Math Camp

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

September 8th, 2015
Multivariate Integration

Suppose we have a function $f : X \rightarrow \mathbb{R}^1$, with $X \subset \mathbb{R}^2$. 

Area under function.

Suppose that area, $A$, is in 2-dimensions: 

- $A = \{(x, y) : x \in [0, 1], y \in [0, 1]\}$
- $A = \{(x, y) : x^2 + y^2 \leq 1\}$
- $A = \{(x, y) : x > y, x, y \in (0, 2)\}$
Suppose we have a function $f : X \to \mathbb{R}^1$, with $X \subset \mathbb{R}^2$.
We will integrate a function over an area.
Multivariate Integration

Suppose we have a function $f : X \rightarrow \mathbb{R}^1$, with $X \subset \mathbb{R}^2$.

We will integrate a function over an area. Area under function.
Multivariate Integration

Suppose we have a function \( f : X \rightarrow \mathbb{R}^1 \), with \( X \subset \mathbb{R}^2 \).
We will integrate a function over an area.  
**Area** under function.  
Suppose that area, \( A \), is in 2-dimensions.
Suppose we have a function $f : X \to \mathbb{R}^1$, with $X \subset \mathbb{R}^2$.
We will integrate a function over an area. Area under function.
Suppose that area, $A$, is in 2-dimensions
- $A = \{x, y : x \in [0, 1], y \in [0, 1]\}$
Multivariate Integration

Suppose we have a function $f : X \to \mathbb{R}^1$, with $X \subset \mathbb{R}^2$.
We will integrate a function over an area.
Area under function.
Suppose that area, $A$, is in 2-dimensions
- $A = \{x, y : x \in [0, 1], y \in [0, 1]\}$
- $A = \{x, y : x^2 + y^2 \leq 1\}$
Suppose we have a function \( f : X \rightarrow \mathbb{R}^1 \), with \( X \subset \mathbb{R}^2 \).
We will integrate a function over an area.
Area under function.
Suppose that area, \( A \), is in 2-dimensions
- \( A = \{ x, y : x \in [0, 1], y \in [0, 1] \} \)
- \( A = \{ x, y : x^2 + y^2 \leq 1 \} \)
- \( A = \{ x, y : x > y, x, y \in (0, 2) \} \)
Suppose we have a function \( f : X \to \mathbb{R}^1 \), with \( X \subset \mathbb{R}^2 \).

We will integrate a function over an area. 

Area under function.

Suppose that area, \( A \), is in 2-dimensions

- \( A = \{ x, y : x \in [0, 1], y \in [0, 1] \} \)
- \( A = \{ x, y : x^2 + y^2 \leq 1 \} \)
- \( A = \{ x, y : x > y, x, y \in (0, 2) \} \)

How do calculate the area under the function over these regions?
Multivariate Integration

Definition

Suppose $f: X \to \mathbb{R}$ where $X \subset \mathbb{R}^n$. We will say that $f$ is integrable over $A \subset X$ if we are able to calculate its area with refined partitions of $A$ and we will write the integral $I = \int_A f(x) dA$. 

That's horribly abstract. There is an extremely helpful theorem that makes this manageable.

Theorem

Fubini's Theorem

Suppose $A = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ and that $f: A \to \mathbb{R}$ is integrable. Then

$$\int_A f(x) dA = \int_{b_n}^{a_n} \int_{b_{n-1}}^{a_{n-1}} \ldots \int_{b_1}^{a_1} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$
Multivariate Integration

Definition

Suppose $f : X \rightarrow \mathbb{R}$ where $X \subset \mathbb{R}^n$. We will say that $f$ is integrable over $A \subset X$ if we are able to calculate its area with refined partitions of $A$ and we will write the integral $I = \int_A f(x) dA$

That's horribly abstract. There is an extremely helpful theorem that makes this manageable.

Theorem

Fubini's Theorem

Suppose $A = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ and that $f : A \rightarrow \mathbb{R}$ is integrable. Then

$\int_A f(x) dA = \int_{b_n}^{b_1} \int_{b_{n-1}}^{b_{n-2}} \ldots \int_{b_2}^{b_1} \int_{b_1}^{b_1} f(x) dx_1 dx_2 \ldots dx_n$
Multivariate Integration

Definition

Suppose $f : X \to \mathbb{R}$ where $X \subset \mathbb{R}^n$. We will say that $f$ is integrable over $A \subset X$ if we are able to calculate its area with refined partitions of $A$ and we will write the integral $I = \int_A f(x) \, dA$

That’s horribly abstract. There is an extremely helpful theorem that makes this manageable.

Theorem

**Fubini’s Theorem** Suppose $A = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ and that $f : A \to \mathbb{R}$ is integrable. Then

$$\int_A f(x) \, dA = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \ldots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x) \, dx_1 \, dx_2 \ldots dx_{n-1} \, dx_n$$
Multivariate Integration Recipe

\[ \int_A f(x) \, dA = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \ldots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x) \, dx_1 \, dx_2 \ldots \, dx_{n-1} \, dx_n \]
Multivariate Integration Recipe

\[
\int_A f(x) \, dA = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x) \, dx_1 \, dx_2 \cdots dx_{n-1} \, dx_n
\]

1) Start with the inside integral \( x_1 \) is the variable, everything else a constant
Multivariate Integration Recipe

\[ \int_A f(x) \, dA = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \ldots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x) \, dx_1 \, dx_2 \ldots \, dx_{n-1} \, dx_n \]

1) Start with the inside integral \( x_1 \) is the variable, everything else a constant
2) Work inside to out, iterating
Multivariate Integration Recipe

\[ \int_A f(x) dA = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x) dx_1 dx_2 \cdots dx_{n-1} dx_n \]

1) Start with the inside integral \( x_1 \) is the variable, everything else a constant

2) Work inside to out, **iterating**

3) At the last step, we should arrive at a number
Intuition: Three Dimensional Jello Molds, a discussion
Multivariate Uniform Distribution

Suppose $f : [0, 1] \times [0, 1] \to \mathbb{R}$ and $f(x_1, x_2) = 1$ for all $x_1, x_2 \in [0, 1] \times [0, 1]$. What is $\int_0^1 \int_0^1 f(x) \, dx_1 \, dx_2$?

\[
\int_0^1 \int_0^1 f(x) \, dx_1 \, dx_2 = \int_0^1 \int_0^1 1 \, dx_1 \, dx_2 \\
= \int_0^1 x_1 |_0^1 \, dx_2 \\
= \int_0^1 (1 - 0) \, dx_2 \\
= \int_0^1 1 \, dx_2 \\
= x_2 |_0^1 \\
= 1
\]
Example 2

Suppose \( f: [a_1, b_1] \times [a_2, b_2] \rightarrow \mathbb{R} \) is given by

\[
f(x_1, x_2) = x_1 x_2
\]
Example 2
Suppose \( f : [a_1, b_1] \times [a_2, b_2] \rightarrow \mathbb{R} \) is given by

\[
f(x_1, x_2) = x_1 x_2
\]

Find \( \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 \)
Example 2

Suppose $f : [a_1, b_1] \times [a_2, b_2] \rightarrow \mathbb{R}$ is given by

$$f(x_1, x_2) = x_1 x_2$$

Find $\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2$

$$\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 = \int_{a_2}^{b_2} \int_{a_1}^{b_1} x_2 x_1 \, dx_1 \, dx_2$$
Example 2
Suppose $f : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}$ is given by

$$f(x_1, x_2) = x_1 x_2$$

Find $\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2$

$$\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 = \int_{a_2}^{b_2} \int_{a_1}^{b_1} x_2 x_1 \, dx_1 \, dx_2$$

$$= \int_{a_2}^{b_2} \frac{x_1^2}{2} x_2 \bigg|_{a_1}^{b_1} \, dx_2$$
Example 2

Suppose \( f : [a_1, b_1] \times [a_2, b_2] \rightarrow \mathbb{R} \) is given by

\[
f(x_1, x_2) = x_1 x_2
\]

