10-301/601: Introduction to Machine Learning Lecture 9 – Logistic Regression

Henry Chai

5/19/25

Recall: Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \rightarrow \mathcal{Y}$
  - Classifier, $h : \mathcal{X} \rightarrow \mathcal{Y}$
  - Goal: find a classifier, h, that best approximates c^*
- Now:
  - (Unknown) Target *distribution*, $y \sim P^*(Y|\boldsymbol{x})$
  - Distribution, $P(Y|\boldsymbol{x})$
  - Goal: find a distribution, P, that best approximates P^*

Building a Probabilistic Classifier

1. Define a decision rule
   - Given a test data point $\boldsymbol{x}'$, predict its label $\hat{y}$ using the *posterior distribution* $P(Y = y | X = \boldsymbol{x}')$
   - Common choice: $\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y = y | X = \boldsymbol{x}')$
2. Model the posterior distribution
   - Option 1 - Model $P(Y|X)$ directly as some function of X (today!)
   - Option 2 - Use Bayes' rule (later):
     - $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \propto P(X|Y)\, P(Y)$

Modelling the Posterior

- Suppose we have binary labels $y \in \{0,1\}$ and D-dimensional inputs $\boldsymbol{x} = [1, x_1, \dots, x_D]^T \in \mathbb{R}^{D+1}$
- **Assume**

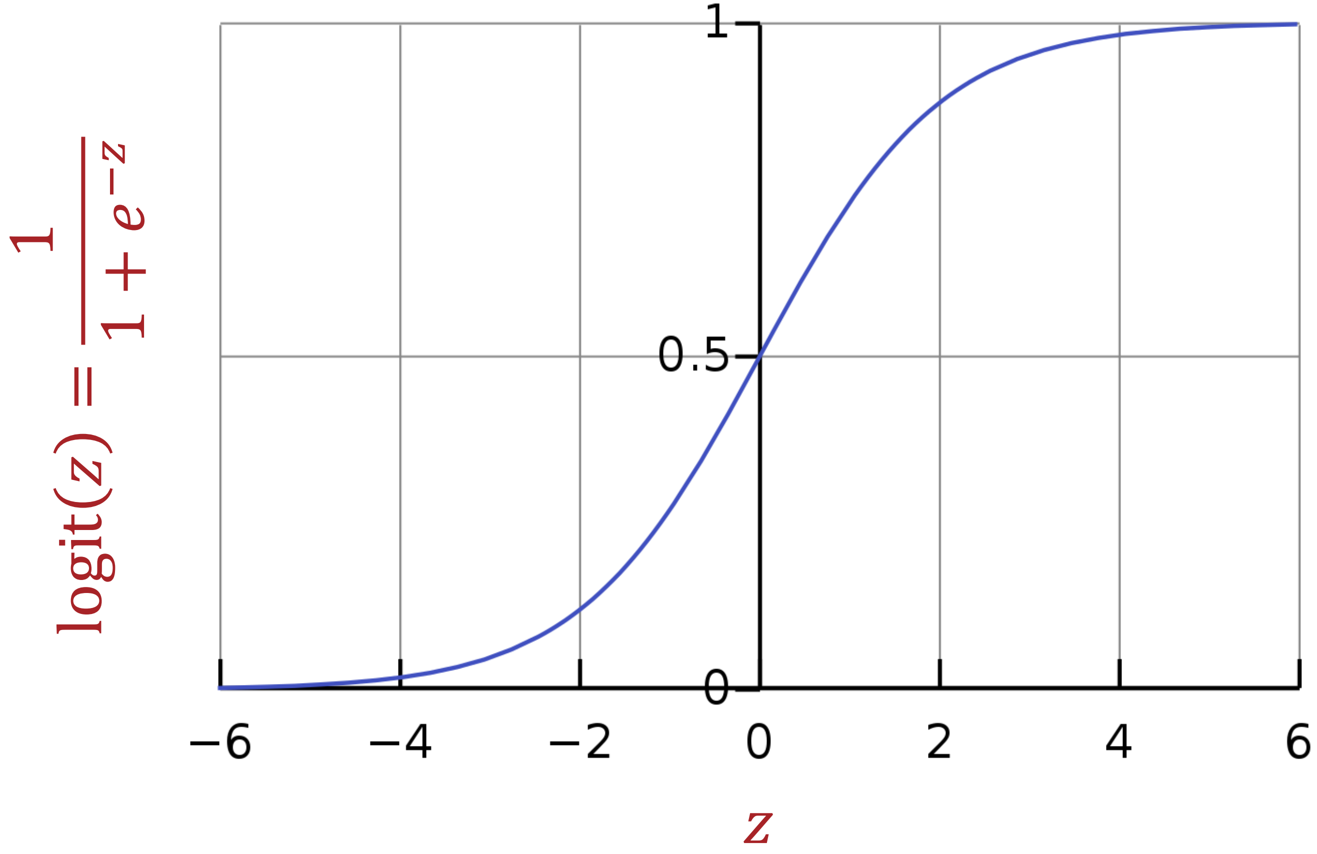
$$P(Y = 1|\boldsymbol{x}) = \text{logit}(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})}$$

$$= \frac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{\exp(\boldsymbol{w}^T\boldsymbol{x}) + 1}$$

- This implies two useful facts:

