Classification: Logistic Regression from Data

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder

LECTURE 3

Slides adapted from Emily Fox

Reminder: Logistic Regression

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

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

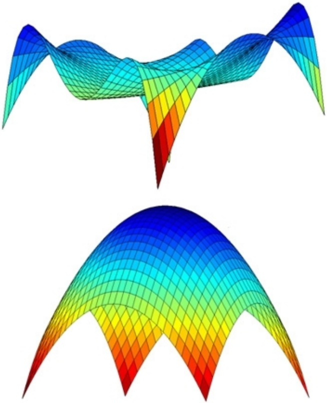
- Discriminative prediction: $p(y|x)$
- Classification uses: ad placement, spam detection
- What we didn't talk about is how to learn β from data

Logistic Regression: Objective Function

$$\mathscr{L} \equiv \ln p(Y \,|\, X, \beta) = \sum_j \ln p(y^{(j)} \,|\, x^{(j)}, \beta) \tag{3}$$

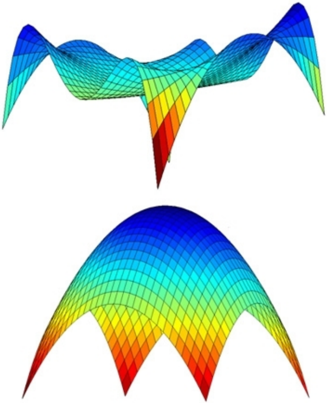
$$= \sum_j y^{(j)} \left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) - \ln \left[1 + \exp\left(\beta_0 + \sum_i \beta_i x_i^{(j)} \right) \right] \tag{4}$$

Convexity



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

Convexity

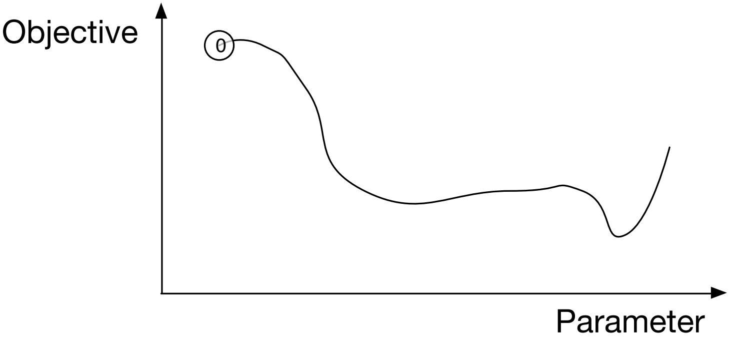


- Convex function
- Doesn't matter where you start, if you "walk up" objective
- Gradient!

Gradient Descent (non-convex)

Goal

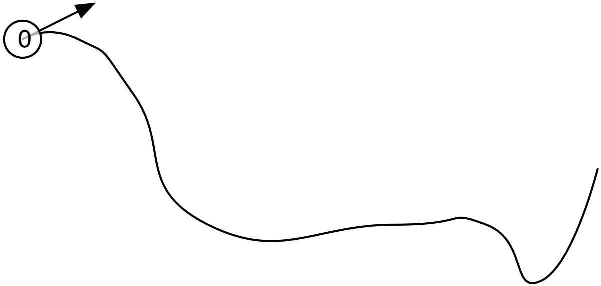
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

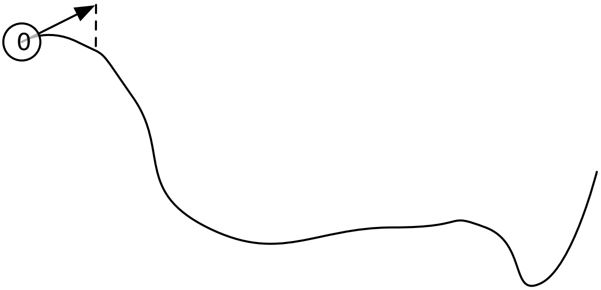
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

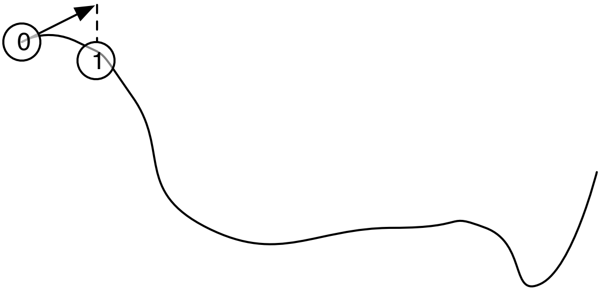
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

