Gradient Descent

Minimizing loss functions

Terence Parr
MSDS program
University of San Francisco

Minimizing the loss: How we train (many) models

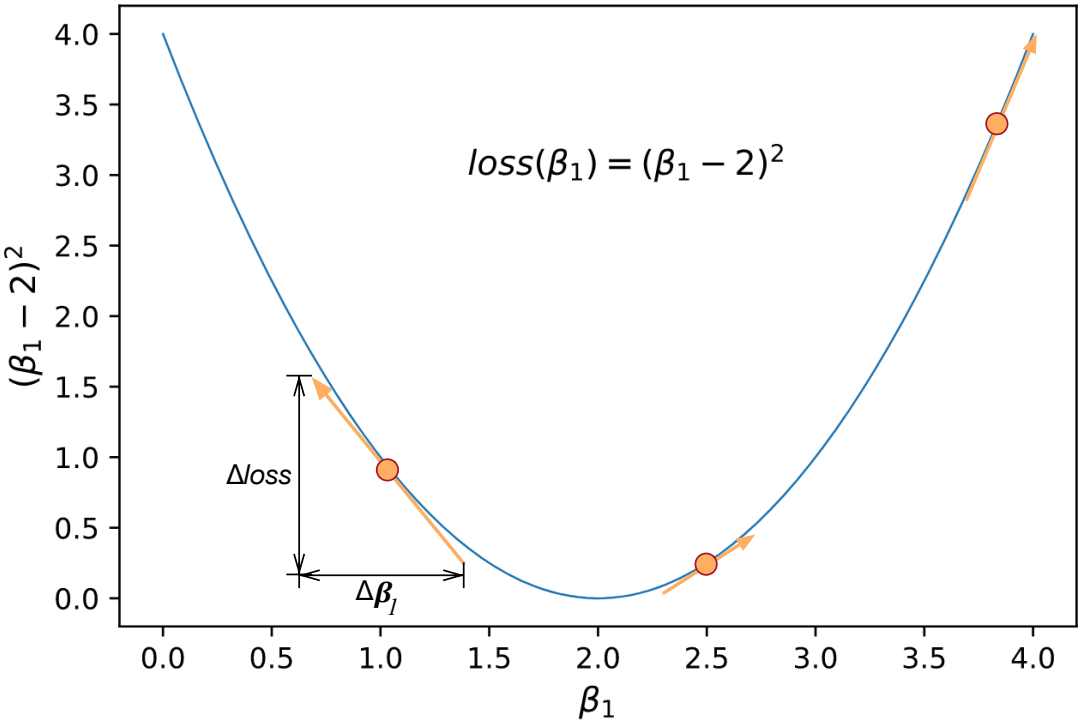
- We need a way to find β such that: $\arg\min_{\beta} \mathscr{L}(\beta)$
- Could try random β vectors and choose the β with lowest loss
- Or, better yet, choose a random β and then tweak with some $\Delta\beta$ in the downhill loss direction until any tweak would increase loss

$$\beta^{(t+1)} = \beta^{(t)} + \Delta\beta^{(t)}$$

- We use information about the loss function in the neighborhood of current β to decide which direction shifts towards smaller loss
- When loss goes up or doesn't change, we're done

How do we pick a direction to move?

- Use information (*gradient*) from loss function in vicinity of current β_1



- Derivative/slope of loss(β_1) is 2(β_1-2), which points in direction of increased loss (up)
- What is derivative of loss at β_1=2?
- Direction of min loss is opposite/negative of derivative
- Derivative also has magnitude, which makes us go faster when slope is steeper
- **How to move**: $\beta_1 = \beta_1 - slope$

Taking steps in right direction

- Direction of min loss is opposite of derivative so let's step in negative of derivative and scale it with a learning rate η:

$$\beta^{(t+1)} = \beta^{(t)} - \eta \frac{d}{d\beta} \mathscr{L}(\beta^{(t)})$$

```
while not_tired:
    b = b - rate * gradient(b)
```

Python gradient descent implementation

- First define a simple loss function and its gradient:

```
def f(b) : return (b-2)**2
def gradient(b): return 2*(b-2)
```

- Then, pick a random starting point and pick a learning rate

```
b = np.random.uniform(0,4)
rate = .2
```