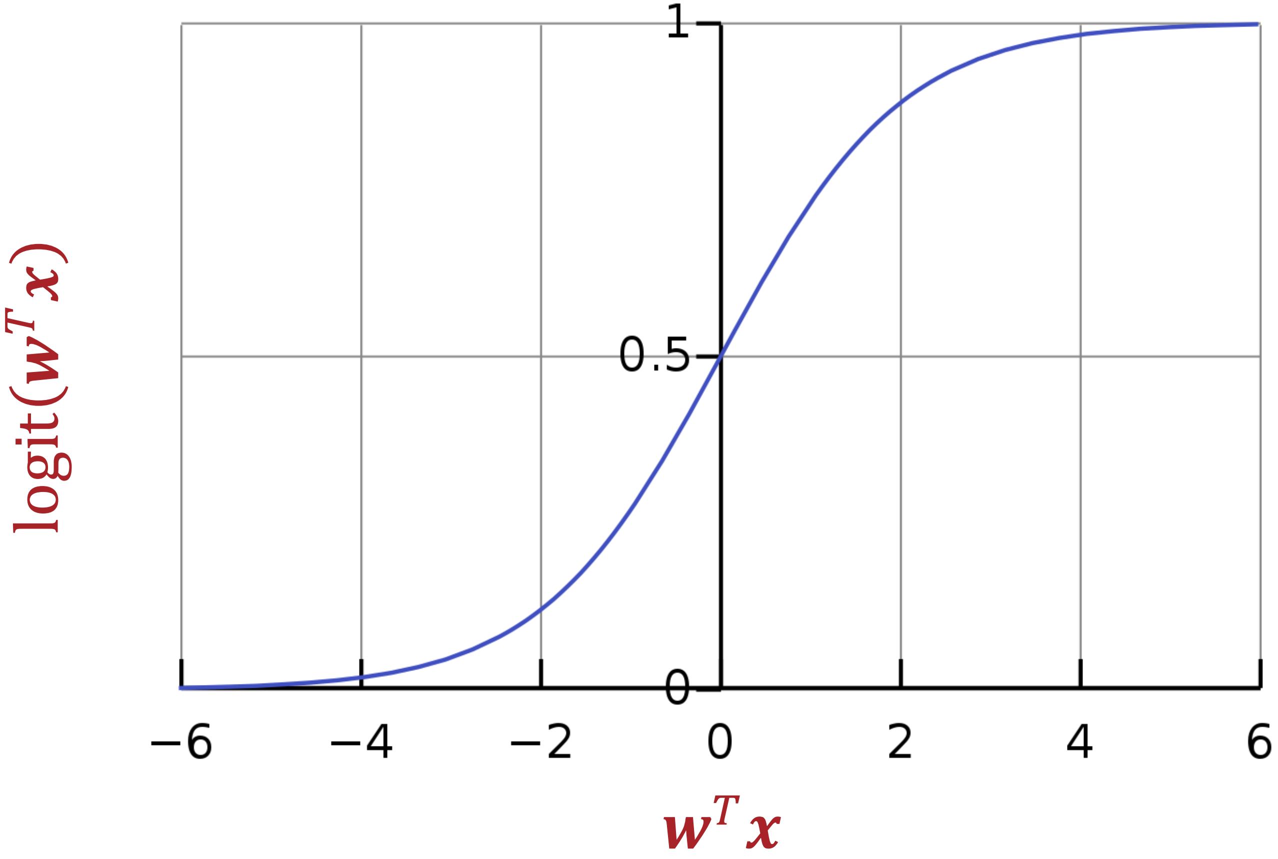
1. $P(Y = 0|\boldsymbol{x}) = 1 - P(Y = 1|\boldsymbol{x}) = \dfrac{1}{\exp(\boldsymbol{w}^T\boldsymbol{x}) + 1}$
2. $\dfrac{P(Y = 1|\boldsymbol{x})}{P(Y = 0|\boldsymbol{x})} = \exp(\boldsymbol{w}^T\boldsymbol{x}) \to \log\dfrac{P(Y = 1|\boldsymbol{x})}{P(Y = 0|\boldsymbol{x})} = \boldsymbol{w}^T\boldsymbol{x}$

Logistic Function



Source: https://en.wikipedia.org/wiki/Logistic_function#/media/File:Logistic-curve.svg

Why use the Logistic Function?



- Differentiable everywhere
- $\text{logit}: \mathbb{R} \to [0, 1]$
- The decision boundary is linear in $\boldsymbol{x}$!

Source: https://en.wikipedia.org/wiki/Logistic_function#/media/File:Logistic-curve.svg

Logistic Regression Decision Boundary

$$\hat{y} = \begin{cases} 1 \text{ if } P(Y = 1|\boldsymbol{x}) \geq \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

$$P(Y = 1|\boldsymbol{x}) = \text{logit}(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})} \geq \frac{1}{2}$$

