Logistic Regression

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM HINRICH SCHÜTZE

What are we talking about?

- Statistical classification: $p(y|x)$
- y is typically a Bernoulli or multinomial outcome
- Classification uses: ad placement, spam detection
- Building block of other machine learning methods

Logistic Regression: Definition

- Weight vector β_i
- Observations X_i
- "Bias" β_0 (like intercept in linear regression)

$$P(Y=0|X) = \frac{1}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]} \quad (1)$$

$$P(Y=1|X) = \frac{\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]} \quad (2)$$

- For shorthand, we'll say that

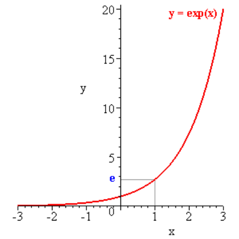
$$P(Y=0|X) = \sigma(-(\beta_0 + \sum_i \beta_i X_i)) \quad (3)$$

$$P(Y=1|X) = 1 - \sigma(-(\beta_0 + \sum_i \beta_i X_i)) \quad (4)$$

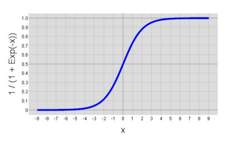
- Where $\sigma(z) = \frac{1}{1+exp[-z]}$

What's this "exp" doing?

Exponential



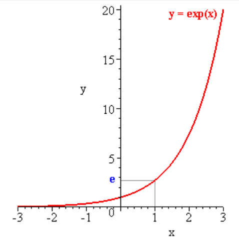
Logistic



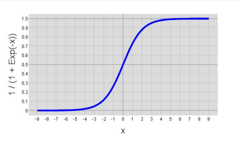
- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - e^x is the limit of compound interest formula as compounds become infinitely small
  - It's the function whose derivative is itself
- The "logistic" function is $\sigma(z) = \frac{1}{1+e^{-z}}$
- Looks like an "S"
- Always between 0 and 1.

What's this "exp" doing?

Exponential



Logistic



- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - e^x is the limit of compound interest formula as compounds become infinitely small
  - It's the function whose derivative is itself
- The "logistic" function is $\sigma(z) = \frac{1}{1+e^{-z}}$
- Looks like an "S"
- Always between 0 and 1.
  - Allows us to model probabilities
  - Different from **linear** regression

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	−1.0
"work"	β_3	−0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	–1.0
"work"	β_3	–0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1]} =$
- $P(Y = 1) = \frac{\exp[0.1]}{1+\exp[0.1]} =$

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	−1.0
"work"	β_3	−0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1]} = 0.48$
- $P(Y = 1) = \frac{\exp[0.1]}{1+\exp[0.1]} = 0.52$
- Bias β_0 encodes the prior probability of a class

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	−1.0
"work"	β_3	−0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	–1.0
"work"	β_3	–0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0+3.0]} =$
- $P(Y=1) = \frac{\exp[0.1-1.0+3.0]}{1+\exp[0.1-1.0+3.0]} =$
- Include bias, and sum the other weights

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	−1.0
"work"	β_3	−0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0+3.0]} = 0.11$
- $P(Y=1) = \frac{\exp[0.1-1.0+3.0]}{1+\exp[0.1-1.0+3.0]} = 0.88$
- Include bias, and sum the other weights

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	−1.0
"work"	β_3	−0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 3

$X = \{\text{Mother}, \text{Work}, \text{Viagra}, \text{Mother}\}$

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	–1.0
"work"	β_3	–0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 3

$X = \{\text{Mother}, \text{Work}, \text{Viagra}, \text{Mother}\}$

- $P(Y = 0) = \frac{1}{1 + \exp[0.1 - 1.0 - 0.5 + 2.0 - 1.0]} =$
- $P(Y = 1) = \frac{\exp[0.1 - 1.0 - 0.5 + 2.0 - 1.0]}{1 + \exp[0.1 - 1.0 - 0.5 + 2.0 - 1.0]} =$
- Multiply feature presence by weight