- Loop until we've made some progress or until gradient(b)==0

```
for t in range(10): # for awhile
    b = b - rate * gradient(b)
```

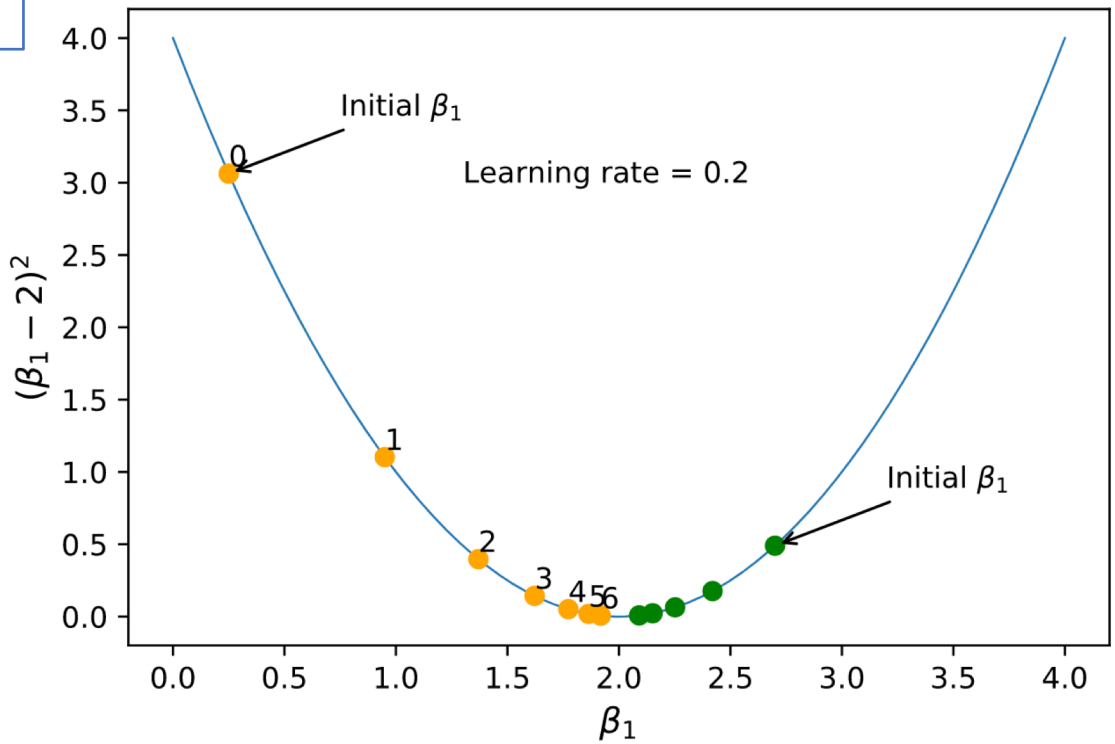

See https://github.com/parrt/msds621/blob/master/notebooks/linear-models/viz-gradient-descent.ipynb

Sample 1D gradient descent run

```
for t in range(7):
    b = b - 0.2 * gradient(b)
```

	beta_1	loss
0	0.055312	3.781813
1	0.833187	1.361453
2	1.299912	0.490123
3	1.579947	0.176444
4	1.747968	0.063520
5	1.848781	0.022867
6	1.909269	0.008232
7	1.945561	0.002964

Notice β_1 accelerates and then slows down. Why?