$$2 \geq 1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})$$

$$1 \geq \exp(-\boldsymbol{w}^T\boldsymbol{x})$$

$$\log(1) \geq -\boldsymbol{w}^T\boldsymbol{x}$$

$$0 \leq \boldsymbol{w}^T\boldsymbol{x}$$

Logistic Regression Decision Boundary

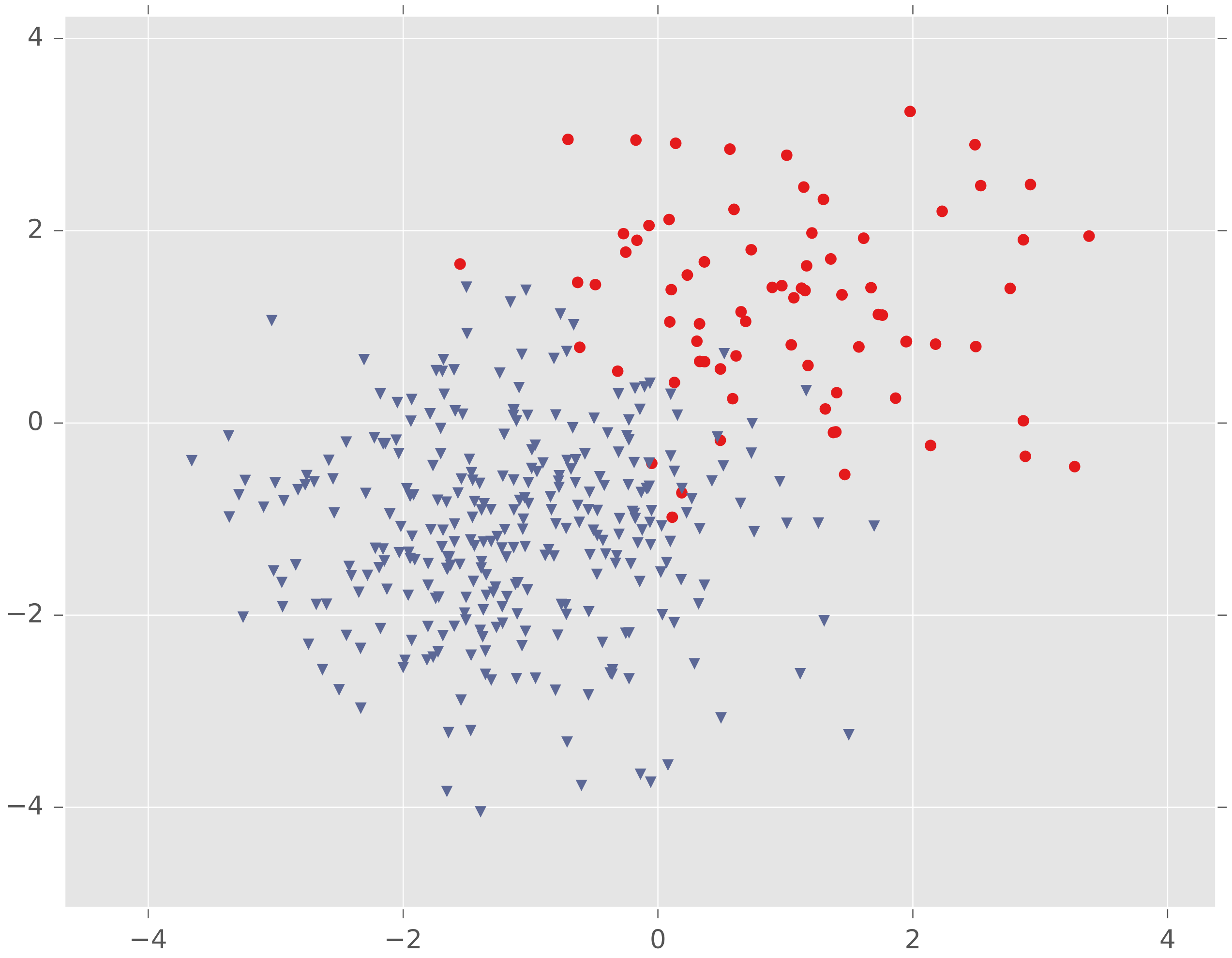


Figure courtesy of Matt Gormley

Logistic Regression Decision Boundary

Figure courtesy of Matt Gormley

Logistic Regression Decision Boundary

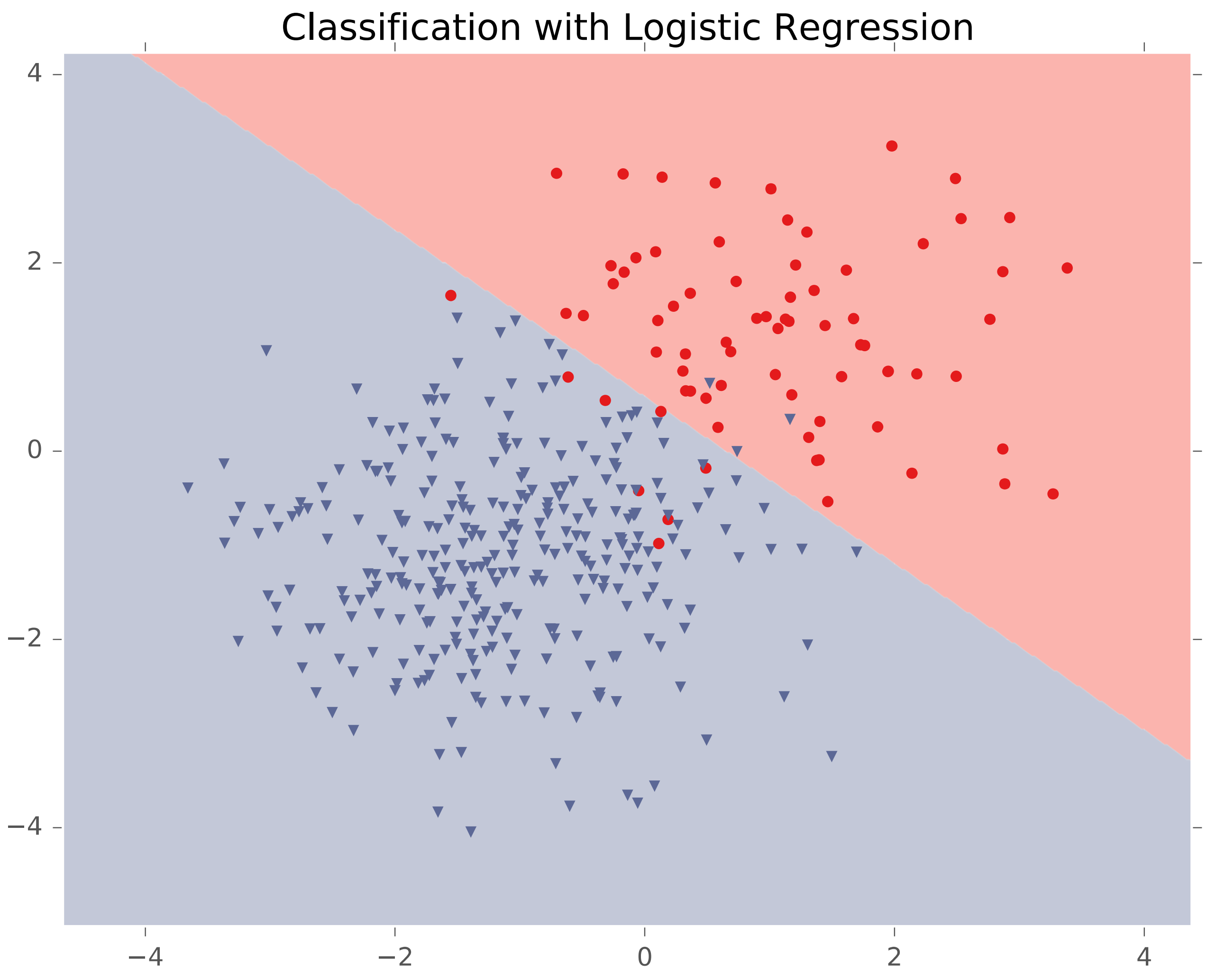


Figure courtesy of Matt Gormley

General Recipe for Machine Learning

- Define a model and model parameters

- Write down an objective function

- Optimize the objective w.r.t. the model parameters

Recipe for Logistic Regression

- Define a model and model parameters
  - Assume independent, identically distributed (iid) data
  - Assume $P(Y = 1|X) = \text{logit}(\boldsymbol{w}^T \boldsymbol{x})$
  - Parameters: $\boldsymbol{\theta} = [w_0, w_1, \ldots, w_D]$
- Write down an objective function
  - ~~Maximize the *conditional* log-likelihood~~
  - Minimize the negative conditional log-likelihood
- Optimize the objective w.r.t. the model parameters
  - ???

Setting the Parameters via Minimum Negative Conditional (log-)Likelihood Estimation (MCLE)

Find $\boldsymbol{\theta}$ that minimizes

$$\ell_{\mathcal{D}}(\boldsymbol{\theta}) = -\log P\left(y^{(1)}, \dots, y^{(N)} \middle| \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}, \boldsymbol{\theta}\right) = -\log \prod_{n=1}^{N} P\left(y^{(n)} \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)$$

$$= -\log \prod_{n=1}^{N} P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)^{y^{(n)}} \left(P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)\right)^{1-y^{(n)}}$$

$$= -\sum_{i=1}^{N} y^{(n)} \log P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right) + \left(1 - y^{(n)}\right) \log P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)$$