Find \( \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 \)

\[
\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 = \int_{a_2}^{b_2} \int_{a_1}^{b_1} x_2 x_1 \, dx_1 \, dx_2
\]

\[
= \int_{a_2}^{b_2} \frac{x_1^2}{2} \bigg|_{a_1}^{b_1} \, dx_2
\]

\[
= \frac{b_2^2 - a_2^2}{2} \int_{a_2}^{b_2} x_2 \, dx_2
\]
Example 2

Suppose $f : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}$ is given by

$$f(x_1, x_2) = x_1 x_2$$

Find $\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2$

$$\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 = \int_{a_2}^{b_2} \int_{a_1}^{b_1} x_2 x_1 \, dx_1 \, dx_2$$

$$= \int_{a_2}^{b_2} \frac{x_1^2}{2} x_2 \bigg|_{a_1}^{b_1} \, dx_2$$

$$= \frac{b_1^2 - a_1^2}{2} \int_{a_2}^{b_2} x_2 \, dx_2$$

$$= \frac{b_1^2 - a_1^2}{2} \left( \frac{x_2^2}{2} \bigg|_{a_2}^{b_2} \right)$$
Example 2
Suppose \( f : [a_1, b_1] \times [a_2, b_2] \to \Re \) is given by

\[
f(x_1, x_2) = x_1 x_2
\]

Find \( \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 \)

\[
\int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) \, dx_1 \, dx_2 = \int_{a_2}^{b_2} \int_{a_1}^{b_1} x_2 x_1 \, dx_1 \, dx_2
\]

\[
= \int_{a_2}^{b_2} \frac{x_1^2}{2} \bigg|_{a_1}^{b_1} x_2 \, dx_2
\]

\[
= \frac{b_1^2 - a_1^2}{2} \int_{a_2}^{b_2} x_2 \, dx_2
\]

\[
= \frac{b_1^2 - a_1^2}{2} \left( \frac{x_2^2}{2} \bigg|_{a_2}^{b_2} \right)
\]

\[
= \frac{b_1^2 - a_1^2}{2} \frac{b_2^2 - a_2^2}{2}
\]
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}^2_+ \rightarrow \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)\]
Example 3: Exponential Distributions

Suppose $f : \mathbb{R}^2_+ \to \mathbb{R}$ and that

$$f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)$$

Find:
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}^2_+ \to \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)
\]

Find:

\[
\int_0^\infty \int_0^\infty f(x_1, x_2) = \\
= \\
= \\
= \\
= \\
=
\]

\[
2 \left[ \lim_{x_1 \to \infty} \exp(-x_1) + 1 \right] \left[ \lim_{x_2 \to \infty} \frac{1}{2} \exp(-2x_2) + 1 \right] = 1
\]
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}_+^2 \rightarrow \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)
\]

Find:

\[
\int_0^\infty \int_0^\infty f(x_1, x_2) = 2 \int_0^\infty \int_0^\infty \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2
\]
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}_+^2 \to \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)
\]

Find:

\[
\int_0^\infty \int_0^\infty f(x_1, x_2) = 2 \int_0^\infty \int_0^\infty \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2
\]

\[
= 2 \int_0^\infty \exp(-x_1) \, dx_1 \int_0^\infty \exp(-2x_2) \, dx_2
\]

\[
= \left[ -\exp(-x_1) \right]_0^\infty \left[ -\frac{1}{2} \exp(-2x_2) \right]_0^\infty
\]

\[
= 1
\]
Example 3: Exponential Distributions

Suppose $f : \mathbb{R}^2_+ \rightarrow \mathbb{R}$ and that

$$f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)$$

Find:

$$\int_0^{\infty} \int_0^{\infty} f(x_1, x_2) = 2 \int_0^{\infty} \int_0^{\infty} \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2$$

$$= 2 \int_0^{\infty} \exp(-x_1) \, dx_1 \int_0^{\infty} \exp(-2x_2) \, dx_2$$

$$= 2(-\exp(-x)|_0^{\infty})(-\frac{1}{2} \exp(-2x_2)|_0^{\infty})$$

$$=$$

$$=$$

$$= $$
Example 3: Exponential Distributions

Suppose $f : \mathbb{R}^2_+ \rightarrow \mathbb{R}$ and that

$$f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)$$

Find:

$$\int_0^\infty \int_0^\infty f(x_1, x_2) \, dx_1 \, dx_2 = 2 \int_0^\infty \int_0^\infty \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2$$

$$= 2 \int_0^\infty \exp(-x_1) \, dx_1 \int_0^\infty \exp(-2x_2) \, dx_2$$

$$= 2(- \exp(-x)|_0^\infty)(- \frac{1}{2} \exp(-2x_2)|_0^\infty)$$

$$= 2 \left[ (- \lim_{x_1 \to \infty} \exp(-x_1) + 1)(- \frac{1}{2} \lim_{x_2 \to \infty} \exp(-2x_2) + \frac{1}{2}) \right]$$

$$= 1$$
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}^2_+ \to \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)
\]

Find:

\[
\int_0^\infty \int_0^\infty f(x_1, x_2) = 2 \int_0^\infty \int_0^\infty \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2
\]

\[
= 2 \int_0^\infty \exp(-x_1) \, dx_1 \int_0^\infty \exp(-2x_2) \, dx_2
\]

\[
= 2(- \exp(-x)|_0^\infty)(-\frac{1}{2} \exp(-2x_2)|_0^\infty)
\]

\[
= 2 \left[ ( - \lim_{x_1 \to \infty} \exp(-x_1) + 1)(-\frac{1}{2} \lim_{x_2 \to \infty} \exp(-2x_2) + \frac{1}{2}) \right]
\]

\[
= 2[\frac{1}{2}]
\]

\[
= 1
\]
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}^2_+ \rightarrow \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)
\]

Find:

\[
\int_0^\infty \int_0^\infty f(x_1, x_2) \, dx_1 \, dx_2 = 2 \int_0^\infty \int_0^\infty \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2
\]

\[
= 2 \int_0^\infty \exp(-x_1) \, dx_1 \int_0^\infty \exp(-2x_2) \, dx_2
\]

\[
= 2 \left( \lim_{x_1 \to \infty} \exp(-x) \bigg|_0^\infty \right) \left( \frac{1}{2} \lim_{x_2 \to \infty} \exp(-2x_2) \bigg|_0^\infty \right)
\]

\[
= 2 \left[ (\lim_{x_1 \to \infty} \exp(-x_1) + 1) \left( \frac{1}{2} \lim_{x_2 \to \infty} \exp(-2x_2) + \frac{1}{2} \right) \right]
\]

\[
= 2 \left[ \frac{1}{2} \right]
\]

\[
= 1
\]
Example 3: Exponential Distributions

Suppose \( f : \mathbb{R}^2_+ \rightarrow \mathbb{R} \) and that

\[
f(x_1, x_2) = 2 \exp(-x_1) \exp(-2x_2)
\]

Find:

\[
\int_0^\infty \int_0^\infty f(x_1, x_2) = 2 \int_0^\infty \int_0^\infty \exp(-x_1) \exp(-2x_2) \, dx_1 \, dx_2
\]

\[
= 2 \int_0^\infty \exp(-x_1) \, dx_1 \int_0^\infty \exp(-2x_2) \, dx_2
\]

\[
= 2 \left[ - \exp(-x) \big|_0^\infty \right] \left( - \frac{1}{2} \exp(-2x_2) \big|_0^\infty \right)
\]

\[
= 2 \left[ \left( - \lim_{x_1 \rightarrow \infty} \exp(-x_1) + 1 \right) \left( - \frac{1}{2} \lim_{x_2 \rightarrow \infty} \exp(-2x_2) + \frac{1}{2} \right) \right]
\]

\[
= 2 \left[ \frac{1}{2} \right]
\]

\[
= 1
\]
Challenge Problems

1) Find $\int_0^1 \int_0^1 x_1 + x_2 \, dx_1 \, dx_2$

2) Demonstrate that

$$\int_0^b \int_0^a x_1 - 3x_2 \, dx_1 \, dx_2 = \int_0^a \int_0^b x_1 - 3x_2 \, dx_2 \, dx_1$$
More Complicated Bounds of Integration

So far, we have integrated over rectangles. But often, we are interested in more complicated regions.