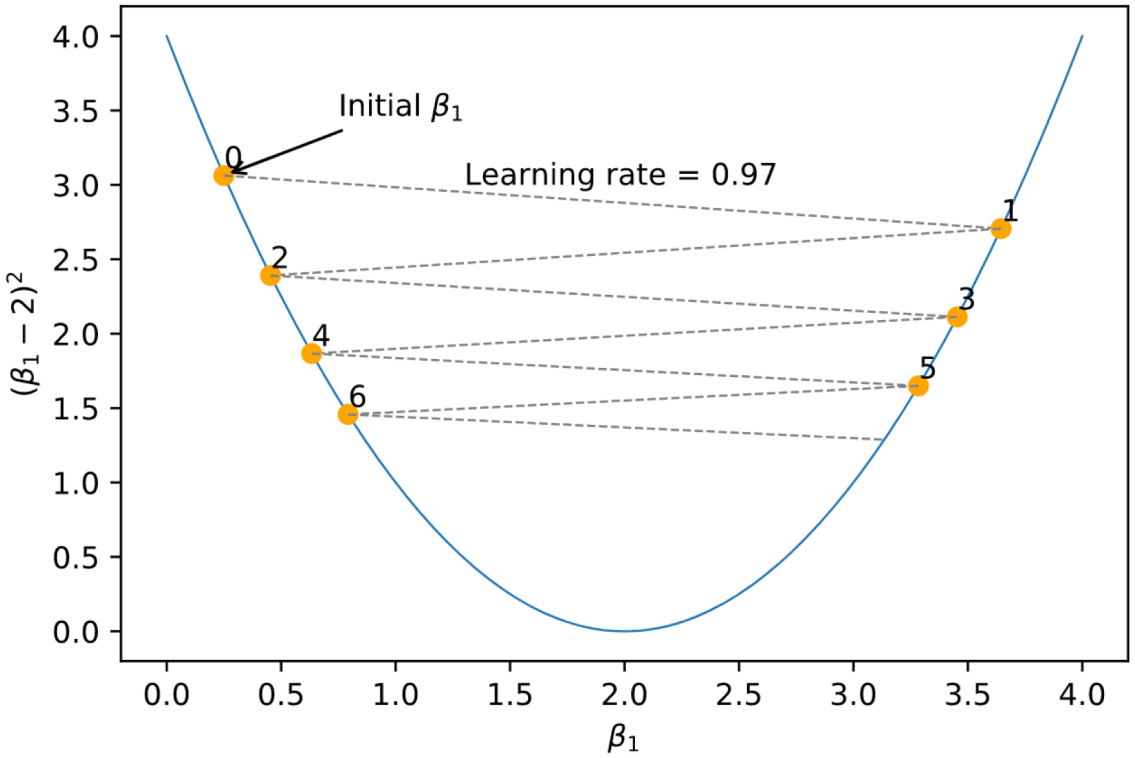
See https://github.com/parrt/msds621/blob/master/notebooks/linear-models/viz-gradient-descent.ipynb

1D function minimization in action

- See “Animation 1D quadratic example” in notebook:
  https://github.com/parrt/msds621/blob/master/notebooks/linear-models/viz-gradient-descent.ipynb

What if we crank up learning rate?

- β_1 oscillates across valley
- Picking learning rate is trial and error for our purposes but small like η=.00001 is a reasonable guess to start out
- If too small, we don't make progress to min