$$= -\sum_{i=1}^{N} y^{(n)} \log \frac{P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)}{P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)} + \log P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)$$

$$= -\sum_{i=1}^{N} y^{(n)} \boldsymbol{\theta}^{T} \boldsymbol{x}^{(n)} - \log\left(1 + \exp\left(\boldsymbol{\theta}^{T} \boldsymbol{x}^{(n)}\right)\right)$$

Setting the Parameters via MAP?

Stay tuned for regularization!

Find $\boldsymbol{\theta}$ that minimizes

$$\ell_{\mathcal{D}}(\boldsymbol{\theta}) = -\log P\left(y^{(1)}, \dots, y^{(N)} \middle| \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}, \boldsymbol{\theta}\right) = -\log \prod_{n=1}^{N} P\left(y^{(n)} \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)$$

$$= -\log \prod_{n=1}^{N} P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)^{y^{(n)}} \left(P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)\right)^{1-y^{(n)}}$$

$$= -\sum_{i=1}^{N} y^{(n)} \log P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right) + \left(1 - y^{(n)}\right) \log P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)$$

$$= -\sum_{i=1}^{N} y^{(n)} \log \frac{P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)}{P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)} + \log P\left(Y = 0 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right)$$

$$= -\sum_{i=1}^{N} y^{(n)} \boldsymbol{\theta}^T \boldsymbol{x}^{(n)} - \log\left(1 + \exp\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(n)}\right)\right)$$

Minimizing the Negative Conditional (log-)Likelihood

$$\ell_{\mathcal{D}}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} y^{(n)}\boldsymbol{\theta}^T\boldsymbol{x}^{(n)} - \log\left(1 + \exp\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(n)}\right)\right)$$

