Recitation 1

Gradients and Directional Derivatives

Brett Bernstein

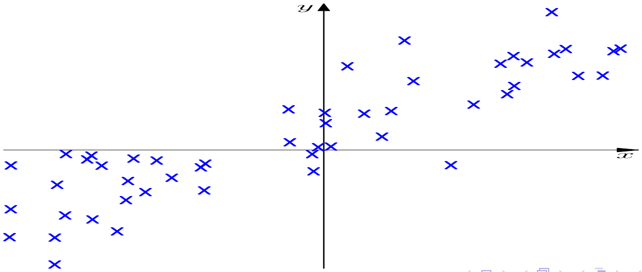
CDS at NYU

January 29, 2019

Intro Question

Question

We are given the data set $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We want to fit a linear function to this data by performing empirical risk minimization. More precisely, we are using the hypothesis space $\mathcal{F} = \{h_\theta(x) = \theta^T x \mid \theta \in \mathbb{R}^d\}$ and the loss function $\ell(a, y) = (a - y)^2$. Given an initial guess $\tilde{\theta}$ for the empirical risk minimizing parameter vector, how could we improve our guess?



Intro Solution

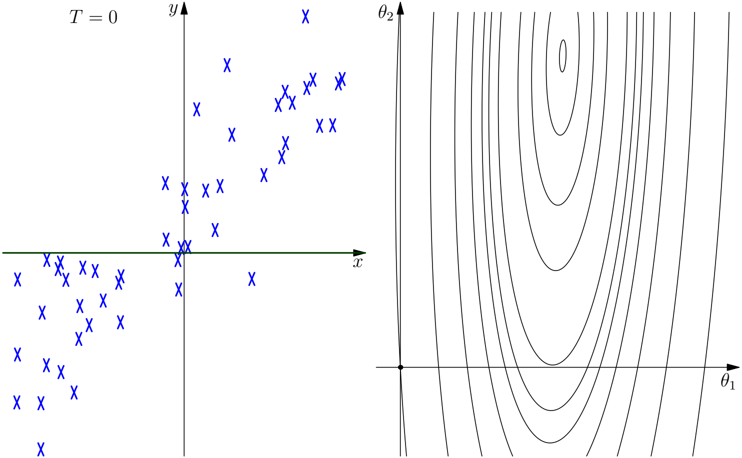
Solution

- The empirical risk is given by

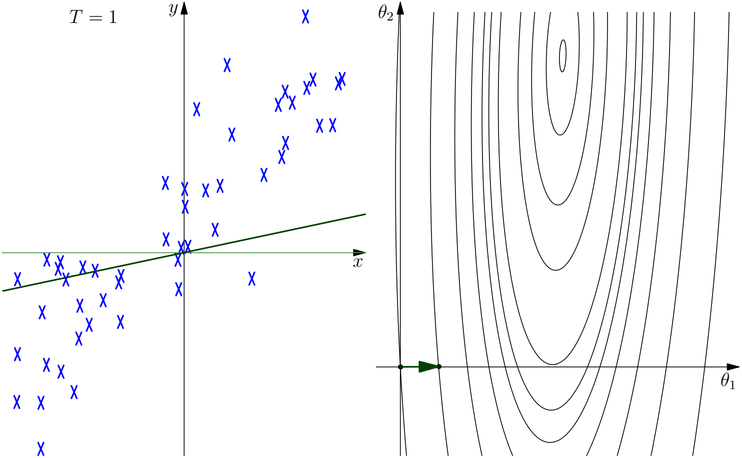
$$J(\theta) := \hat{R}_n(h_\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(h_\theta(x_i), y_i) = \frac{1}{n}\sum_{i=1}^{n}(\theta^T x_i - y_i)^2 = \frac{1}{n}\|X\theta - y\|_2^2,$$

  where $X \in \mathbb{R}^{n\times d}$ is the matrix whose ith row is given by x_i^T.
- Can improve a non-optimal guess $\tilde{\theta}$ by taking a small step in the direction of the negative gradient $-\nabla J(\theta)$.