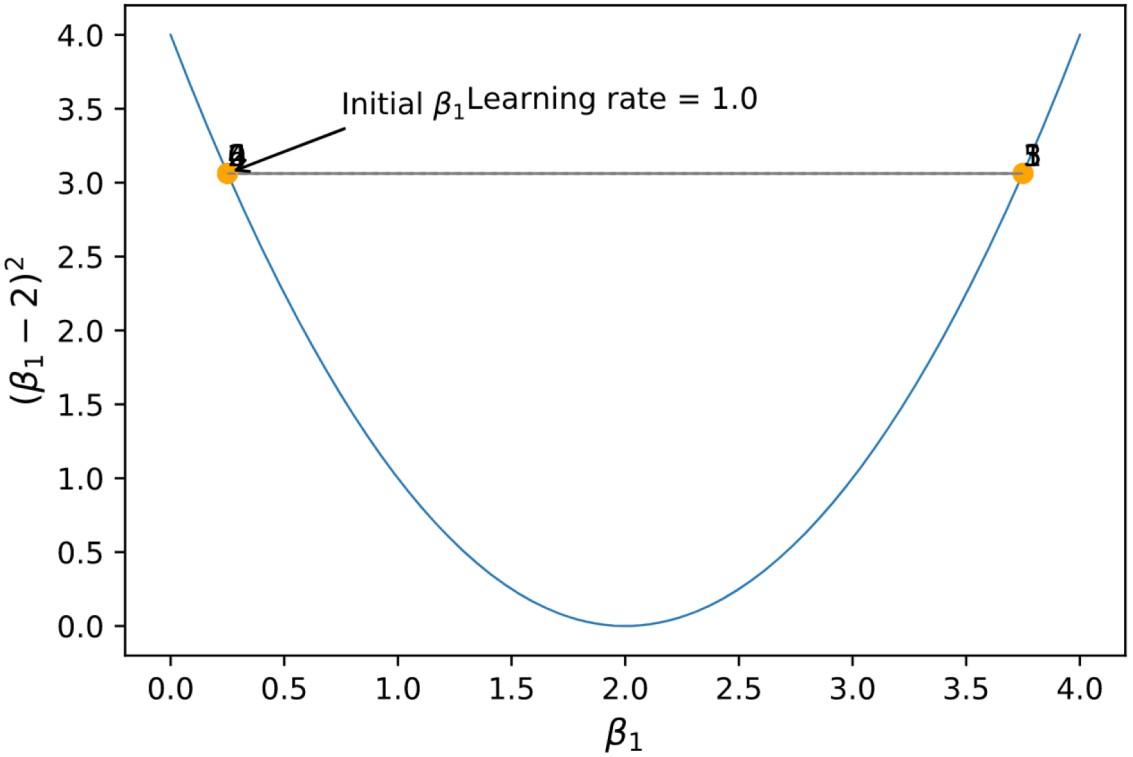


What if learning rate is really too high?

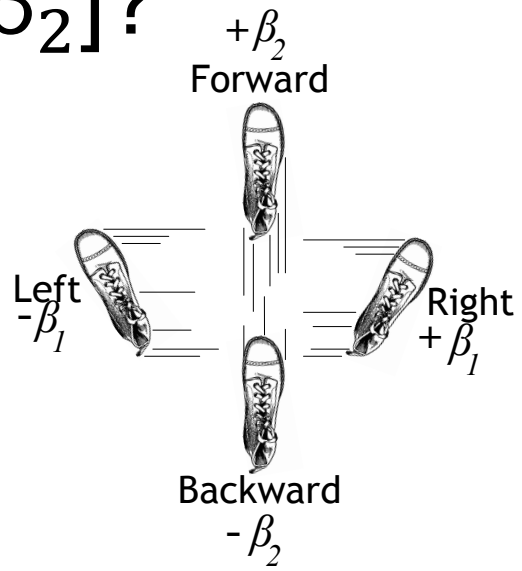
- We get nowhere:

	beta_1	loss
0	0.495633	2.263119
1	3.504367	2.263119
2	0.495633	2.263119
3	3.504367	2.263119
4	0.495633	2.263119