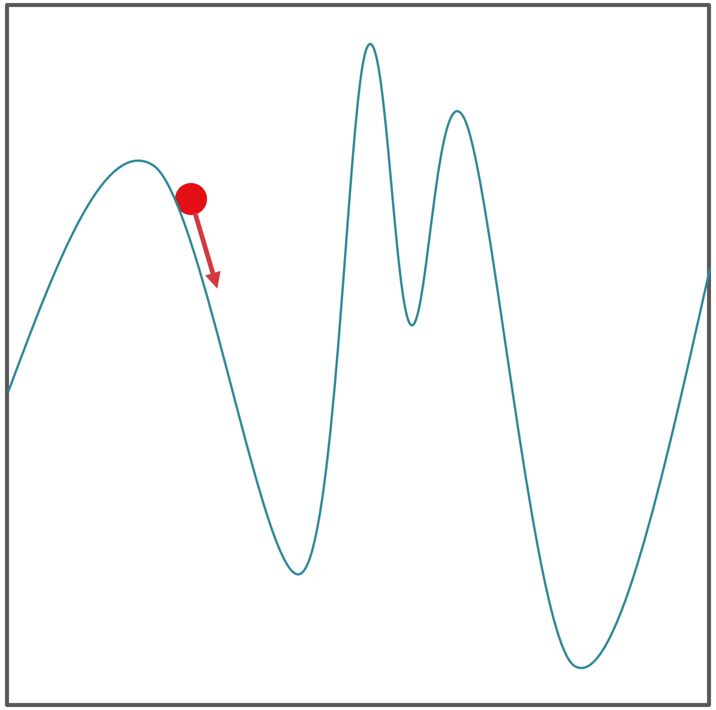
$$\nabla_{\boldsymbol{\theta}}\ell_{\mathcal{D}}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} y^{(n)}\nabla_{\boldsymbol{\theta}}\boldsymbol{\theta}^T\boldsymbol{x}^{(n)} - \nabla_{\boldsymbol{\theta}}\log\left(1 + \exp\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(n)}\right)\right)$$

$$= -\sum_{n=1}^{N} y^{(n)}\boldsymbol{x}^{(n)} - \frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(n)}\right)}{1 + \exp\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(n)}\right)}\boldsymbol{x}^{(n)}$$

$$= \sum_{n=1}^{N} \boldsymbol{x}^{(n)}\left(P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{\theta}\right) - y^{(n)}\right)$$

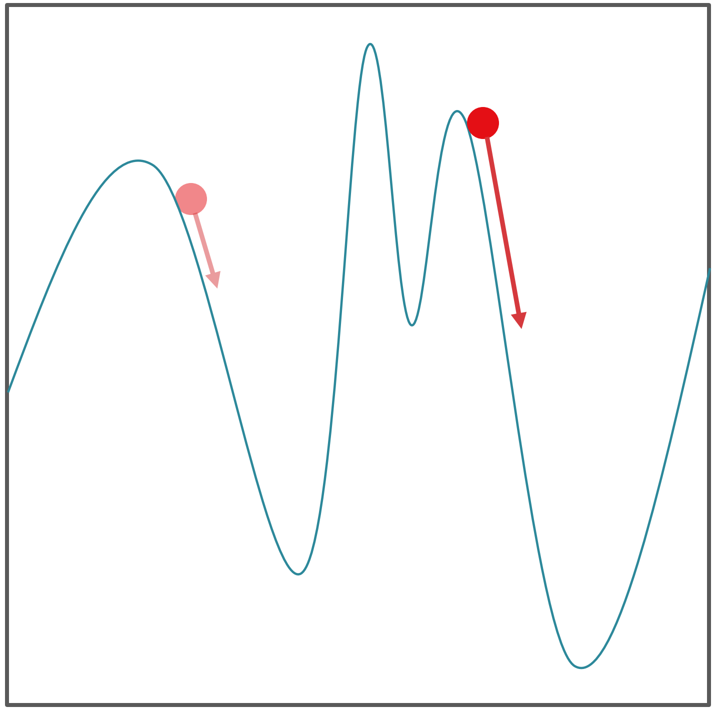
Gradient Descent: Intuition

- An iterative method for minimizing functions
- Requires the gradient to exist everywhere



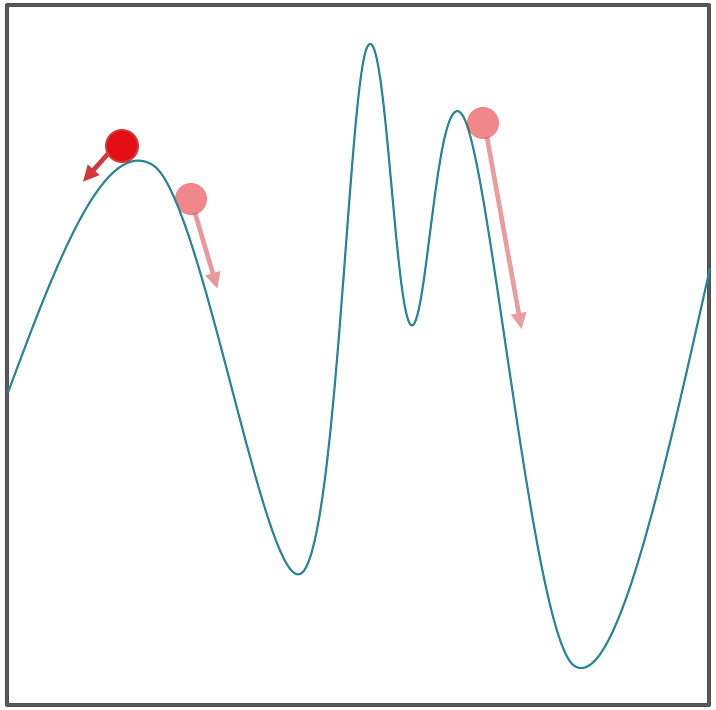
Gradient Descent: Intuition

- An iterative method for minimizing functions
- Requires the gradient to exist everywhere



