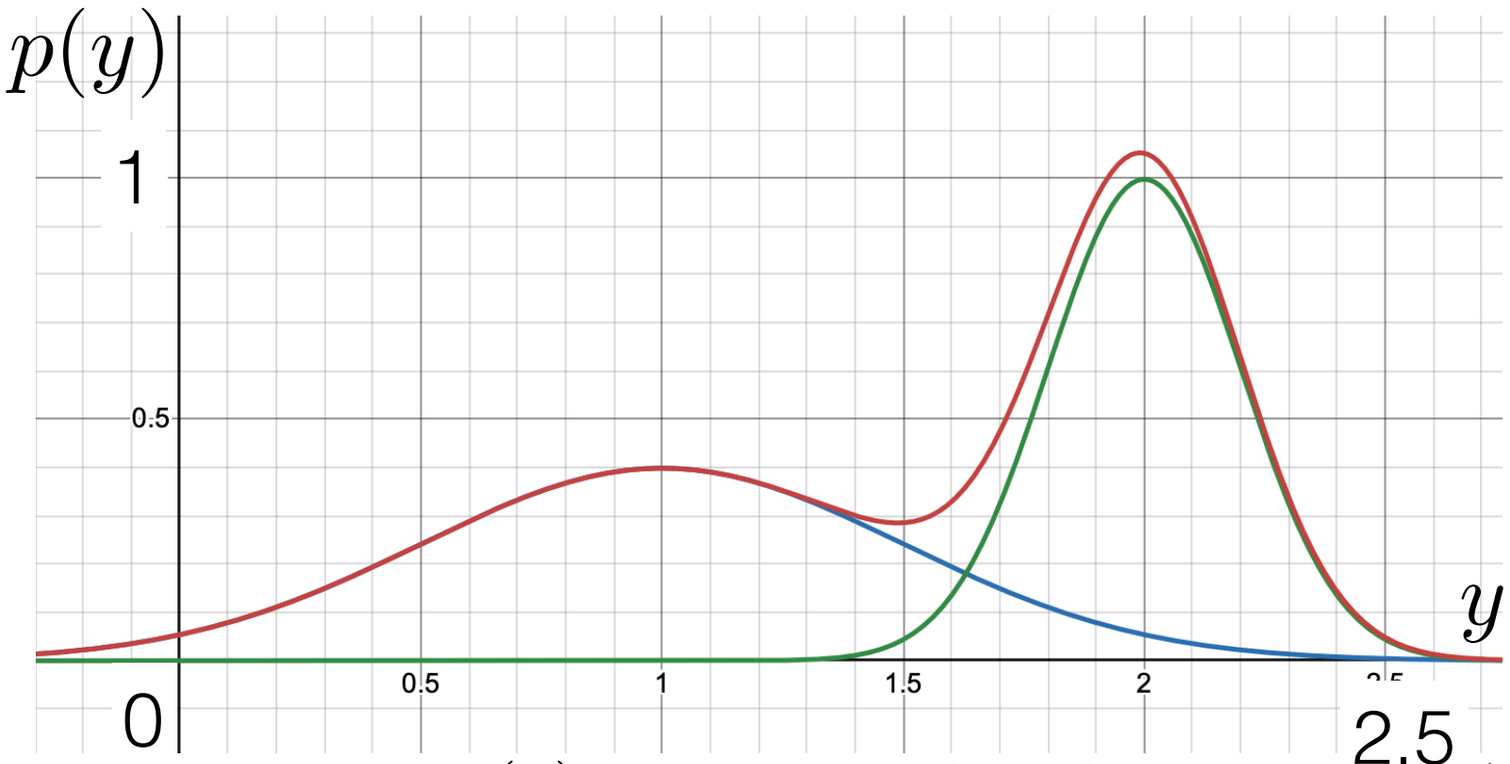
- We'd like to find the MLE for *N* iid data
- If we find the MLE, no other parameter setting should have higher likelihood
- Set $\mu_1^* = y^{(1)}$
- Then $p(y^{(1)}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}$ and

$$p(\mathcal{D}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \left(\pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}\right) \prod_{n=2}^{N} \pi_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y^{(n)}-\mu_2)^2}{2\sigma_2^2}\right)$$

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0,1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$



- We'd like to find the MLE for *N* iid data
- If we find the MLE, no other parameter setting should have higher likelihood
- Set $\mu_1^* = y^{(1)}$
- Then $p(y^{(1)}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}$ and

$$p(\mathcal{D}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \left(\pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}\right) \prod_{n=2}^{N} \pi_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y^{(n)}-\mu_2)^2}{2\sigma_2^2}\right)$$

- So we can make the likelihood arbitrarily large by making the variance σ_1^2 arbitrarily small

Potential issues with maximum likelihood

- Example: mixture of 2 Gaussians: fixed proportions $\pi_k \in (0,1)$

$$y^{(n)} \in \mathbb{R}, p(y|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

- We'd like to find the MLE for *N* iid data
- If we find the MLE, no other parameter setting should have higher likelihood
- Set $\mu_1^* = y^{(1)}$
- Then $p(y^{(1)}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}$ and

$$p(\mathcal{D}|\mu_1^*, \mu_2, \sigma_1^2, \sigma_2^2) \geq \left(\pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}\right) \prod_{n=2}^{N} \pi_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y^{(n)}-\mu_2)^2}{2\sigma_2^2}\right)$$

- So we can make the likelihood arbitrarily large by making the variance σ_1^2 arbitrarily small
- The resulting predictor seems unlikely to generalize well

Bonus: Bernoulli MLE in more detail

- Example: $y^{(n)} \in \{0,1\}, y^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta), \theta \in [0,1]$
- Case A: $\theta \in (0,1)$

  $p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}}(1-\theta)^{1-y^{(n)}}$

  $\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log\theta + (1-y^{(n)})\log(1-\theta)]$

  $\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \theta^{-1}\sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-1}\sum_{n=1}^{N}(1-y^{(n)}) \overset{\text{set}}{=} 0$

  $\frac{d^2 \log p(\mathcal{D}|\theta)}{d\theta^2} = -\theta^{-2}\sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-2}\sum_{n=1}^{N}(1-y^{(n)})$

  - So the optimizer, if it's in (0,1), is $\hat{\theta} = N^{-1}\sum_{n=1}^{N} y^{(n)}$
- Case B: $\theta = 0$
  - Then $p(\mathcal{D}|\theta) =$ 1 if all the data points are 0; else, 0
  - Since 1 is the largest possible likelihood value and 0 is the smallest, $\hat{\theta} = 0$ if and only if the data points are all 0.
  - The $\theta = 1$ case is similar
- Noting that the average of all 0s is 0, we can write:

$$\hat{\theta} = \arg\max_{\theta\in[0,1]} \log p(\mathcal{D}|\theta) = N^{-1}\sum_{n=1}^{N} y^{(n)}$$