- It can even diverge, exploding β_1

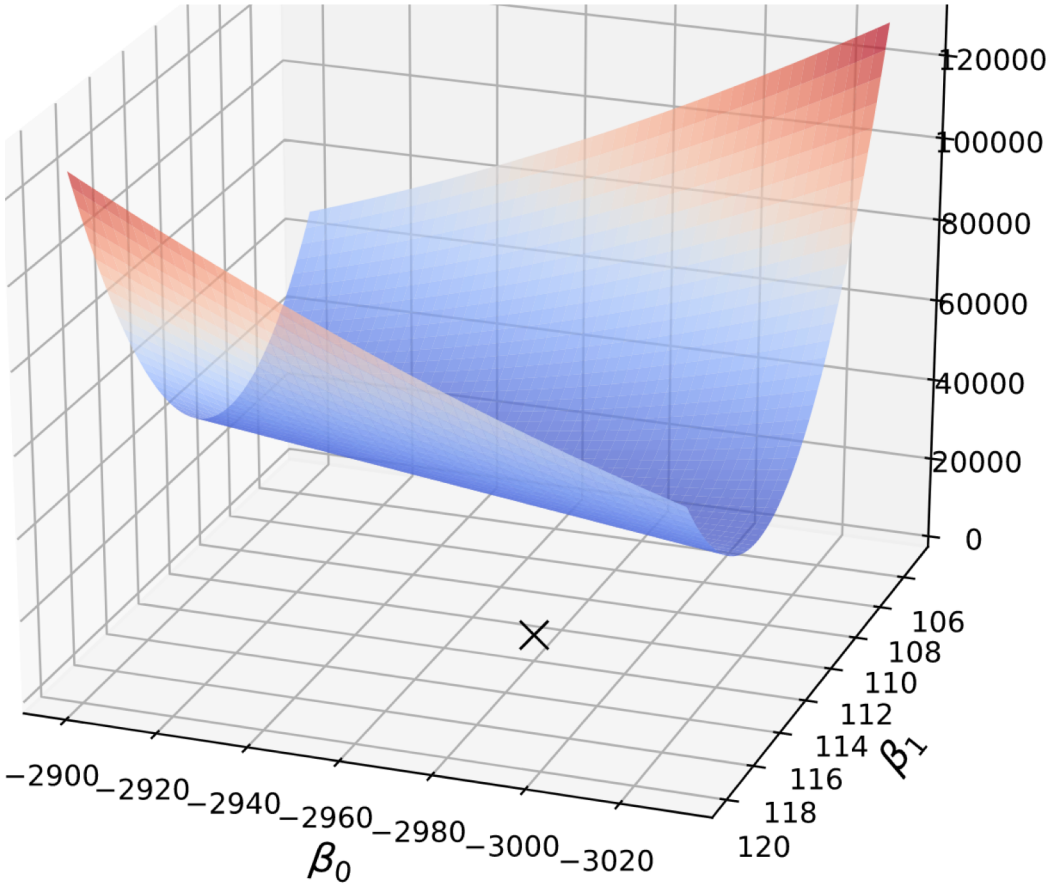


What happens in 2D for $\beta = [\beta_1, \beta_2]$?



- Imagine you're stuck on a mountain in the dark and need to get to the bottom
- Take steps to left, right, forward, backward or at an angle to minimize the "elevation function"
- Treat each direction separately, then combine them to obtain the best step direction
- Each direction's slope is a *partial derivative* and, combined, are the *gradient* vector

Loss function: 1-var regr. w/2 coeff (β_0, β_1)

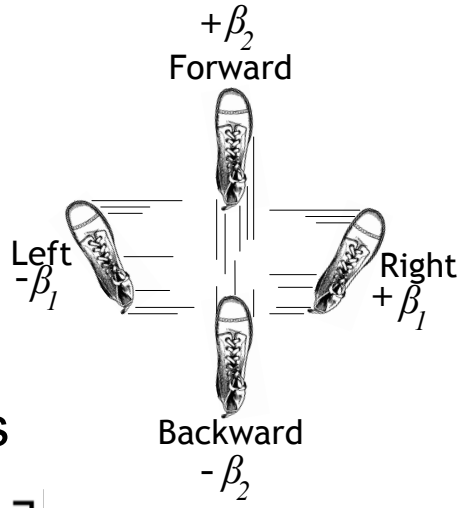


- Shallow in β_0 dir
- Steep in β_1 dir
- This plot show loss for non-standardized variables so a unit change in β_0 doesn't change loss nearly as much for β_1

Notation and finite difference approximation

- "Rise over run" is the derivative/slope of $f(x)$ at x:

$$\frac{d}{dx}f(x) = \frac{\partial}{\partial x}f(x) \approx \frac{f(x+h) - f(x)}{h}$$



- Gradient of p>1 $\boldsymbol{x}$ vector has p partial derivative entries

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{x_1}f(\mathbf{x}) \\ \frac{\partial}{x_2}f(\mathbf{x}) \end{bmatrix} \approx \begin{bmatrix} \frac{f([x_1+h, x_2]) - f(\mathbf{x})}{h} \\ \frac{f([x_1, x_2+h]) - f(\mathbf{x})}{h} \end{bmatrix}$$

- The partial derivative is just the slope in 1 dir, holding others constant

See https://explained.ai/matrix-calculus/index.html

General gradient descent

- Partial derivative is rate of change in one direction: $\frac{\partial}{\partial \beta_i} \mathscr{L}(\beta)$
- Combining partial derivatives into vector gives the *gradient:* ∇_β
- Gradient points in direction of increased loss, so must go in negative gradient direction to decrease loss as before:

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla_\beta \mathscr{L}(\beta^{(t)})$$ where η is a learning rate

- Gradients have magnitude and direction
- E.g., β=[-1,2] means take a step to left, but bigger step forward
- Take that single step: $\beta = \beta - \eta^*$ [-1, 2]
- In each direction, the partial derivative of loss function is 0 when flat
- When gradient vector = 0 vector, we're at min loss; choose that β

Update equation needs gradient:

$$\beta^{(t+1)} = \beta^{(t)} - \nabla_\beta \mathscr{L}(\beta^{(t)})$$

Gradient of $\mathscr{L}(\beta) = (\mathbf{y} - \mathbf{X}'\beta) \cdot (\mathbf{y} - \mathbf{X}'\beta)$ for lin regression is

$$\nabla_\beta \mathscr{L}(\beta) = -2\mathbf{X}'^T(\mathbf{y} - \mathbf{X}'\beta)$$