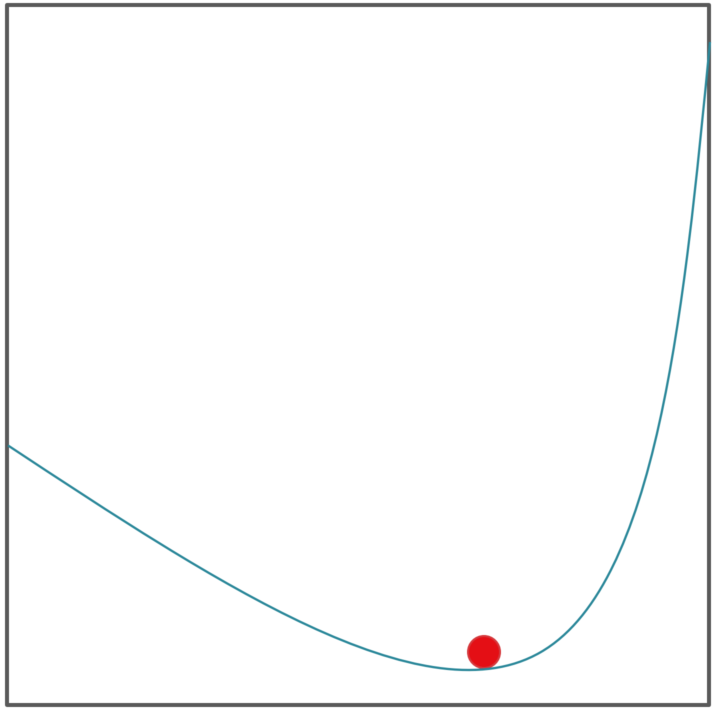
Gradient Descent: Intuition

- An iterative method for minimizing functions
- Requires the gradient to exist everywhere



Gradient Descent: Intuition

- An iterative method for minimizing functions
- Requires the gradient to exist everywhere



- Good news: the negative conditional log-likelihood is *convex*!

Gradient Descent: Step Direction

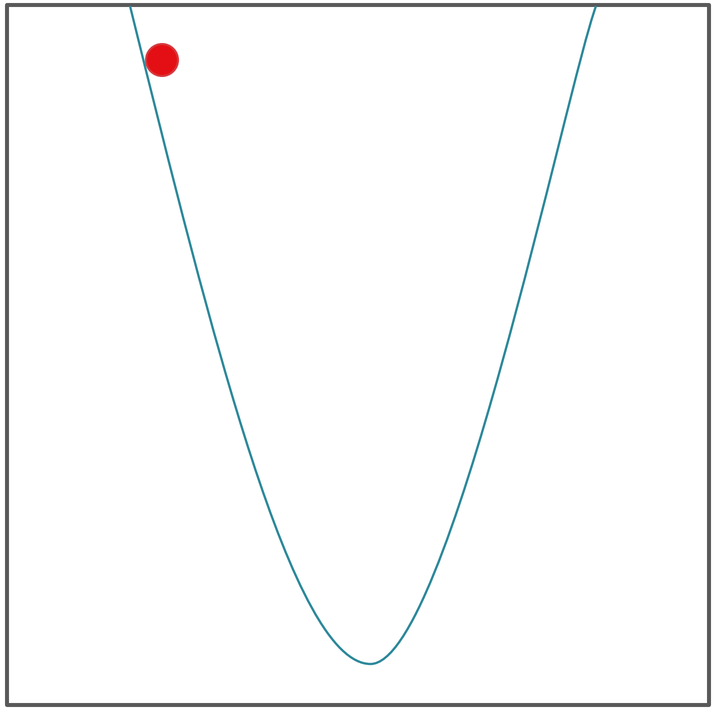
- Suppose the current parameter vector is $\boldsymbol{\theta}^{(t)}$
- Move some distance, η, in the "most downhill" direction, $\widehat{\boldsymbol{v}}$:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta \widehat{\boldsymbol{v}}$$

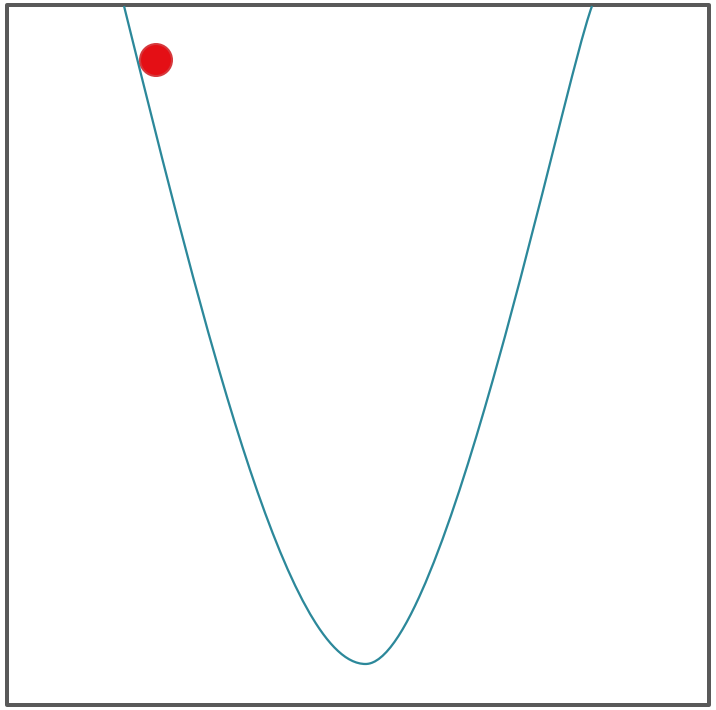
- The gradient points in the direction of steepest *increase* …
- … so $\widehat{\boldsymbol{v}}$ is a unit vector pointing in the opposite direction:

$$\widehat{\boldsymbol{v}}^{(t)} = -\frac{\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)}{\left\|\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|}$$