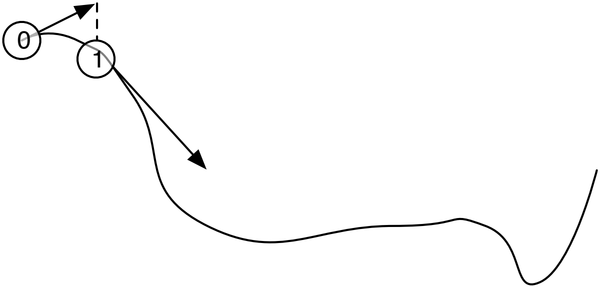
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

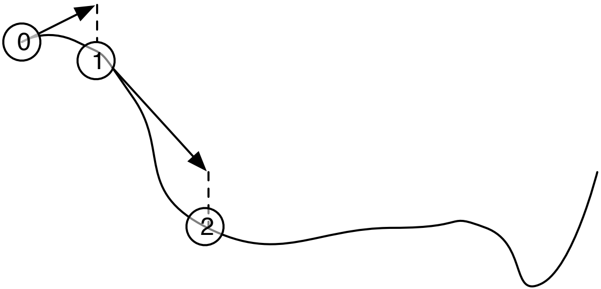
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

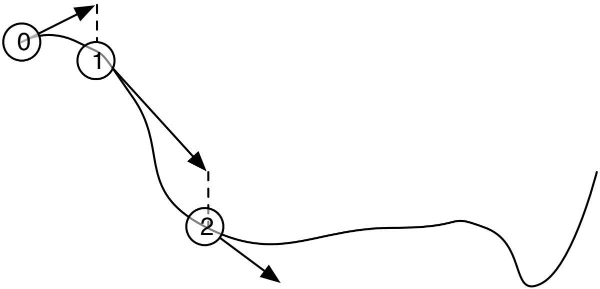
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

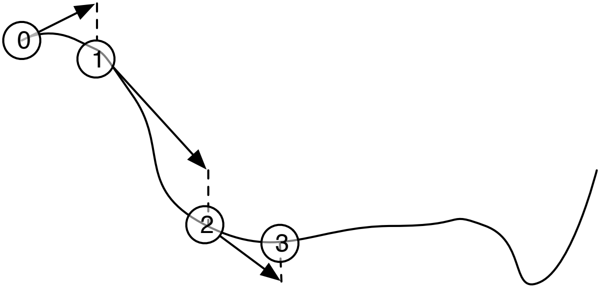
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

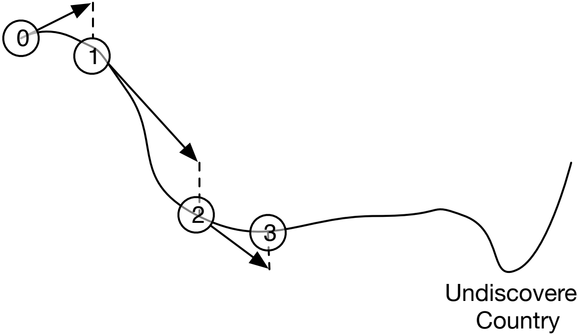
Optimize log likelihood with respect to variables *W* and *b*



Gradient Descent (non-convex)

Goal

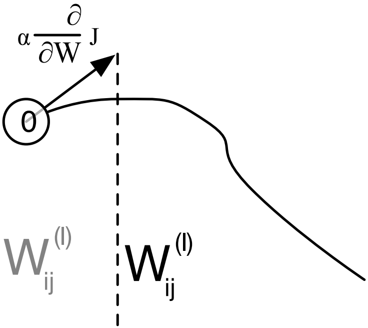
Optimize log likelihood with respect to variables W and b



Gradient Descent (non-convex)

Goal

Optimize log likelihood with respect to variables *W* and *b*



Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n} \right] \tag{5}$$

Update

$$\Delta \beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{6}$$

$$\beta_i \leftarrow \beta_i' + \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{7}$$