Logistic Regression Example

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	−1.0
"work"	β_3	−0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 3

$X = \{\text{Mother}, \text{Work}, \text{Viagra}, \text{Mother}\}$

- $P(Y=0) = \frac{1}{1+\exp\left[0.1-1.0-0.5+2.0-1.0\right]} = 0.60$
- $P(Y=1) = \frac{\exp\left[0.1-1.0-0.5+2.0-1.0\right]}{1+\exp\left[0.1-1.0-0.5+2.0-1.0\right]} = 0.30$
- Multiply feature presence by weight

Logistic Regression

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

ABC

Logistic Regression: Objective Function

$$\ell \equiv \ln p(Y \mid X, \beta) = \sum_j \ln p(y^{(j)} \mid x^{(j)}, \beta) \tag{1}$$

$$= \sum_j y^{(j)} \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) - \ln \left[1 + \exp \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) \right] \tag{2}$$

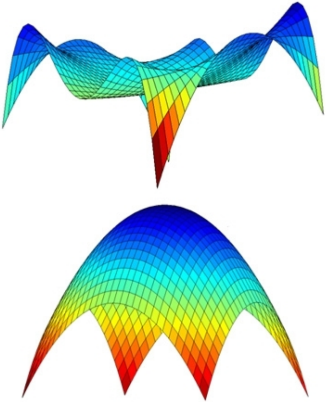
Logistic Regression: Objective Function

$$\ell \equiv \ln p(Y \mid X, \beta) = \sum_j \ln p(y^{(j)} \mid x^{(j)}, \beta) \tag{1}$$

$$= \sum_j y^{(j)} \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) - \ln \left[1 + \exp \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) \right] \tag{2}$$

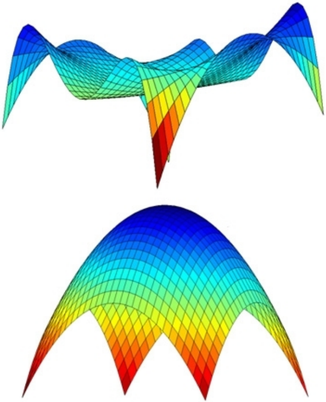
Training data (y, x) are fixed. Objective function is a function of β ... what values of β give a good value.

Convexity



- Convex function
- Doesn’t matter where you start, if you “walk up” objective

Convexity

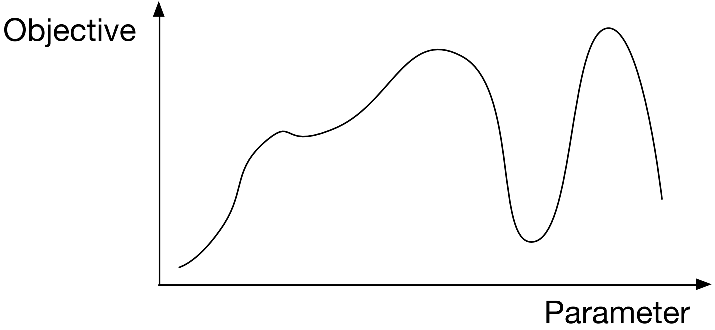


- Convex function
- Doesn't matter where you start, if you "walk up" objective
- Gradient!

Gradient Ascent (non-convex)

Goal

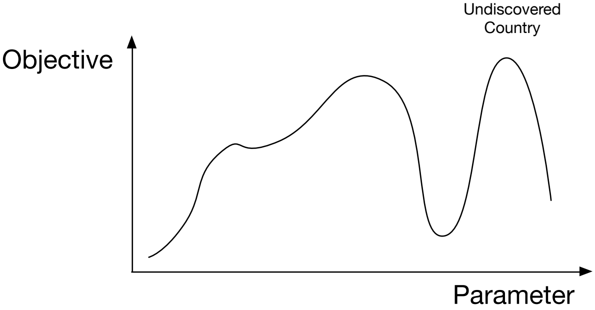
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

