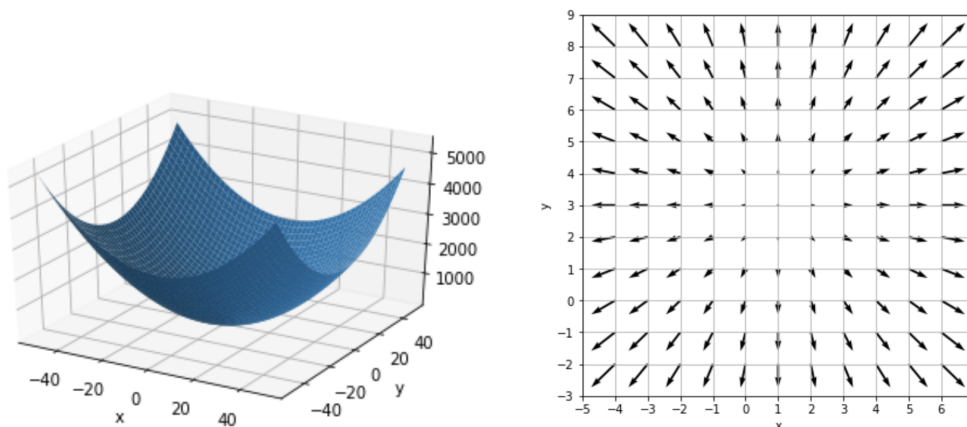


## Discussion #11

Name: Raguvir Kunani

## Gradients

1. On the left is a 3D plot of  $f(x, y) = (x - 1)^2 + (y - 3)^2$ . On the right is a plot of its **gradient field**. Note that the arrows show the relative magnitudes of the gradient vector.



- (a) From the visualization, what do you think is the minimal value of this function and where does it occur?

$$x=1, y=3$$

- (b) Calculate the gradient  $\nabla f = \left[ \frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y} \right]^T$ .  $f(x, y) = (x-1)^2 + (y-3)^2$

$$\frac{\partial f}{\partial x} = 2(x-1)$$

$$\frac{\partial f}{\partial y} = 2(y-3)$$

$$\Rightarrow \nabla f = \begin{bmatrix} 2(x-1) \\ 2(y-3) \end{bmatrix}$$

- (c) When  $\nabla f = 0$ , what are the values of  $x$  and  $y$ ?

$\nabla f = 0$  at the minimum of  $f$ ; from (a) this is  $x=1, y=3$

Algebraically,  $\nabla f = \vec{0} \Rightarrow \begin{bmatrix} 2(x-1) \\ 2(y-3) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{matrix} x=1 \\ y=3 \end{matrix}$

General Gradient Update:

$$\beta^{t+1} = \beta^t - \alpha \nabla_{\beta} L(\beta, \vec{x}, \vec{y}) \Big|_{\beta = \beta^t}$$

## Gradient Descent Algorithm

2. Given the following loss function and  $\mathbf{x} = (x_i)_{i=1}^n$ ,  $\mathbf{y} = (y_i)_{i=1}^n$ ,  $\beta^t$ , explicitly write out the update equation for  $\beta^{t+1}$  in terms of  $x_i$ ,  $y_i$ ,  $\beta^t$ , and  $\alpha$ , where  $\alpha$  is the step size.

$$L(\beta, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n (\beta^2 x_i^2 - \log(y_i))$$

→  $\beta$  scalar  $\Rightarrow \nabla_{\beta} L(\beta, \mathbf{x}, \mathbf{y}) = \frac{\partial L(\beta, \mathbf{x}, \mathbf{y})}{\partial \beta}$

$$\begin{aligned} \frac{\partial L(\beta, \mathbf{x}, \mathbf{y})}{\partial \beta} &= \frac{\partial}{\partial \beta} \left[ \frac{1}{n} \sum_{i=1}^n (\beta^2 x_i^2 - \log(y_i)) \right] = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta} [\beta^2 x_i^2 - \log(y_i)] \\ &= \frac{1}{n} \sum_{i=1}^n 2\beta x_i^2 = \frac{2\beta}{n} \sum_{i=1}^n x_i^2 \end{aligned}$$

Plug into update formula:  $\beta^{t+1} = \beta^t - \alpha \frac{2\beta}{n} \sum_{i=1}^n x_i^2 \Big|_{\beta = \beta^t} \Rightarrow \boxed{\beta^{t+1} = \beta^t - \alpha \frac{2\beta^t}{n} \sum_{i=1}^n x_i^2}$

## Convexity

3. Convexity allows optimization problems to be solved more efficiently and for global optimums to be realized. Mainly, it gives us a nice way to minimize loss (i.e. gradient descent). There are three ways to informally define convexity.
- Walking in a straight line between points on the function keeps you above the function. This works for any function.
  - The tangent line at any point lies below the function (globally). The function must be differentiable.
  - The second derivative is non-negative everywhere (aka "concave up" everywhere). The function must be twice differentiable.