So update equation becomes (adding *learning rate* η):

$$\beta^{(t+1)} = \beta^{(t)} + \eta\mathbf{X}'^T(\mathbf{y} - \mathbf{X}'\beta^{(t)})$$

η scales the step we take each at each step (fold 2 into η)

Simplest gradient descent algorithm

Algorithm: *basic_minimize*($\mathbf{X}$, $\mathbf{y}$, $\mathscr{L}$, η) **returns** coefficients $\vec{b}$

Let $\vec{b} \sim 2N(0,1) - 1$ (*init b with random p + 1-sized vector with elements in [-1,1)*)

$\mathbf{X}' = (\vec{\mathbf{1}}, \mathbf{X})$ (*Add first column of 1s to data*) except for L1/L2 regression

repeat

$\nabla_{\vec{b}} = -\mathbf{X}'^T(\mathbf{y} - \mathbf{X}'\vec{b})$

$\vec{b} = \vec{b} - \eta\nabla_{\vec{b}}$ new direction

until $|(\mathscr{L}(\vec{b}) - \mathscr{L}(\vec{b}_{prev}))| < precision$;

return $\vec{b}$

Adding momentum to particle update

- Reinforce movement in same direction; add fraction of previous dir

Algorithm: *minimize*$(\mathbf{X}, \mathbf{y}, \mathscr{L}, \eta, \gamma)$ **returns** coefficients $\vec{b}$

Let $\vec{b} \sim 2N(0,1) - 1$ *(random $p+1$-sized vector with elements in [-1,1))*

$\mathbf{X}' = (\vec{\mathbf{1}}, \mathbf{X})$ *(Add first column of 1s)* except for L1/L2 regression

repeat

$\nabla_{\vec{b}} = -\mathbf{X}'^T(\mathbf{y} - \mathbf{X}'\vec{b})$

old direction new direction

$\vec{v} = \gamma\vec{v} + \eta\nabla_{\vec{b}}$

$\vec{b} = \vec{b} - \vec{v}$

γ is a new hyper parameter

until $|(\mathscr{L}(\vec{b}) - \mathscr{L}(\vec{b}_{prev}))| < precision;$

return $\vec{b}$

General Adagrad gradient descent

- Single learning rate for all dimensions is brutally slow
- Imagine long shallow valley with steep walls
- η small enough for steep walls is way too slow for other, shallow dimension

Algorithm: *adagrad_minimize*($\mathbf{X}$, $\mathbf{y}$, $\mathscr{L}$, $\nabla\mathscr{L}$, η, ϵ=1e-5) **returns** coefficients $\vec{b}$

Let $\vec{b} \sim 2N(0,1) - 1$ (*random $p+1$-sized vector with elements in [-1,1)*)

$h = \vec{0}$ (*$p+1$-sized sum of squared gradient history*)

$\mathbf{X}' = (\vec{1}, \mathbf{X})$ (*Add first column of 1s*) except for L1/L2 regression

repeat

$\vec{h} \mathrel{+}= \nabla\mathscr{L} \otimes \nabla\mathscr{L}$ (*track sum of squared partials, use element-wise product*)

$\vec{b} = \vec{b} - \eta * \frac{\nabla\mathscr{L}}{(\sqrt{\vec{h}+\epsilon})}$ ← adjust w/update per β_i; low h(istory) for β_i increases its learning rate (ϵ avoids divide by 0)

until $|(\mathscr{L}(\vec{b}) - \mathscr{L}(\vec{b}_{prev}))| < precision$;

return $\vec{b}$

See http://cs231n.github.io/neural-networks-3/#ada

Loss, gradient functions for minimization

- Linear regression

$$\mathscr{L}(\beta) = (\mathbf{y} - \mathbf{X}'\beta) \cdot (\mathbf{y} - \mathbf{X}'\beta)$$

$$\nabla_\beta \mathscr{L}(\beta) = -2\mathbf{X}'^T(\mathbf{y} - \mathbf{X}'\beta)$$