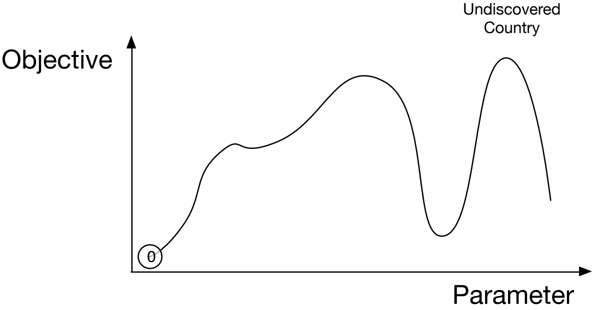
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

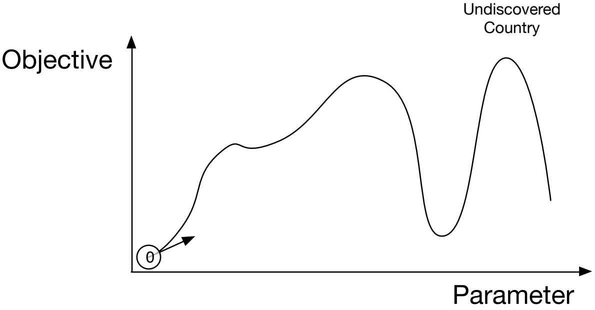
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

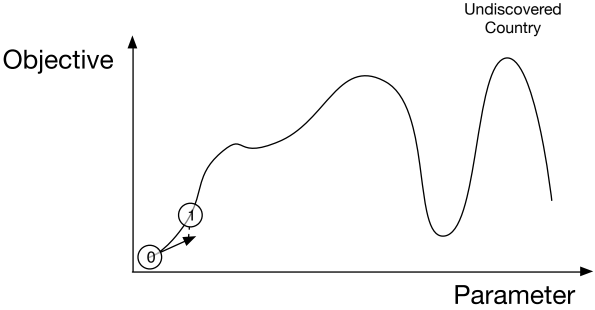
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

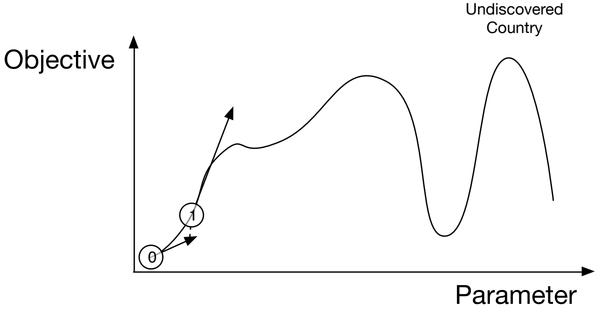
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

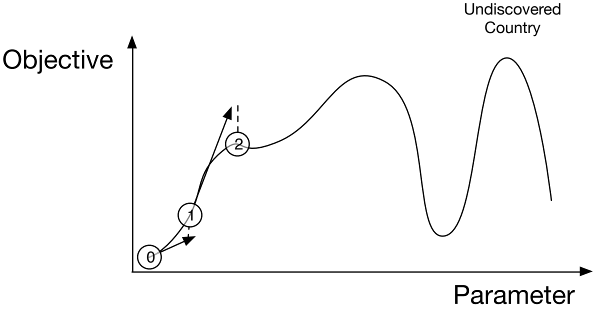
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

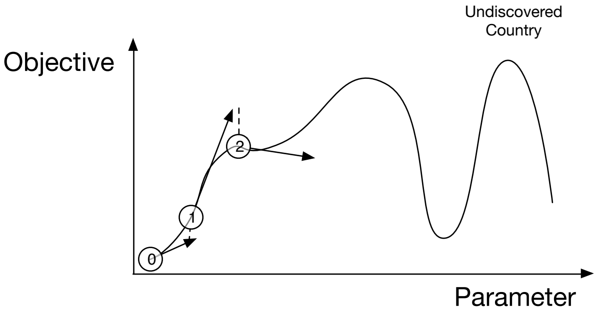
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