How do we do this?
More Complicated Bounds of Integration

So far, we have integrated over rectangles. But often, we are interested in more complicated regions.

How do we do this?
Example 4: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, $f(x_1, x_2) = x_1 + x_2$. Find area of function where $x_1 < x_2$. 

Trick: we need to determine bound. If $x_1 < x_2$, $x_1$ can take on any value 

$$\int \int_{x_1 < x_2} f(x) \, dx_1 \, dx_2 = \int_0^1 \int_0^{x_2} x_1 + x_2 \, dx_1 \, dx_2 = \int_0^1 x_2^2 \, dx_2 + \int_0^1 x_2^2 \, dx_2 = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$
Example 4: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \to \mathbb{R}$, $f(x_1, x_2) = x_1 + x_2$. Find area of function where $x_1 < x_2$.

**Trick:** we need to determine bound. If $x_1 < x_2$, $x_1$ can take on any value from 0 to $x_2$.
Example 4: More Complicated Regions

Suppose \( f : [0, 1] \times [0, 1] \rightarrow \mathbb{R}, f(x_1, x_2) = x_1 + x_2 \). Find area of function where \( x_1 < x_2 \).

**Trick:** we need to determine bound. If \( x_1 < x_2 \), \( x_1 \) can take on any value from 0 to \( x_2 \)

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_0^{x_2} (x_1 + x_2) \, dx_1 \, dx_2
\]
Example 4: More Complicated Regions

Suppose \( f : [0, 1] \times [0, 1] \to \mathbb{R}, \ f(x_1, x_2) = x_1 + x_2 \). Find area of function where \( x_1 < x_2 \).

**Trick:** we need to determine bound. If \( x_1 < x_2 \), \( x_1 \) can take on any value from 0 to \( x_2 \).

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_0^{x_2} (x_1 + x_2) \, dx_1 \, dx_2
\]

\[
= \int_0^1 x_2 \int_0^{x_2} x_1 \, dx_1 \, dx_2 + \int_0^1 x_2 \int_0^{x_2} \frac{x_1^2}{2} \, dx_1 \, dx_2
\]

\[
= \left[ \frac{x_1^2}{2} \right]_0^{x_2} \int_0^1 x_2 \, dx_2 + \left[ \frac{x_1^2}{4} \right]_0^{x_2} \int_0^1 x_2 \, dx_2
\]

\[
= \frac{1}{2} x_2^3 + \frac{1}{4} x_2^3 = \frac{3}{4} x_2^3
\]
Example 4: More Complicated Regions

Suppose \( f : [0, 1] \times [0, 1] \to \mathbb{R}, f(x_1, x_2) = x_1 + x_2 \). Find area of function where \( x_1 < x_2 \).

**Trick:** we need to determine bound. If \( x_1 < x_2 \), \( x_1 \) can take on any value from 0 to \( x_2 \)

\[
\begin{align*}
\int\int_{x_1 < x_2} f(x) &= \int_0^1 \int_0^{x_2} x_1 + x_2 \, dx_1 \, dx_2 \\
&= \int_0^1 x_2 x_1 \bigg|_0^{x_2} \, dx_2 + \int_0^1 \frac{x_1^2}{2} \bigg|_0^{x_2} \, dx_2 \\
&= \int_0^1 x_2^2 \, dx_2 + \int_0^1 \frac{x_2^2}{2} \end{align*}
\]
Example 4: More Complicated Regions

Suppose \( f : [0, 1] \times [0, 1] \to \mathbb{R} \), \( f(x_1, x_2) = x_1 + x_2 \). Find area of function where \( x_1 < x_2 \).

**Trick:** we need to determine bound. If \( x_1 < x_2 \), \( x_1 \) can take on any value from 0 to \( x_2 \).

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_0^{x_2} x_1 + x_2 \, dx_1 \, dx_2
\]

\[
= \int_0^1 x_2 x_1 \bigg|_0^{x_2} \, dx_2 + \int_0^1 \frac{x_1^2}{2} \bigg|_0^{x_2} \, dx_2
\]

\[
= \int_0^1 x_2^2 \, dx_2 + \int_0^1 \frac{x_2^2}{2} \, dx_2
\]

\[
= \frac{x_2^3}{3} \bigg|_0^1 + \frac{x_2^3}{6} \bigg|_0^1
\]

\[
= 1 + \frac{1}{2} = \frac{3}{2}
\]
Example 4: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, $f(x_1, x_2) = x_1 + x_2$. Find area of function where $x_1 < x_2$.

Trick: we need to determine bound. If $x_1 < x_2$, $x_1$ can take on any value from 0 to $x_2$

\[
\begin{align*}
\int \int_{x_1 < x_2} f(x) &= \int_{0}^{1} \int_{0}^{x_2} x_1 + x_2 \, dx_1 \, dx_2 \\
&= \int_{0}^{1} x_2 x_1 \bigg|_{0}^{x_2} \, dx_2 + \int_{0}^{1} \frac{x_1^2}{2} \bigg|_{0}^{x_2} \, dx_2 \\
&= \int_{0}^{1} x_2^2 \, dx_2 + \int_{0}^{1} \frac{x_2^2}{2} \, dx_2 \\
&= \frac{x_2^3}{3} \bigg|_{0}^{1} + \frac{x_2^3}{6} \bigg|_{0}^{1} \\
&= \frac{1}{3} + \frac{1}{6} \\
&= \frac{1}{3} + \frac{1}{6} 
\end{align*}
\]
Example 4: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, $f(x_1, x_2) = x_1 + x_2$. Find area of function where $x_1 < x_2$.

Trick: we need to determine bound. If $x_1 < x_2$, $x_1$ can take on any value from 0 to $x_2$

$$\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_0^{x_2} x_1 + x_2 \, dx_1 \, dx_2$$

$$= \int_0^1 x_2 x_1 \bigg|_0^{x_2} \, dx_2 + \int_0^1 \frac{x_1^2}{2} \bigg|_0^{x_2} \, dx_2$$

$$= \int_0^1 x_2^2 \, dx_2 + \int_0^1 \frac{x_2^2}{2} \, dx_2$$

$$= \frac{x_2^3}{3} \bigg|_0^1 + \frac{x_2^3}{6} \bigg|_0^1$$

$$= \frac{1}{3} + \frac{1}{6}$$

$$= \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{2}$$
Consider the same function and let’s switch the bounds.
Consider the same function and let’s switch the bounds.

\[ \int\int_{x_1<x_2} f(x) = \int_0^1 \int_{x_1}^1 x_1 + x_2 \, dx_2 \, dx_1 \]
Consider the same function and let’s switch the bounds.

\[
\iiint_{x_1<x_2} f(x) = \int_0^1 \int_{x_1}^1 x_1 + x_2 \, dx_2 \, dx_1 \\
= \int_0^1 x_1 x_2 \big|_{x_1}^1 + \int_0^1 \frac{x_2^2}{2} \big|_{x_1}^1 \, dx_1
\]
Consider the same function and let’s switch the bounds.

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_{x_1}^1 x_1 + x_2 \, dx_2 \, dx_1
\]

\[
= \int_0^1 x_1 x_2 \bigg|_{x_1}^1 + \int_0^1 \frac{x_2^2}{2} \bigg|_{x_1}^1 \, dx_1
\]

\[
= \int_0^1 x_1 - x_1^2 + \int_0^1 \frac{1}{2} - \frac{x_1^2}{2} \, dx_1
\]
Consider the same function and let’s switch the bounds.

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_{x_1}^1 x_1 + x_2 \, dx_2 \, dx_1
\]

\[
= \int_0^1 x_1 x_2 \bigg|_{x_1}^1 + \int_0^1 \frac{x_2^2}{2} \bigg|_{x_1}^1 \, dx_1
\]

\[
= \int_0^1 x_1 - x_1^2 + \int_0^1 \frac{1}{2} - \frac{x_1^2}{2} \, dx_1
\]

\[
= \frac{x_1^2}{2} \bigg|_0^1 - \frac{x_1^3}{3} \bigg|_0^1 + \frac{x_1}{2} \bigg|_0^1 - \frac{x_1^3}{6} \bigg|_0^1
\]
Consider the same function and let’s switch the bounds.