- Logistic regression

$$\mathscr{L}(\beta) = \sum_{i=1}^{n} \left\{ y^{(i)} \mathbf{x}'^{(i)} \beta - log(1 + e^{\mathbf{x}'\beta}) \right\}$$

$$\nabla_\beta \mathscr{L}(\beta) = -\mathbf{X}'^T(\mathbf{y} - \sigma(\mathbf{X}' \cdot \beta))$$

L1, L2 regression loss, gradient functions

- L2 (Ridge); 0-center x_i then β_0 = mean(**y**), find $\beta_{1..p}$ via:

$$\mathscr{L}(\beta) = (\mathbf{y} - \mathbf{X}\beta) \cdot (\mathbf{y} - \mathbf{X}\beta) + \lambda\beta \cdot \beta$$

$$\nabla_\beta \mathscr{L}(\beta) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) + 2\lambda\beta$$

- L1 (Lasso); 0-center x_i then β_0 = mean(**y**), find $\beta_{1..p}$ via:

$$\mathscr{L}(\beta) = (\mathbf{y} - \mathbf{X}\beta) \cdot (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\nabla_\beta \mathscr{L}(\beta) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \, \text{sign}(\beta)$$

L1 logistic loss, gradient functions

- Must compute β_0 differently; partial β_0 is a function of β_0

$$\frac{\partial}{\partial \beta_0} \mathscr{L}(\beta, \lambda) = mean(\mathbf{y} - \sigma(\mathbf{X}' \cdot \beta))$$

- Other β_i are functions of β_0 but not within the penalty term

$$\nabla_{\beta_{1..p}} \mathscr{L}(\beta, \lambda) = \frac{1}{n} \left\{ \mathbf{X}^T (\mathbf{y} - \sigma(\mathbf{X}' \cdot \beta')) - \lambda \, \text{sign}(\beta) \right\}$$

- Combine to get full gradient vector

L1 logistic gradient algorithm

(review from regularization lecture)

Algorithm: $L1NegLogLikelihood(\mathbf{X}', \mathbf{y}, \beta')$

$err = \mathbf{y} - \sigma(\mathbf{X}' \cdot \beta')$ $(error\ vector \text{ is } n \times 1\ column\ vector)$

$\frac{\partial}{\partial \beta_0} = mean(err)$ $(usual\ log\text{-}likelihood\ gradient;\ use\ current\ \beta')$

$r = \lambda \operatorname{sign}(\beta')$ $(regularization\ term\ p+1 \times 1\ column\ vector)$

$r[0] = 0$ $(kill\ \beta_0\ position\ but\ keep\ as\ p+1 \times 1\ vector)$

$\nabla = \frac{1}{n}\left\{\mathbf{X}'^T err - r\right\}$

$$\textbf{return } -\begin{bmatrix} \frac{\partial}{\partial \beta_0} \\ \nabla_1 \\ \vdots \\ \nabla_p \end{bmatrix}$$

Key takeaways

- Move towards lower loss; consider each direction separately
- Slope in direction β_i is partial derivative: $\frac{\partial}{\partial \beta_i} \mathscr{L}(\beta)$
- Gradient is p or p+1 dimensional vector of partial derivatives
- Gradients point "upwards" towards higher cost/loss
- Coefficients should therefore move in opposite direction of gradient
- Gradient is the 0 vector at the minimum loss; i.e., flat
- Can stop optimizing when gradient is close to 0 vector or when $\beta_i^{(t+1)}$ is very close to $\beta_i^{(t)}$ or after fixed number of iterations

More key takeaways

- Coefficient update equation:

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla_{\beta} \mathscr{L}(\beta^{(t)})$$ where η is a learning rate

- If η is "small enough," $\beta_i^{(t+1)}$ will converge to a solution vector (maybe slowly)
- If too big, will bounce back and forth across valleys or diverge
- Adagrad
  - Single learning rate too slow; need a rate per dimension
  - Increases update step size for dimensions with shallow slopes historically
  - Slows down across all dimensions over time as history sum *h* gets bigger

$$\vec{b} = \vec{b} - \eta * \frac{\nabla \mathscr{L}}{(\sqrt{\vec{h}+\epsilon})}$$

- L1, L2 linear regression doesn't optimize β_0, it's just mean(**y**), if we 0-center x_i
- L1, L2 logistic regression optimizes $\beta_{0..p}$ but β_0 differently than $\beta_{1..p}$