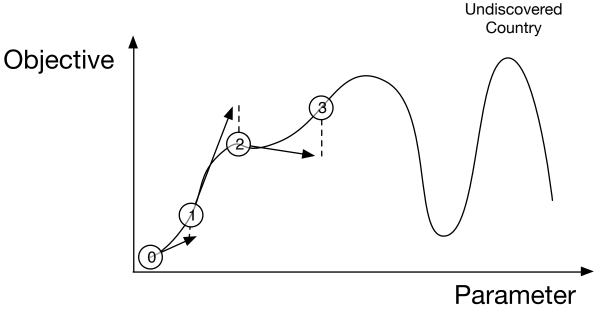
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

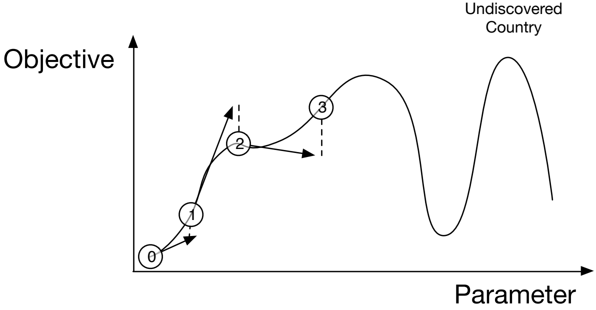
Optimize log likelihood with respect to variables β



Gradient Ascent (non-convex)

Goal

Optimize log likelihood with respect to variables β



Luckily, (vanilla) logistic regression is convex

Logistic Regression

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM WILLIAM COHEN

Gradient for Logistic Regression

To ease notation, let's define

$$\pi_i = \frac{\exp \beta^T x_i}{1 + \exp \beta^T x_i} \tag{1}$$

Our objective function is

$$\ell = \sum_i \log p(y_i \,|\, x_i) = \sum_i \ell_i = \sum_i \begin{cases} \log \pi_i & \text{if } y_i = 1 \\ \log(1 - \pi_i) & \text{if } y_i = 0 \end{cases} \tag{2}$$

Taking the Derivative

Apply chain rule:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{\partial \ell_i(\vec{\beta})}{\partial \beta_j} = \sum_i \begin{cases} \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \beta_j} & \text{if } y_i = 1 \\ \frac{1}{1-\pi_i} \left(-\frac{\partial \pi_i}{\partial \beta_j} \right) & \text{if } y_i = 0 \end{cases} \tag{3}$$

If we plug in the derivative,

$$\frac{\partial \pi_i}{\partial \beta_j} = \pi_i (1 - \pi_i) x_j, \tag{4}$$

we can merge these two cases

$$\frac{\partial \ell_i}{\partial \beta_j} = (y_i - \pi_i) x_j. \tag{5}$$

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \ell(\vec{\beta}) = \left[\frac{\partial \ell(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \ell(\vec{\beta})}{\partial \beta_n}\right] \tag{6}$$

Update

$$\Delta\beta \equiv \eta \nabla_\beta \ell(\vec{\beta}) \tag{7}$$

$$\beta_i' \leftarrow \beta_i + \eta \frac{\partial \ell(\vec{\beta})}{\partial \beta_i} \tag{8}$$

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \ell(\vec{\beta}) = \left[\frac{\partial \ell(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \ell(\vec{\beta})}{\partial \beta_n} \right] \tag{6}$$

Update

$$\Delta\beta \equiv \eta \nabla_\beta \ell(\vec{\beta}) \tag{7}$$

$$\beta_i' \leftarrow \beta_i + \eta \frac{\partial \ell(\vec{\beta})}{\partial \beta_i} \tag{8}$$

Why are we adding? What would well do if we wanted to do **descent**?