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_{x_1}^1 x_1 + x_2 \, dx_2 \, dx_1
\]

\[
= \int_0^1 x_1 x_2 \bigg|_{x_1}^1 \, dx_1 + \int_0^1 \frac{x_2^2}{2} \bigg|_{x_1}^1 \, dx_1
\]

\[
= \int_0^1 x_1 - x_1^2 + \int_0^1 \frac{1}{2} - \frac{x_1^2}{2} \, dx_1
\]

\[
= \frac{x_1^2}{2} \bigg|_0^1 - \frac{x_1^3}{3} \bigg|_0^1 + \frac{x_1}{2} \bigg|_0^1 - \frac{x_1^3}{6} \bigg|_0^1
\]

\[
= \frac{1}{2} - \frac{1}{3} + \frac{1}{2} - \frac{1}{6}
\]
Consider the same function and let’s switch the bounds.

\[
\int \int_{x_1 < x_2} f(x) = \int_0^1 \int_{x_1}^1 x_1 + x_2 \, dx_2 \, dx_1
\]

\[
= \int_0^1 x_1 x_2 \bigg|_{x_1}^1 + \int_0^1 \frac{x_2^2}{2} \bigg|_1^{x_1} \, dx_1
\]

\[
= \int_0^1 x_1 - x_1^2 + \int_0^1 \frac{1}{2} - \frac{x_1^2}{2} \, dx_1
\]

\[
= \frac{x_1^2}{2} \bigg|_0^1 - \frac{x_1^3}{3} \bigg|_0^1 + \frac{x_1^3}{2} \bigg|_0^1 - \frac{x_1^3}{6} \bigg|_0^1
\]

\[
= \frac{1}{2} - \frac{1}{3} + \frac{1}{2} - \frac{1}{6}
\]

\[
= 1 - \frac{3}{6}
\]

\[
= \frac{1}{2}
\]
Example 5: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \to \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$?
Example 5: More Complicated Regions

Suppose $f([0, 1] \times [0, 1]) \rightarrow \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$? Where is $x_1 + x_2 < 1$?
Example 5: More Complicated Regions

Suppose $f[0, 1] \times [0, 1] \rightarrow \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$? Where is $x_1 + x_2 < 1$? Where, $x_1 < 1 - x_2$
Example 5: More Complicated Regions

Suppose $f[0, 1] \times [0, 1] \to \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$? Where is $x_1 + x_2 < 1$? Where, $x_1 < 1 - x_2$

$$\iint_{x_1 + x_2 < 1} f(x) \, dx = \int_0^1 \int_0^{1-x_2} 1 \, dx_1 \, x_2$$
Example 5: More Complicated Regions

Suppose \( f[0, 1] \times [0, 1] \rightarrow \mathbb{R}, f(x_1, x_2) = 1 \). What is the area of \( x_1 + x_2 < 1 \)? Where is \( x_1 + x_2 < 1 \)? Where, \( x_1 < 1 - x_2 \)

\[
\int \int_{x_1+x_2<1} f(x) \, dx = \int_0^1 \int_0^{1-x_2} 1 \, dx_1 x_2 \\
= \int_0^1 x_1 |_{0}^{1-x_2} \, dx_2
\]
Example 5: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \to \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$? Where is $x_1 + x_2 < 1$? Where, $x_1 < 1 - x_2$

$$\int \int_{x_1 + x_2 < 1} f(x) \, dx = \int_0^1 \int_0^{1-x_2} 1 \, dx_1 \, x_2$$

$$= \int_0^1 x_1 \bigg|_0^{1-x_2} \, dx_2$$

$$= \int_0^1 (1 - x_2) \, dx_2$$
Example 5: More Complicated Regions

Suppose $f[0, 1] \times [0, 1] \to \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$? Where is $x_1 + x_2 < 1$? Where, $x_1 < 1 - x_2$

\[
\int \int_{x_1 + x_2 < 1} f(x) \, dx = \int_0^1 \int_0^{1-x_2} 1 \, dx_1 \, x_2
\]

\[
= \int_0^1 x_1 \bigg|_0^{1-x_2} \, dx_2
\]

\[
= \int_0^1 (1 - x_2) \, dx_2
\]

\[
= x_2 \bigg|_0^1 - \frac{x_2^2}{2} \bigg|_0^1
\]

\[
= 1 - \frac{1}{2}
\]
Example 5: More Complicated Regions

Suppose $f : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, $f(x_1, x_2) = 1$. What is the area of $x_1 + x_2 < 1$? Where is $x_1 + x_2 < 1$? Where, $x_1 < 1 - x_2$

\[
\int \int_{x_1 + x_2 < 1} f(x) \, dx = \int_0^1 \int_0^{1-x_2} 1 \, dx_1 \, dx_2
\]

\[
= \int_0^1 x_1 \bigg|_0^{1-x_2} \, dx_2
\]

\[
= \int_0^1 (1 - x_2) \, dx_2
\]

\[
= x_2 \bigg|_0^1 - \frac{x_2^2}{2} \bigg|_0^1
\]

\[
= 1 - \left( \frac{1}{2} \right)
\]
Example 5: More Complicated Regions

Suppose \( f[0, 1] \times [0, 1] \rightarrow \mathbb{R}, f(x_1, x_2) = 1 \). What is the area of \( x_1 + x_2 < 1 \)? Where is \( x_1 + x_2 < 1 \)? Where, \( x_1 < 1 - x_2 \)

\[
\int \int_{x_1 + x_2 < 1} f(x) \, dx = \int_{0}^{1} \int_{0}^{1-x_2} 1 \, dx_1 \, x_2
\]
\[
= \int_{0}^{1} x_1 \bigg|_{0}^{1-x_2} \, dx_2
\]
\[
= \int_{0}^{1} (1 - x_2) \, dx_2
\]
\[
= x_2 \bigg|_{0}^{1} - \frac{x_2^2}{2} \bigg|_{0}^{1}
\]
\[
= 1 - \frac{1}{2}
\]
\[
= \frac{1}{2}
\]
Example 5: More Complicated Regions

Suppose \( f[0, 1] \times [0, 1] \rightarrow \mathbb{R} \), \( f(x_1, x_2) = 1 \). What is the area of \( x_1 + x_2 < 1 \)? Where is \( x_1 + x_2 < 1 \)? Where, \( x_1 < 1 - x_2 \)

\[
\int \int_{x_1 + x_2 < 1} f(x) \, dx = \int_0^1 \int_0^{1-x_2} 1 \, dx_1 \, x_2
\]

\[
= \int_0^1 x_1 |_{0}^{1-x_2} \, dx_2
\]

\[
= \int_0^1 (1 - x_2) \, dx_2
\]

\[
= x_2 |_{0}^{1} - \frac{x_2^2}{2} |_{0}^{1}
\]

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

\[
= \frac{1}{2}
\]
Multivariate Optimization

Optimizing multivariate functions

- Parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ such that $f(\beta | X, Y)$ is maximized
- Policy $x \in \mathbb{R}^n$ that maximizes $U(x)$
- Weights $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ such that a weighted average of forecasts $f = (f_1, f_2, \ldots, f_k)$ have minimum loss

$$
\min_{\pi} = -(\sum_{j=1}^{K} \pi_j f_j - y)^2
$$

Today we’ll describe analytic and computational approaches to optimization

- Analytic recipe for optimization
- Computational optimization
  - Multivariate Newton-Raphson
  - BFGS
  - Approximate Optimization: k-means
Multivariate Optimization

Definition

Let \( x \in \mathbb{R}^n \) and let \( \delta > 0 \). Define a neighborhood of \( x \), \( B(x, \delta) \), as the set of points such that,

\[
B(x, \delta) = \{ y \in \mathbb{R}^n : \|x - y\| < \delta \}
\]

Definition

Suppose \( f : X \rightarrow \mathbb{R} \) with \( X \subset \mathbb{R}^n \). A vector \( x^* \in X \) is a global maximum if, for all other \( x \in X \)

\[
f(x^*) > f(x)
\]

A vector \( x^{\text{local}} \) is a local maximum if there is a neighborhood around \( x^{\text{local}} \), \( Q \subset X \) such that, for all \( x \in Q \),

\[
f(x^{\text{local}}) > f(x)
\]
Multivariate Optimization

Definition

A set $X \subset \mathbb{R}^n$ is compact if it is closed and bounded

Theorem

**Multivariate Extreme Value Theorem** Suppose $f : X \to \mathbb{R}$ be continuous and $X \subset \mathbb{R}^n$ and $X$ compact. Then $f$ takes on its maximum and minimum values on $X$.