Gradient Descent: Step Size

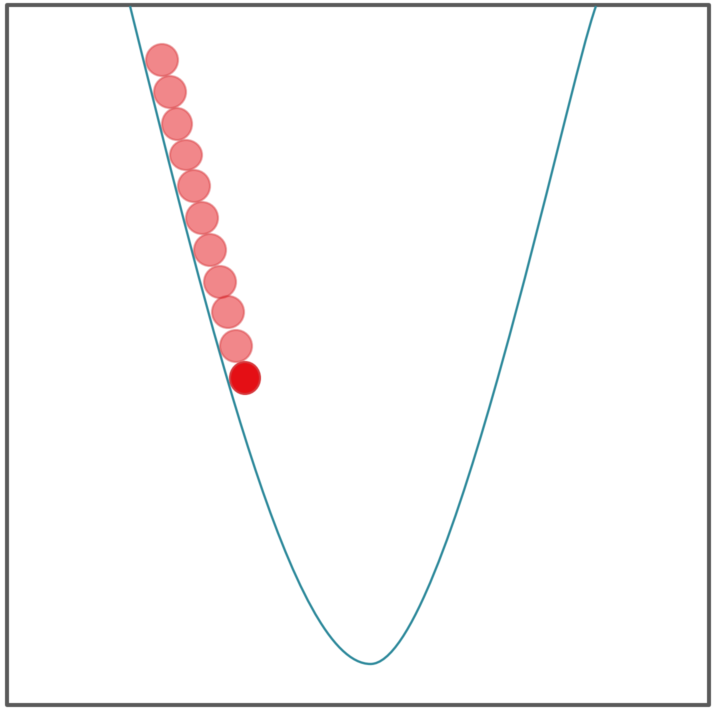


Small η

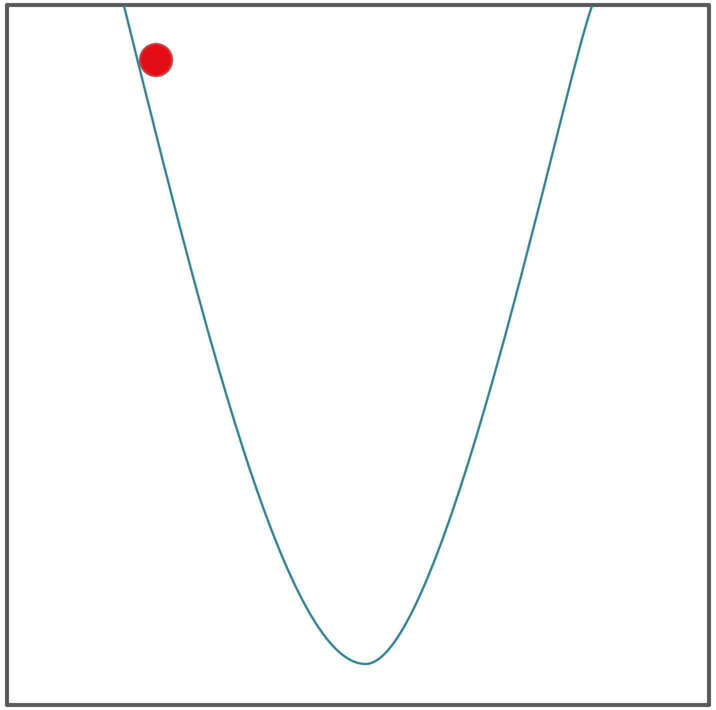


Large η

Gradient Descent: Step Size

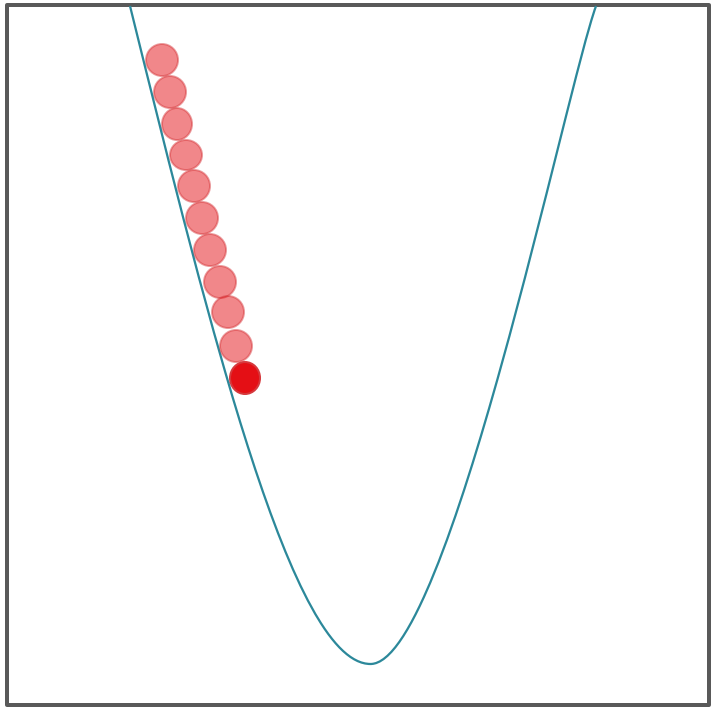


Small η

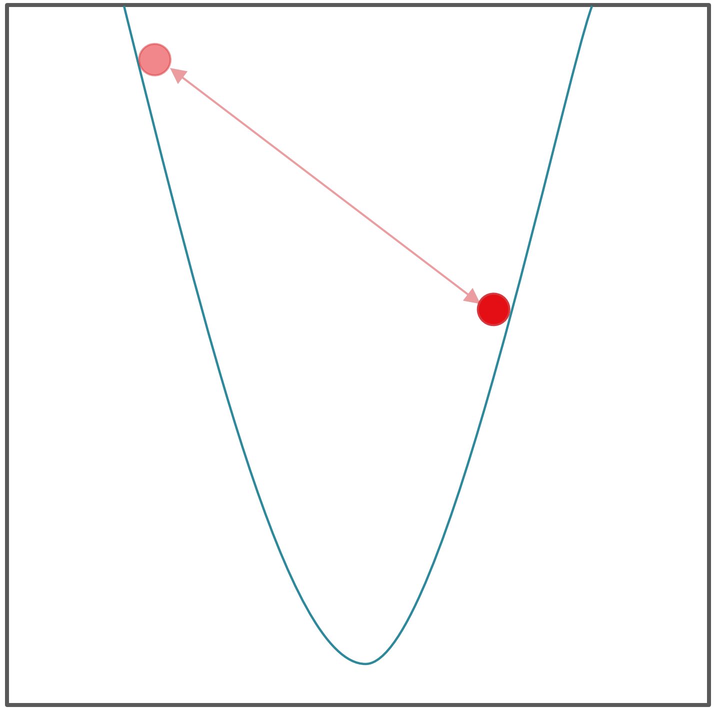


Large η

Gradient Descent: Step Size



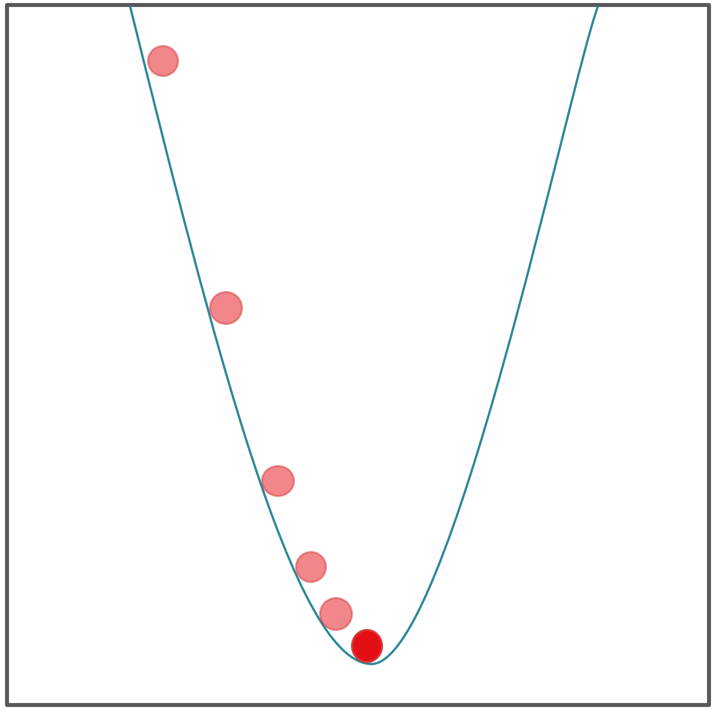
Small η



Large η

Gradient Descent: Step Size

- Use a variable $\eta^{(t)}$ instead of a fixed η!



- Set $\eta^{(t)} = \eta^{(0)} \left\| \nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}} \left(\boldsymbol{\theta}^{(t)} \right) \right\|$
- $\left\| \nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}} \left(\boldsymbol{\theta}^{(t)} \right) \right\|$ decreases as $\ell_{\mathcal{D}}$ approaches its minimum $\rightarrow \eta^{(t)}$ (hopefully) decreases over time

Gradient Descent

- $\hat{\boldsymbol{v}}^{(t)} = -\frac{\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)}{\left\|\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|}$

- $\eta^{(t)} = \eta^{(0)}\left\|\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|$

- $$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + \eta^{(t)}\hat{\boldsymbol{v}}^{(t)} \\ &= \boldsymbol{\theta}^{(t)} + \left(\eta^{(0)}\left\|\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|\right)\left(-\frac{\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)}{\left\|\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)\right\|}\right) \\ &= \boldsymbol{\theta}^{(t)} - \eta^{(0)}\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)\end{aligned}$$

Gradient Descent

- Input: $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}, \eta^{(0)}$

1. Initialize the parameters $\boldsymbol{\theta}^{(0)}$ and set $t = 0$
2. While TERMINATION CRITERION is not satisfied
   a. Compute the gradient:
      $\nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)$
   b. Update $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta^{(0)} \nabla_{\boldsymbol{\theta}} \ell_{\mathcal{D}}\left(\boldsymbol{\theta}^{(t)}\right)$
   c. Increment t: $t \leftarrow t + 1$

- Output: $\boldsymbol{\theta}^{(t)}$

Key Takeaways

- Logistic regression
  - Logistic function induces a linear decision boundary
  - Conditional likelihood maximization
- Gradient descent
  - Effect of step size
  - Termination criteria