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n} \right] \tag{5}$$

Update

$$\Delta \beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{6}$$

$$\beta_i \leftarrow \beta_i' + \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{7}$$

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

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n} \right] \tag{5}$$

Update

$$\Delta \beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{6}$$

$$\beta_i \leftarrow \beta_i' + \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{7}$$

η: step size, must be greater than zero

Gradient for Logistic Regression

Gradient

$$\nabla_\beta \mathscr{L}(\vec{\beta}) = \left[\frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_0}, \dots, \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_n} \right] \tag{5}$$

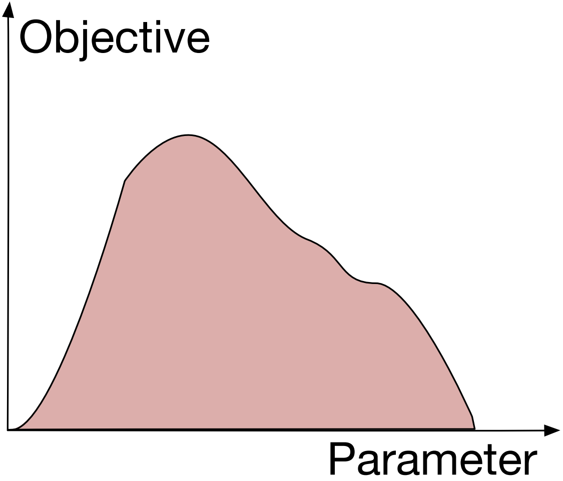
Update

$$\Delta\beta \equiv \eta \nabla_\beta \mathscr{L}(\vec{\beta}) \tag{6}$$

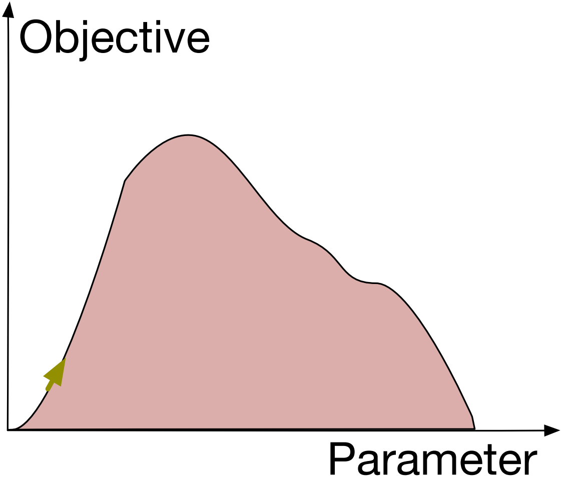
$$\beta_i \leftarrow \beta_i' + \eta \frac{\partial \mathscr{L}(\vec{\beta})}{\partial \beta_i} \tag{7}$$