We’re going to come up with the multivariate equivalent of the first order and second order conditions now
Gradient

Definition

Suppose $f : X \rightarrow \mathbb{R}^n$ with $X \subset \mathbb{R}^1$ is a differentiable function. Define the gradient vector of $f$ at $x_0$, $\nabla f(x_0)$ as,

$$\nabla f(x_0) = \left( \frac{\partial f(x_0)}{\partial x_1}, \frac{\partial f(x_0)}{\partial x_2}, \frac{\partial f(x_0)}{\partial x_3}, \ldots, \frac{\partial f(x_0)}{\partial x_n} \right)$$
Gradient First Order Condition

**Theorem**

Suppose $f : X \to \mathbb{R}^1$, $X \subset \mathbb{R}^n$. Suppose $a \in X$ is a local extremum. Then,

$$\nabla f(a) = 0$$

$$= (0, 0, \ldots, 0)$$

- Proof (intuition): same as one dimensional case (left-hand, right hand), just do it dimension by dimension

- **Critical Values:**
  1) Maximum
  2) Minimum
  3) Saddle point

- **Second Derivative Test!**
Definition

Suppose \( f : X \rightarrow \mathbb{R}^1 , \ X \subset \mathbb{R}^n \), with \( f \) a twice differentiable function. We will define the **Hessian** matrix as the matrix of second derivatives at \( x^* \in X \),

\[
H(f)(x^*) = \begin{pmatrix}
\frac{\partial^2 f}{\partial x_1 \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x^*) \\
\frac{\partial^2 f}{\partial x_2 \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x^*) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x^*)
\end{pmatrix}
\]

**General test** ⇔ **Two Dimensional Test** ⇔ **Example**
Hessians

Definition

Consider \( n \times n \) matrix \( A \). If, for all \( x \in \mathbb{R}^n \) where \( x \neq 0 \):

\[
x'Ax > 0 \quad \text{A is positive definite}
\]

\[
x'Ax < 0 \quad \text{A is negative definite}
\]

If \( x'Ax > 0 \) for some \( x \) and \( x'Ax < 0 \) for other \( x \), then we say \( A \) is indefinite

Justin Grimmer (Stanford University)
Methodology I
September 8th, 2015 20 / 50
Approximating functions and second order conditions

Theorem

**Taylor’s Theorem** Suppose \( f : \mathbb{R} \rightarrow \mathbb{R} \), \( f(x) \) is infinitely differentiable function. Then, the taylor expansion of \( f(x) \) around \( a \) is given by

\[
f(x) = f(a) + \frac{f'(a)}{1!} (x - a) + \frac{f''(a)}{2!} (x - a)^2 + \frac{f'''(a)}{3!} (x - a)^3 + \ldots
\]

\[
f(x) = \sum_{n=0}^{\infty} \frac{f^n(a)}{n!} (x - a)^n
\]
Example Function

Suppose \( a = 0 \) and \( f(x) = e^x \). Then,

\[
\begin{align*}
    f'(x) &= e^x \\
    f''(x) &= e^x \\
    &\vdots \\
    f^n(x) &= e^x
\end{align*}
\]

This implies

\[
e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots
\]
Multivariate Taylor’s Theorem

**Theorem**

Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) is a three-times continuously differentiable function, then around \( a \in \mathbb{R}^n \),

\[
f(x) = f(a) + \nabla f(a)(x - a) + \frac{1}{2}(x - a)' H(f)(a)(x - a) + R(a, x)
\]

where \( \frac{R(x,a)}{||x-a||^2} \to 0 \) as \( x \to a \)
Intuition for Quadratic Form

Suppose $x^*$ is some critical value,

$$f(x) = f(x^*) + \nabla f(x^*)(x - x^*) + (x - \frac{1}{2}x^*)H(f)(x^*)(x - x^*) + R(x^*, x)$$

$$f(x) - f(x^*) = 0(x - x^*) + (x - \frac{1}{2}x^*)H(f)(x^*)(x - x^*) + R(x^*, x)$$

For $x$ near $x^*$, $R(x^*, x) \approx 0$

$H(f)(x^*)$ positive definite $\rightarrow f(x) > f(x^*) \rightarrow$ local minimum

$H(f)(x^*)$ negative definite $\rightarrow f(x) < f(x^*) \rightarrow$ local maximum
Theorem

Second Derivative Test

- If $H(f)(a)$ is positive definite then $a$ is a local minimum
- If $H(f)(a)$ is negative definite then $a$ is a local maximum
- If $H(f)(a)$ is indefinite then $a$ is a saddle point
Second Derivative Test

Many ways to assess definiteness \( \rightsquigarrow \) use determinant

**Theorem**

*Two Dimensional, Second Derivative Test.* Suppose \( f : X \to \mathbb{R} \) with \( X \subset \mathbb{R}^2 \) and \( f \) twice differentiable. Write the Hessian of \( f \) at a critical value \( a \),

\[
H(f)(a) = \begin{pmatrix} A & B \\ B & C \end{pmatrix}
\]

Then, we can conduct the second derivative test as:

- \( AC - B^2 > 0 \) and \( A > 0 \) \( \rightsquigarrow \) **positive definite** \( \rightsquigarrow \) \( a \) is a local minimum
- \( AC - B^2 > 0 \) and \( A < 0 \) \( \rightsquigarrow \) **negative definite** \( \rightsquigarrow \) \( a \) is a local maximum
- \( AC - B^2 < 0 \) \( \rightsquigarrow \) **indefinite** \( \rightsquigarrow \) saddle point
- \( AC - B^2 = 0 \) **inconclusive**
Multivariate Recipe

1) Calculate gradient
2) Set equal to zero, solve system of equations
3) Calculate Hessian
4) Assess Hessian at critical values
5) Boundary values? (if relevant)
Example 1: A Simple Optimization Problem

Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x_1, x_2) = 3(x_1 + 2)^2 + 4(x_2 + 4)^2$$

Calculate gradient

$$\nabla f(x) = (6x_1 + 12, 8x_2 + 32)$$

$$0 = (6x_1^* + 12, 8x_2^* + 32)$$

We now solve the system of equations to yield $x_1^* = -2$ and $x_2^* = -4$
Example 1: A Simple Optimization Problem

\[ H(f)(x^*) = \begin{pmatrix} 6 & 0 \\ 0 & 8 \end{pmatrix} \]

\[
\det(H(f)(x^*)) = 48 \text{ and } 6 > 0 \text{ so } H(f)(x^*) \text{ is positive definite. local minimum}
\]
Example 2: Two Dimensional Ideal Points

Suppose legislators are considering legislation $x \in \mathbb{R}^2$. And suppose legislator $i$ has utility function $U_i : \mathbb{R}^2 \to \mathbb{R}$,

$$U(x)_i = - (x_1 - \mu_1)^2 - (x_2 - \mu_2)^2$$

What is legislator $i$’s optimal policy?

$$\nabla f(x) = (-2(x_1 - \mu_1), -2(x_2 - \mu_2))$$

$$\nabla f(x) = 0$$

Solving yields $x_1^* = \mu_1$ and $x_2^* = \mu_2$. 
Example 2: Two Dimensional Ideal Points

\[ U(x)_i = -(x_1 - \mu_1)^2 - (x_2 - \mu_2)^2 \]

Call \( \mu = (\mu_1, \mu_2) \)