- (a) Is the function described in question 1 convex? Make an argument visually.

Yes; the line between any 2 points on the bowl goes above the bowl

- (b) Find a counterexample for the claim that the composition of two convex functions is also convex.  $h = g(f(x))$

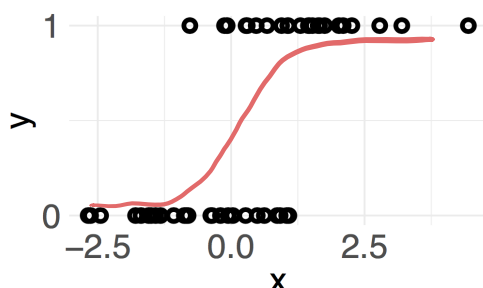
$$g(x) = -x \quad f(x) = x^2 \quad \Rightarrow \quad h(x) = -f(x) = -x^2$$

↓  
CONCAVE!

## Logistic Regression

The next two questions refer to a binary classification problem with a single feature  $x$ .

4. Based on the scatter plot of the data below, draw a reasonable approximation of the logistic regression probability estimates for  $\mathbb{P}(Y = 1 | x)$ .



The approach for this is we want our logistic model (the  $\sigma$  function) to have high probabilities where our data is 1 and low probabilities where our data is 0.

5. You have a classification data set consisting of two  $(x, y)$  pairs  $(1, 0)$  and  $(-1, 1)$ .

The covariate vector  $\mathbf{x}$  for each pair is a two-element column vector  $[1 \ x]^T$ .

You run an algorithm to fit a model for the probability of  $Y = 1$  given  $\mathbf{x}$ :

$$\mathbb{P}(Y = 1 | \mathbf{x}) = \sigma(\mathbf{x}^T \beta)$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Your algorithm returns  $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$

- (a) Calculate  $\hat{\mathbb{P}}(Y = 1 | \mathbf{x} = [1 \ 0]^T)$

$$\mathbf{x} \cdot \hat{\beta} = [1 \ 0]^T \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}(1) + 0(-\frac{1}{2}) = -\frac{1}{2} \Rightarrow \mathbb{P}(Y=1 | \mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}) = \sigma\left(-\frac{1}{2}\right)$$

- (b) The empirical risk using log loss (a.k.a., cross-entropy loss) is given by:

$$\begin{aligned} R(\beta) &= \frac{1}{n} \sum_{i=1}^n -\log \hat{\mathbb{P}}(Y = y_i | \mathbf{x}_i) \\ &= -\frac{1}{n} \sum_{i=1}^n y_i \log \hat{\mathbb{P}}(Y = 1 | \mathbf{x}_i) + (1 - y_i) \log \hat{\mathbb{P}}(Y = 0 | \mathbf{x}_i) \end{aligned}$$

$$\frac{1}{1+e^{-t}} = \frac{e^t}{e^t} \left( \frac{1}{1+e^{-t}} \right) = \frac{e^t}{e^t + 1}$$

And  $\hat{\mathbb{P}}(Y = 1 | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)}$  while  $\hat{\mathbb{P}}(Y = 0 | \mathbf{x}_i) = \frac{1}{1 + \exp(\mathbf{x}_i^T \beta)}$ . Therefore,

$$\begin{aligned} R(\beta) &= -\frac{1}{n} \sum_{i=1}^n y_i \log \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} + (1 - y_i) \log \frac{1}{1 + \exp(\mathbf{x}_i^T \beta)} \\ &= -\frac{1}{n} \sum_{i=1}^n y_i \mathbf{x}_i^T \beta + \log(\sigma(-\mathbf{x}_i^T \beta)) \end{aligned}$$

Let  $\beta = [\beta_0 \ \beta_1]$ . Explicitly write out the empirical risk for the data set  $(1, 0)$  and  $(-1, 1)$  as a function of  $\beta_0$  and  $\beta_1$ .

- (c) Calculate the empirical risk for  $\hat{\beta} = \left[-\frac{1}{2} \quad -\frac{1}{2}\right]^T$  and the two observations  $(1, 0)$  and  $(-1, 1)$ .

Don't worry about this, just know how to do calculations with normal model when you're given an  $x$  and  $\beta$