Gradient for Logistic Regression

Gradient

$$\nabla_{\beta} \ell(\vec{\beta}) = \left[\frac{\partial \ell(\vec{\beta})}{\partial \beta_0}, \ldots, \frac{\partial \ell(\vec{\beta})}{\partial \beta_n} \right] \tag{6}$$

Update

$$\Delta \beta \equiv \eta \nabla_{\beta} \ell(\vec{\beta}) \tag{7}$$

$$\beta_i' \leftarrow \beta_i + \eta \frac{\partial \ell(\vec{\beta})}{\partial \beta_i} \tag{8}$$

η: step size, must be greater than zero

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \ell(\vec{\beta}) = \left[\frac{\partial \ell(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \ell(\vec{\beta})}{\partial \beta_n} \right] \tag{6}$$

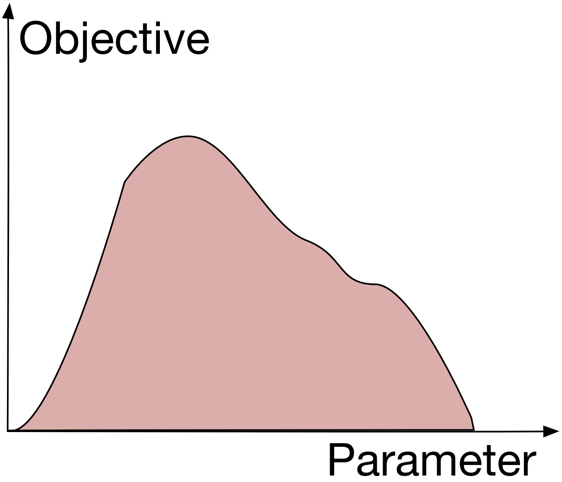
Update

$$\Delta \beta \equiv \eta \nabla_\beta \ell(\vec{\beta}) \tag{7}$$

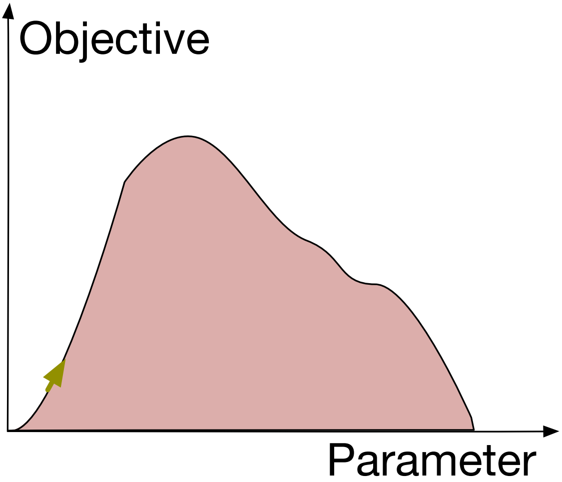
$$\beta_i' \leftarrow \beta_i + \eta \frac{\partial \ell(\vec{\beta})}{\partial \beta_i} \tag{8}$$