The Hessian at the critical value is

\[
H(f)(\mu) = \begin{pmatrix}
\frac{\partial^2 U_i}{\partial x_1 \partial x_1}(\mu) & \frac{\partial^2 U_i}{\partial x_1 \partial x_2}(\mu) \\
\frac{\partial^2 U_i}{\partial x_2 \partial x_1}(\mu) & \frac{\partial^2 U_i}{\partial x_2 \partial x_2}(\mu)
\end{pmatrix}
\]

\[
= \begin{pmatrix}
-2 & 0 \\
0 & -2
\end{pmatrix}
\]

So, \(-2 \times -2 - 0 = 4 > 0\) and \(-2 < 0 \leadsto \text{negative definite, maximum}\)

\( \mu = (\mu_1, \mu_2) \) are legislator \( i \)'s two dimensional ideal point.
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of \( n \) observations from a normal distribution, 

\[ Y_i \sim \text{Normal}(\mu, \sigma^2) \]

\[ Y = (Y_1, Y_2, \ldots, Y_n) \]

Our task:
- Obtain likelihood (summary estimator)
- Derive maximum likelihood estimators for \( \mu \) and \( \sigma^2 \)
- Characterize sampling distribution
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,

\[ Y_i \sim \text{Normal}(\mu, \sigma^2) \]
\[ Y = (Y_1, Y_2, \ldots, Y_n) \]
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

$$Y = (Y_1, Y_2, \ldots, Y_n)$$

Our task:
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$
$$Y = (Y_1, Y_2, \ldots, Y_n)$$

Our task:
- Obtain likelihood (summary estimator)
Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

$$Y = (Y_1, Y_2, \ldots, Y_n)$$

Our task:

- Obtain likelihood (summary estimator)
- Derive maximum likelihood estimators for $\mu$ and $\sigma^2$
Example 3: Maximum Likelihood Estimation, Normal Distribution

Suppose that we draw an independent and identically distributed random sample of $n$ observations from a normal distribution,

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$$

Our task:

- Obtain likelihood (summary estimator)
- Derive maximum likelihood estimators for $\mu$ and $\sigma^2$
- Characterize sampling distribution
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ L(\mu, \sigma^2 | Y) \propto n \prod_{i=1}^{n} f(Y_i | \mu, \sigma^2) \]
\[ \propto N \prod_{i=1}^{n} \exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right] \sqrt{\frac{2}{\pi\sigma^2}} \]

Taking the logarithm, we have
\[ l(\mu, \sigma^2 | Y) = -n \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + c \]
\[ = -n \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) + c' \]

Justin Grimmer (Stanford University)
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[
L(\mu, \sigma^2 | Y) \propto \prod_{i=1}^{n} f(Y_i | \mu, \sigma^2)
\]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ L(\mu, \sigma^2 | Y) \propto \prod_{i=1}^{n} f(Y_i | \mu, \sigma^2) \]

\[ \propto \prod_{i=1}^{n} \exp \left[ -\frac{(Y_i - \mu)^2}{2\sigma^2} \right] \frac{1}{\sqrt{2\pi\sigma^2}} \]

Taking the logarithm, we have

\[ l(\mu, \sigma^2 | Y) = -\frac{1}{2} n \sum_{i=1}^{n} \left( Y_i - \mu \right)^2 - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + c = -\frac{1}{2} n \sum_{i=1}^{n} \left( Y_i - \mu \right)^2 - \frac{n}{2} \log(\sigma^2) + c' \]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ L(\mu, \sigma^2 | Y) \propto \prod_{i=1}^{n} f(Y_i | \mu, \sigma^2) \]
\[ \propto \prod_{i=1}^{n} \exp \left[ -\frac{(Y_i - \mu)^2}{2\sigma^2} \right] \]
\[ \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \right] \]
\[ \propto \frac{1}{(2\pi)^{n/2} \sigma^{2n/2}} \]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[
L(\mu, \sigma^2 | Y) \propto \prod_{i=1}^{n} f(Y_i | \mu, \sigma^2) \\
\propto \prod_{i=1}^{N} \exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right] \\
\propto \prod_{i=1}^{N} \frac{\exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right]}{\sqrt{2\pi\sigma^2}} \\
\propto \exp\left[-\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2}\right] \\
\propto \frac{\exp\left[-\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2}\right]}{(2\pi)^{n/2}\sigma^{2n/2}}
\]

Taking the logarithm, we have

Example 3: Maximum Likelihood Estimation, Normal Distribution

\[
L(\mu, \sigma^2 | \mathbf{Y}) \propto \prod_{i=1}^{n} f(Y_i | \mu, \sigma^2) \\
\propto \prod_{i=1}^{N} \frac{\exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right]}{\sqrt{2\pi\sigma^2}} \\
\propto \exp\left[-\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2}\right] \frac{1}{(2\pi)^{n/2}\sigma^{2n/2}}
\]

Taking the logarithm, we have

\[
l(\mu, \sigma^2 | \mathbf{Y}) = -\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + c
\]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[
L(\mu, \sigma^2 \mid Y) \propto \prod_{i=1}^{n} f(Y_i \mid \mu, \sigma^2)
\]

\[
\propto \prod_{i=1}^{n} \frac{\exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right]}{\sqrt{2\pi\sigma^2}}
\]

\[
\propto \exp\left[-\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2}\right] \frac{1}{(2\pi)^{n/2}\sigma^{2n/2}}
\]

Taking the logarithm, we have

\[
l(\mu, \sigma^2 \mid Y) = -\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + c
\]

\[
= -\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) + c'
\]
Example 3: Log-Likelihood Plot

- In R, drew 10,000 realizations from

\[ Y_i \sim \text{Normal}(0, 25), 100) \]

\[ -28800 \]

\[ -28600 \]

\[ -28400 \]

\[ -28200 \]

\[ -2 \]

\[ -1 \]

\[ 0 \]

\[ 1 \]

\[ 2 \]

\[ 80 \]

\[ 100 \]

\[ 120 \]

\[ 140 \]

\[ 160 \]

\[ \mu \]

\[ \sigma \]

Justin Grimmer (Stanford University)

Methodology I

September 8th, 2015

34 / 50
Example 3: Log-Likelihood Plot

- In R, drew 10,000 realizations from

\[ Y_i \sim \text{Normal}(0.25, 100) \]
Example 3: Log-Likelihood Plot

- In R, drew 10,000 realizations from

\[ Y_i \sim \text{Normal}(0.25, 100) \]

- Used realized values \( y_i \) evaluate \( l(\mu, \sigma^2 | \mathbf{y}) \)
Example 3: Log-Likelihood Plot
Example 3: Log-Likelihood Plot
Example 3: Maximum Likelihood Estimation, Normal Distribution

Let’s find $\hat{\mu}$ and $\hat{\sigma}^2$ that maximizes log-likelihood.
Example 3: Maximum Likelihood Estimation, Normal Distribution

Let’s find $\hat{\mu}$ and $\hat{\sigma}^2$ that maximizes log-likelihood.

$$l(\mu, \sigma^2 | Y) = -\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) + c'$$
Example 3: Maximum Likelihood Estimation, Normal Distribution

Let’s find $\hat{\mu}$ and $\hat{\sigma^2}$ that maximizes log-likelihood.

$$l(\mu, \sigma^2 | Y) = - \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) + c'$$

$$\frac{\partial l(\mu, \sigma^2 | Y)}{\partial \mu} = \sum_{i=1}^{n} \frac{2(Y_i - \mu)}{2\sigma^2}$$
Example 3: Maximum Likelihood Estimation, Normal Distribution

Let’s find $\hat{\mu}$ and $\hat{\sigma}^2$ that maximizes log-likelihood.

\[
l(\mu, \sigma^2 | Y) = - \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) + c'
\]

\[
\frac{\partial l(\mu, \sigma^2 | Y)}{\partial \mu} = \sum_{i=1}^{n} \frac{2(Y_i - \mu)}{2\sigma^2}
\]

\[
\frac{\partial l(\mu, \sigma^2 | Y)}{\partial \sigma^2} = - \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \mu)^2
\]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[
0 = - \sum_{i=1}^{n} \frac{2(Y_i - \hat{\mu})}{2\hat{\sigma}^2}
\]

\[
0 = - \frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (Y_i - \mu^*)^2
\]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ 0 = - \sum_{i=1}^{n} \frac{2(Y_i - \hat{\mu})}{2\hat{\sigma}^2} \]

\[ 0 = - \frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (Y_i - \mu^*)^2 \]