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

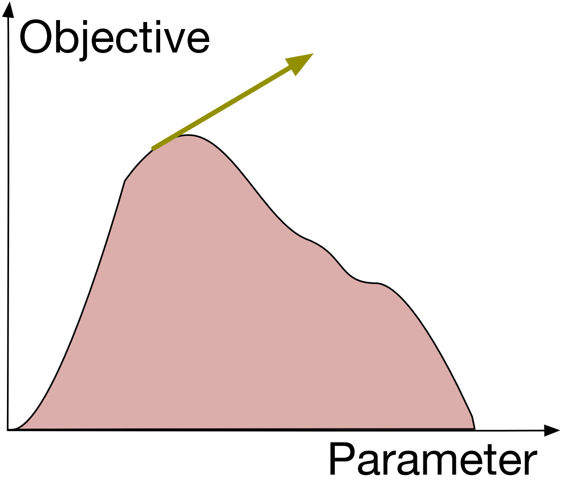
Choosing Step Size



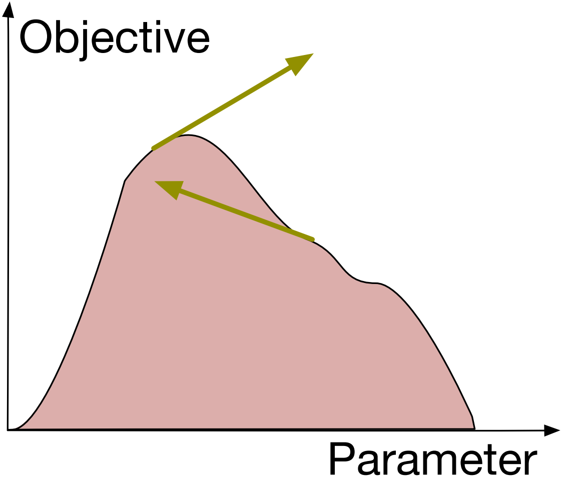
Choosing Step Size



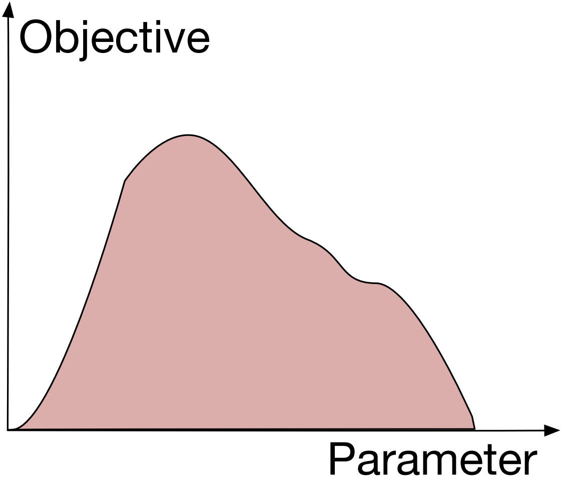
Choosing Step Size



Choosing Step Size



Choosing Step Size



Remaining issues

- When to stop?
- What if β keeps getting bigger?

Regularized Conditional Log Likelihood

Unregularized

$$\beta^* = \arg\max_{\beta} \ln\left[p(y^{(j)} \,|\, x^{(j)}, \beta)\right] \tag{8}$$

Regularized

$$\beta^* = \arg\max_{\beta} \ln\left[p(y^{(j)} \,|\, x^{(j)}, \beta)\right] - \mu \sum_i \beta_i^2 \tag{9}$$

Regularized Conditional Log Likelihood

Unregularized

$$\beta^* = \arg\max_{\beta} \ln \left[p(y^{(j)} \,|\, x^{(j)}, \beta) \right] \tag{8}$$

Regularized

$$\beta^* = \arg\max_{\beta} \ln \left[p(y^{(j)} \,|\, x^{(j)}, \beta) \right] - \mu \sum_{i} \beta_i^2 \tag{9}$$

μ is “regularization” parameter that trades off between likelihood and having small parameters

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

$$\mathscr{L}(\beta) \equiv \mathbb{E}_x [\nabla \mathscr{L}(\beta, x)] \tag{10}$$

- 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

$$\mathcal{L}(\beta) \equiv \mathbb{E}_x \left[\nabla \mathcal{L}(\beta, x) \right] \tag{10}$$

- 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* chosen at random from the dataset,

$$\beta_i \leftarrow \beta_i' + \eta \left(-\mu \beta_i' + x_i \left[y - p(y = 1 \,|\, \vec{x}, \vec{\beta}') \right] \right) \tag{11}$$

Stochastic Gradient for Logistic Regression

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

$$\beta_i \leftarrow \beta_i' + \eta \left(-\mu \beta_i' + x_i \left[y - p(y = 1 \,|\, \vec{x}, \vec{\beta}') \right] \right) \tag{11}$$

Examples in class.

Stochastic Gradient for Regularized Regression

$$\mathscr{L} = \log p(y \,|\, x; \beta) - \mu \sum_j \beta_j^2 \tag{12}$$

Stochastic Gradient for Regularized Regression

$$\mathscr{L} = \log p(y \,|\, x; \beta) - \mu \sum_j \beta_j^2 \tag{12}$$

Taking the derivative (with respect to example x_i)

$$\frac{\partial \mathscr{L}}{\partial \beta_j} = (y_i - p(y_i = 1 \,|\, \vec{x}_i; \beta))x_j - 2\mu\beta_j \tag{13}$$

Proofs about Stochastic Gradient

- Depends on convexity of objective and how close ϵ you want to get to actual answer
- Best bounds depend on changing η over time and **per dimension** (not all features created equal)

In class

- Your questions!
- Working through simple example
- Prepared for logistic regression homework