NB: Conjugate gradient is usually better, but harder to implement

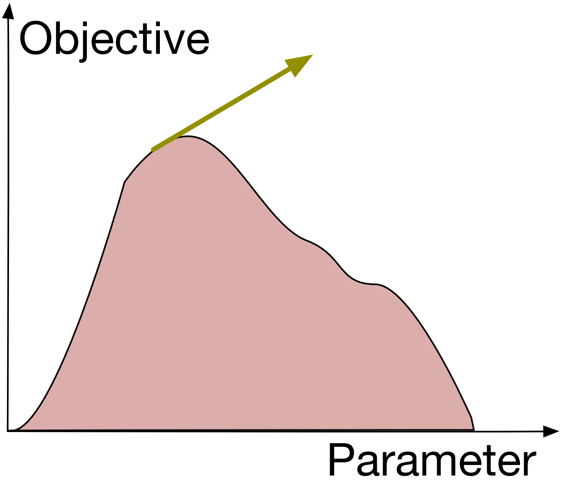
Choosing Step Size



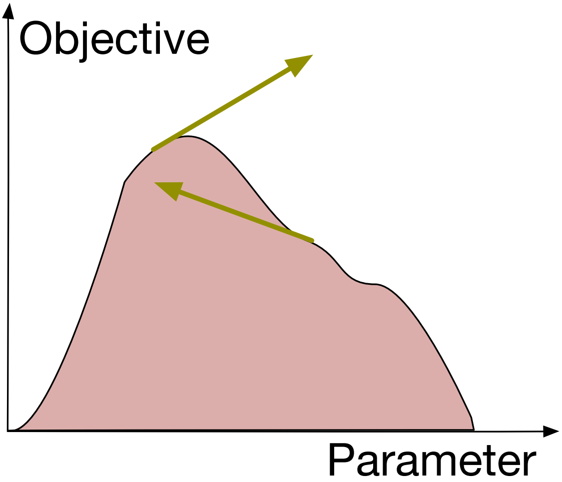
Choosing Step Size



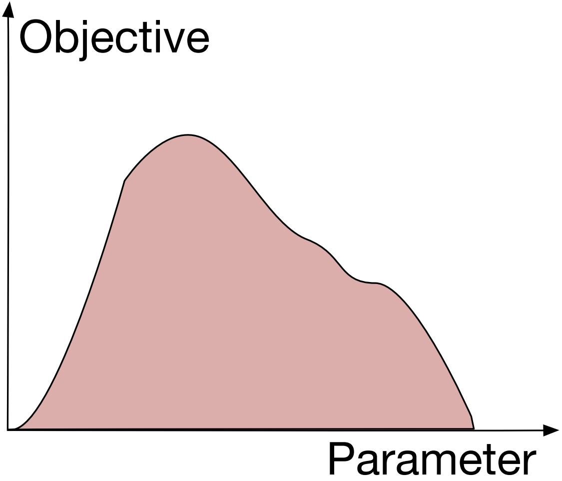
Choosing Step Size



Choosing Step Size



Choosing Step Size



Approximating the Gradient

- Our datasets are big (to fit into memory)
- ...or data are changing / streaming

Approximating the Gradient

- Our datasets are big (to fit into memory)
- ...or data are changing / streaming
- Hard to compute true gradient

$$\ell(\beta) \equiv \mathbb{E}_x \left[\nabla \ell(\beta, x)\right] \tag{9}$$

- Average over all observations

Approximating the Gradient

- Our datasets are big (to fit into memory)
- ... or data are changing / streaming
- Hard to compute true gradient

$$\ell(\beta) \equiv \mathbb{E}_x [\nabla \ell(\beta, x)] \tag{9}$$

- Average over all observations
- What if we compute an update just from one observation?

Getting to Union Station

Pretend it’s a pre-smartphone world and you want to get to Union Station

Stochastic Gradient for Logistic Regression

Given a **single observation** x_i chosen at random from the dataset,

$$\beta_j \leftarrow \beta_j' + \eta \left[y_i - \pi_i \right] x_{i,j} \tag{10}$$

Stochastic Gradient for Logistic Regression

Given a **single observation** x_i chosen at random from the dataset,

$$\beta_j \leftarrow \beta_j' + \eta \left[y_i - \pi_i \right] x_{i,j} \tag{10}$$

Examples in class.

Algorithm

1. Initialize a vector B to be all zeros
2. For $t = 1, \ldots, T$
   - For each example $\vec{x}_i, y_i$ and feature j:
     - Compute $\pi_i \equiv \Pr(y_i = 1 \,|\, \vec{x}_i)$
     - Set $\beta[j] = \beta[j]' + \lambda(y_i - \pi_i)x_i$
3. Output the parameters $\beta_1, \ldots, \beta_d$.

Wrapup

- Logistic Regression: Regression for outputting Probabilities
- Intuitions similar to linear regression
- We'll talk about feature engineering for both next time