Solving for \( \hat{\mu} \) and \( \hat{\sigma}^2 \) yields,
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ 0 = - \sum_{i=1}^{n} \frac{2(Y_i - \hat{\mu})}{2\hat{\sigma}^2} \]

\[ 0 = - \frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (Y_i - \mu^*)^2 \]

Solving for \( \hat{\mu} \) and \( \hat{\sigma}^2 \) yields,

\[ \hat{\mu} = \frac{\sum_{i=1}^{n} Y_i}{n} \]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ 0 = - \sum_{i=1}^{n} \frac{2(Y_i - \hat{\mu})}{2\hat{\sigma}^2} \]

\[ 0 = - \frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (Y_i - \mu^*)^2 \]

Solving for \( \hat{\mu} \) and \( \hat{\sigma}^2 \) yields,

\[ \hat{\mu} = \frac{\sum_{i=1}^{n} Y_i}{n} \]

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]
Example 3: Maximum Likelihood Estimation, Normal Distribution
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ H(f)(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \mu^2} & \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial^2 \sigma^2} \end{pmatrix} \]

Taking derivatives and evaluating at MLE's yields, \[ H(f)(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{n \hat{\sigma}^2}{2} & \frac{-n \hat{\sigma}^2}{2} \\ \frac{-n \hat{\sigma}^2}{2} & \frac{n^2}{\hat{\sigma}^5} \end{pmatrix} \]

and \[ \det(H(f)(\hat{\mu}, \hat{\sigma}^2)) = \frac{n^2}{\hat{\sigma}^5} \]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ H(f)(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \mu^2} & \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \sigma^2} \end{pmatrix} \]

Taking derivatives and evaluating at MLE's yields,
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[
\mathbf{H}(f)(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix}
\frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \mu^2} & \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \sigma^2 \partial \mu} \\
\frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu, \sigma^2 | Y)}{\partial^2 \sigma^2}
\end{pmatrix}
\]

Taking derivatives and evaluating at MLE’s yields,

\[
\mathbf{H}(f)(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix}
\frac{-n}{\hat{\sigma}^2} & 0 \\
0 & \frac{-n}{(\hat{\sigma}^2)^2}
\end{pmatrix}
\]
Example 3: Maximum Likelihood Estimation, Normal Distribution

\[ H(f)(\hat{\mu}, \hat{\sigma}^2) = \left( \begin{array}{cc} \frac{\partial^2 l(\mu, \sigma^2|Y)}{\partial \mu^2} & \frac{\partial^2 l(\mu, \sigma^2|Y)}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu, \sigma^2|Y)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu, \sigma^2|Y)}{\partial \sigma^4} \end{array} \right) \]

Taking derivatives and evaluating at MLE’s yields,

\[
H(f)(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{(\hat{\sigma}^2)^2} \end{pmatrix}
\]

\[
\text{det}(H(f)(\hat{\mu}, \hat{\sigma}^2)) = \frac{n^2}{\hat{\sigma}^5} \text{ and } -\frac{n}{\hat{\sigma}^2} < 0 \implies \text{maximum}
\]
Computational Optimization

Analytic solutions: often hard.
Computational Optimization

Analytic solutions: often hard.
Computational solutions: simplify. Trade offs
Computational Optimization

Analytic solutions: often hard.
Computational solutions: simplify. **Trade offs**
  - Newton-Raphson: expensive
Computational Optimization

Analytic solutions: often hard.
Computational solutions: simplify. Trade offs
- Newton-Raphson: expensive
- BFGS: less expensive
Analytic solutions: often hard.
Computational solutions: simplify. **Trade offs**
- Newton-Raphson: expensive
- BFGS: less expensive
- EM-like optimization: solve intractable problems, parallelizable
Multivariate Newton Raphson

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose we have guess $x_t$.

Then our update is:

$$
x_{t+1} = x_t - H(f)(x_t) - \frac{1}{2} \nabla f(x_t)
$$

Derivation (intuition):
Approximate function with tangent plane.
Find value of $x_{t+1}$ that makes the plane equal to zero. Update again.
Multivariate Newton Raphson

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose we have guess $x_t$. Then our update is:

$$x_{t+1} = x_t - H(f)(x_t) - \frac{1}{\lambda} \nabla f(x_t)$$

Derivation (intuition): Approximate function with tangent plane. Find value of $x_{t+1}$ that makes the plane equal to zero. Update again.
Multivariate Newton Raphson

Suppose $f : \mathbb{R}^n \to \mathbb{R}$. Suppose we have guess $x_t$. Then our update is:

$$x_{t+1} = x_t - H(f)(x_t)^{-1}\nabla f(x_t)$$
Suppose \( f : \mathbb{R}^n \rightarrow \mathbb{R} \). Suppose we have guess \( x_t \). Then our update is:

\[
x_{t+1} = x_t - H(f)(x_t)^{-1}\nabla f(x_t)
\]

Derivation (intuition):

Approximate function with tangent plane. Find value of \( x_{t+1} \) that makes the plane equal to zero. Update again.
Suppose \( f : \mathbb{R}^n \rightarrow \mathbb{R} \). Suppose we have guess \( x_t \). Then our update is:

\[
x_{t+1} = x_t - H(f)(x_t)^{-1} \nabla f(x_t)
\]

Derivation (intuition): Approximate function with tangent plane.
Multivariate Newton Raphson

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Suppose we have guess $x_t$. Then our update is:

$$x_{t+1} = x_t - H(f)(x_t)^{-1}\nabla f(x_t)$$

Derivation (intuition): Approximate function with tangent plane. Find value of $x_{t+1}$ that makes the plane equal to zero. Update again.
Multivariate Newton Raphson

Suppose \( f : \mathbb{R}^n \rightarrow \mathbb{R} \). Suppose we have guess \( x_t \). Then our update is:

\[
x_{t+1} = x_t - H(f)(x_t)^{-1}\nabla f(x_t)
\]

Derivation (intuition): Approximate function with tangent plane. Find value of \( x_{t+1} \) that makes the plane equal to zero. Update again.

R Code
Multivariate Newton Raphson

- Expensive to calculate (requires inverting Hessian)
- Very sensitive to starting points
- Ideally: method that exploits Newton-like structure, but is cheaper and more robust

BFGS: Quasi-Newton method

R code

Justin Grimmer (Stanford University)
Multivariate Newton Raphson

- Expensive to calculate (requires inverting Hessian)
Multivariate Newton Raphson

- Expensive to calculate (requires inverting Hessian)
- Very sensitive to starting points
Multivariate Newton Raphson

- Expensive to calculate (requires inverting Hessian)
- Very sensitive to starting points
- Ideally: method that exploits Newton-like structure, but is cheaper and more robust
Multivariate Newton Raphson

- Expensive to calculate (requires inverting Hessian)
- Very sensitive to starting points
- Ideally: method that exploits Newton-like structure, but is cheaper and more robust

BFGS: Quasi-Newton method
Multivariate Newton Raphson

- Expensive to calculate (requires inverting Hessian)
- Very sensitive to starting points
- Ideally: method that exploits Newton-like structure, but is cheaper and more robust

**BFGS:** Quasi-Newton method

R code
Optimization that is Both Discrete and Continuous

**K-means**: most commonly used clustering algorithm.

Story: Data are grouped in \( K \) clusters and each cluster has a center or mean.

→ Two types of parameters to estimate

1) For each cluster \( j \) (\( j = 1, \ldots, K \)),

\[
\begin{align*}
& r_{ij} = \text{Indicator, Document } i \text{ assigned to cluster } j \\
& r_{j} = (r_{1j}, r_{2j}, \ldots, r_{Nj})
\end{align*}
\]

\( r \) = \( (r_{1}^{'}1, r_{1}^{'}2, \ldots, r_{1}^{'}K) \) (\( N \times K \) matrix)

2) For each cluster \( j \), a cluster center for cluster \( j \).

\[
\mu_{j} = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{Mj})
\]

Notation. Representation of document \( i \):

\[
y_{i} = (y_{i1}, y_{i2}, \ldots, y_{iM})
\]
Optimization that is Both Discrete and Continuous

**K-means**: most commonly used clustering algorithm.

**Story**: Data are grouped in $K$ clusters and each cluster has a center or mean.
Optimization that is Both Discrete and Continuous

**K-means**: most commonly used clustering algorithm.

**Story**: Data are grouped in $K$ clusters and each cluster has a center or mean.

→ Two *types* of parameters to estimate
Optimization that is Both Discrete and Continuous

**K-means**: most commonly used clustering algorithm.

**Story**: Data are grouped in \( K \) clusters and each cluster has a center or mean.

→ **Two types** of parameters to estimate

1) For each cluster \( j \), \( (j = 1, \ldots, K) \)
   
   \[ r_{ij} = \text{Indicator, Document } i \text{ assigned to cluster } j \]

   \[ r_j = (r_{1j}, r_{2j}, \ldots, r_{Nj}) \]

   \[ r = (r'_1, r'_2, \ldots, r'_K) \quad (N \times K \text{ matrix}) \]
Optimization that is Both Discrete and Continuous

**K-means**: most commonly used clustering algorithm.

**Story**: Data are grouped in $K$ clusters and each cluster has a center or mean.

→ **Two types** of parameters to estimate

1) For each cluster $j$, ($j = 1, \ldots, K$)
   - $r_{ij} = $Indicator, Document $i$ assigned to cluster $j$
   - $r_j = (r_{1j}, r_{2j}, \ldots, r_{Nj})$
   - $r = (r_1', r_2', \ldots, r_K')$ ($N \times K$ matrix)

2) For each cluster $j$
   - $\mu_j$ a cluster center for cluster $j$.
   - $\mu_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{Mj})$
Optimization that is Both Discrete and Continuous

**K-means**: most commonly used clustering algorithm.

**Story**: Data are grouped in $K$ clusters and each cluster has a center or mean.

→ **Two types** of parameters to estimate

1) For each cluster $j$, $(j = 1, \ldots, K)$

   $r_{ij}$ = Indicator, Document $i$ assigned to cluster $j$

   $r_j = (r_{1j}, r_{2j}, \ldots, r_{Nj})$

   $r = (r'_1, r'_2, \ldots, r'_K)$ ($N \times K$ matrix)

2) For each cluster $j$

   $\mu_j$ a cluster center for cluster $j$.

   $\mu_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{Mj})$

**Notation.** Representation of document $i$:

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{iM})$$
Specifying the Method

1) Assume Euclidean distance between objects.
2) **Objective function**

\[ f(r, \mu, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right) \]

**Goal:** Choose \( r^* \) and \( \mu^* \) to minimize \( f(r, \mu, y) \)

- If \( K = N \) \( f(r^*, \mu^*, y) = 0 \) (Minimum)
- Each observation in own cluster
- If \( K = 1 \), \( f(r^*, \mu^*, y) = N \times \sigma^2 \)
- Each observation in one cluster
- Center: average of documents
Specifying the Method

1) Assume Euclidean distance between objects.
2) Objective function

\[ f(r, \mu, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right) \]

Goal:
Specifying the Method

1) Assume Euclidean distance between objects.

2) Objective function

\[
f(r, \mu, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right)
\]

Goal:
Choose \( r^* \) and \( \mu^* \) to minimize \( f(r, \mu, y) \)
Specifying the Method

1) Assume Euclidean distance between objects.
2) **Objective function**

\[ f(r, \mu, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right) \]

Goal:
Choose \( r^* \) and \( \mu^* \) to minimize \( f(r, \mu, y) \)

Two observations:
- If \( K = N \) \( f(r^*, \mu^*, y) = 0 \) (Minimum)
  - Each observation in own cluster
  - \( \mu_i = y_i \)
- If \( K = 1 \), \( f(r^*, \mu^*, y) = N \times \sigma^2 \)
  - Each observation in one cluster
  - Center: average of documents
Specifying the Method

1) Assume Euclidean distance between objects
2) Objective function
3) Algorithm for optimization

Iterative algorithm, Each Iteration $t$
- Conditional on $\mu^{t-1}$ (from previous iteration), choose $r^t$
- Conditional on $r^t$, choose $\mu^t$

Repeat until convergence, measured as change in $f$.

\[
\text{Change} = f(\mu^t, r^t, y) - f(\mu^{t-1}, r^{t-1}, y)
\]
Specifying the Method

\[
f(r, \mu, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right)
\]

Algorithm for estimation:
Begin: initialize \( \mu_{1}^{t-1}, \mu_{2}^{t-1}, \ldots, \mu_{K}^{t-1} \)
Choose \( r^{t} \)

\[
r_{ij}^{t} = \begin{cases} 
1 & \text{if } j = \arg \min_{k} \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \\
0 & \text{otherwise}
\end{cases}
\]

In words: Assign each document \( y_{i} \) to the closest center \( \mu_{k} \)
\[ f(r, \mu, y) = \sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right) \]

Conditional on \( r^t \), choose \( \mu^t \)

Let’s focus on \( \mu_k \)

\[ f(r, \mu_k, y)_k = \sum_{i=1}^{N} r_{ik} \left( \sum_{m=1}^{M} (y_{im} - \mu_{km})^2 \right) \]
Focus on just $\mu_{km}$

$$f(r, \mu_{km}, y)_{km} = \sum_{i=1}^{N} r_{ik}(y_{im} - \mu_{km})^2$$

Quadratic: take derivative, set equal to zero (second derivative test works)

$$\frac{\partial f(r, \mu_{km}, y)_{km}}{\partial \mu_{km}} = -2 \sum_{i=1}^{N} r_{ik}(y_{im} - \mu_{km})$$

$$2 \sum_{i=1}^{N} r_{ik}(y_{im} - \mu_{km}^t) = 0$$

$$\sum_{i=1}^{N} r_{ik}y_{im} - \mu_{km}^t \sum_{i=1}^{N} r_{ik} = 0$$

$$\frac{\sum_{i=1}^{N} r_{ik}y_{im}}{\sum_{i=1}^{N} r_{ik}} = \mu_{km}^t$$
\[ \mu_k^t = \frac{\sum_{i=1}^{N} r_{ik} y_i}{\sum_{i=1}^{N} r_{ik}} \]

In words:
- \( \mu_k^t \) is the average of documents assigned to the \( k^{th} \) cluster

Algorithm, In Words
- Conditional on center estimates, assign documents to closest cluster centers
- Conditional on document assignments, cluster centers are averages of documents assigned to the cluster

Expectation-Maximization (EM) [connection guarantees convergence]
- Estimation of \( r \sim \) Expectation step (data augmentation)
- Estimation of \( \mu_k \sim \) Maximization Step
Visual Example
Visual Example
Visual Example
Visual Example
Visual Example
Visual Example
Visual Example
Visual Example
Visual Example
Visual Example
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
- Randomly select most fit, then combine
- Can converge to global maximum, but might require extensive run time
Many Optimization Procedures!!!

Nelder Mead:

- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:

- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:

- Evaluate fitness of solutions
- Randomly select most fit, then combine
- Can converge to global maximum, but might require extensive run time
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
- Randomly select most fit, then combine
- Can converge to global maximum, but might require extensive run time
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
- Randomly select most fit, then combine
- Can converge to global maximum, but might require extensive run time
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
  - Sample a new subset, perform optimization, combine with previous sample
  - Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
- Randomly select most fit, then combine
- Can converge to global maximum, but might require extensive run time
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
- Randomly select most fit, then combine
Many Optimization Procedures!!!

Nelder Mead:
- Evaluate points on a simplex (triangle)
- Either Reflect, Expand, or Contract (based on values)
- Converges to local extrema

Stochastic Optimization:
- Sample a subset of data, perform optimization
- Sample a new subset, perform optimization, combine with previous sample
- Converges on local extrema (given regulatory conditions)

Genetic Optimization:
- Evaluate fitness of solutions
- Randomly select most fit, then combine
- Can converge to **global** maximum, but might require extensive run time
Where We Are Going

- Done with math component
- Start probability tomorrow