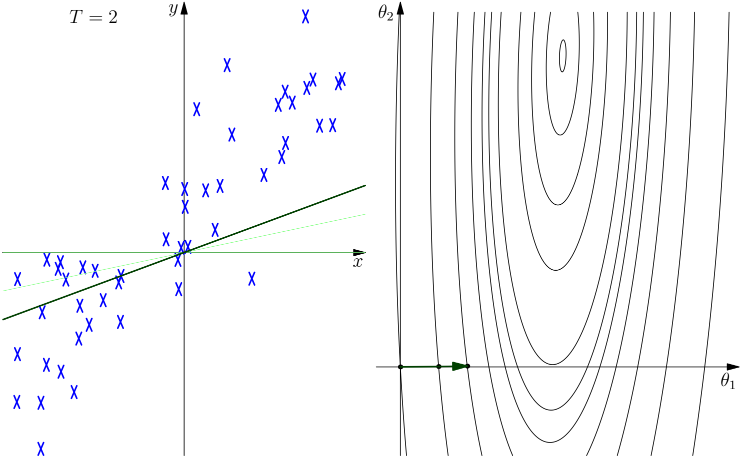
Negative Gradient Steps



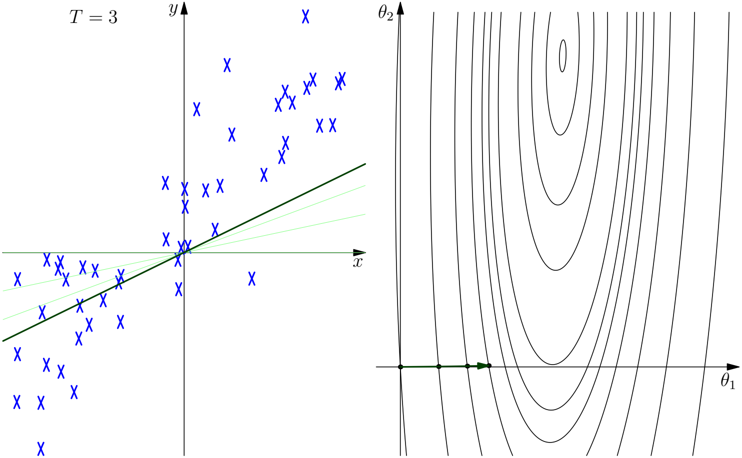
Negative Gradient Steps



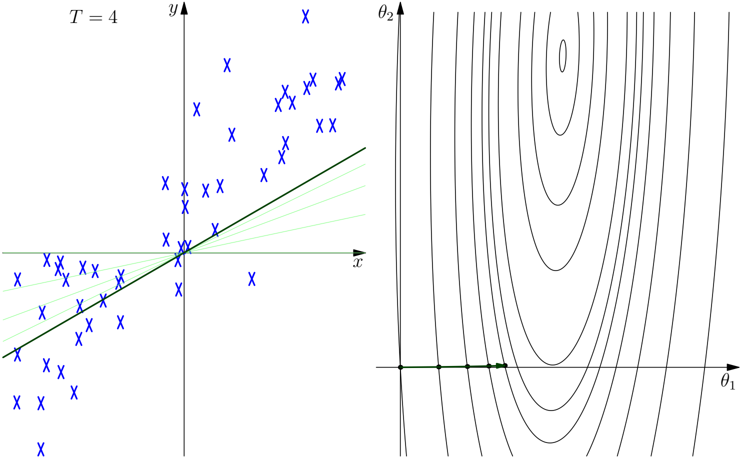
Negative Gradient Steps



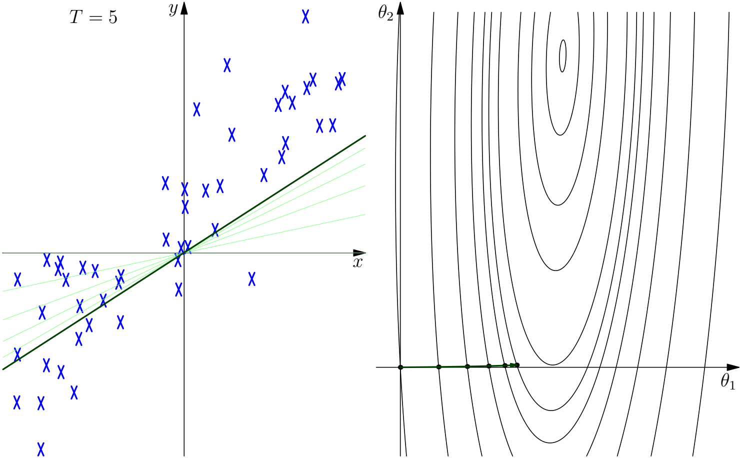
Negative Gradient Steps



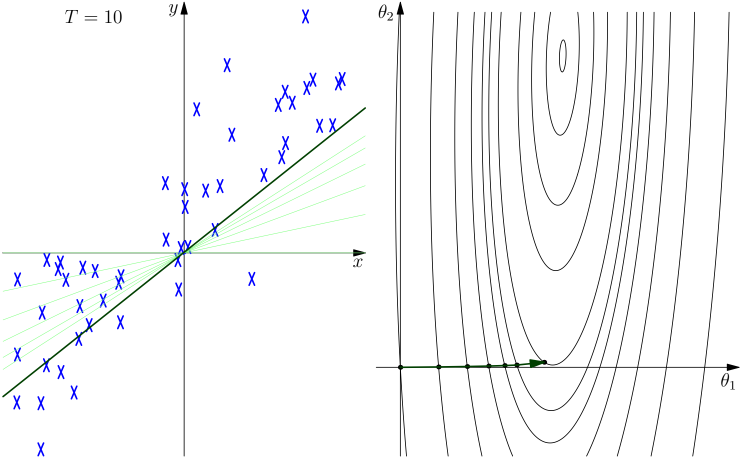
Negative Gradient Steps



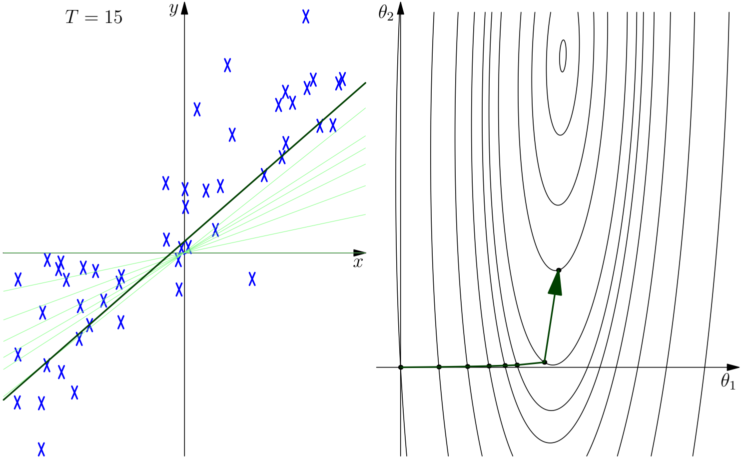
Negative Gradient Steps



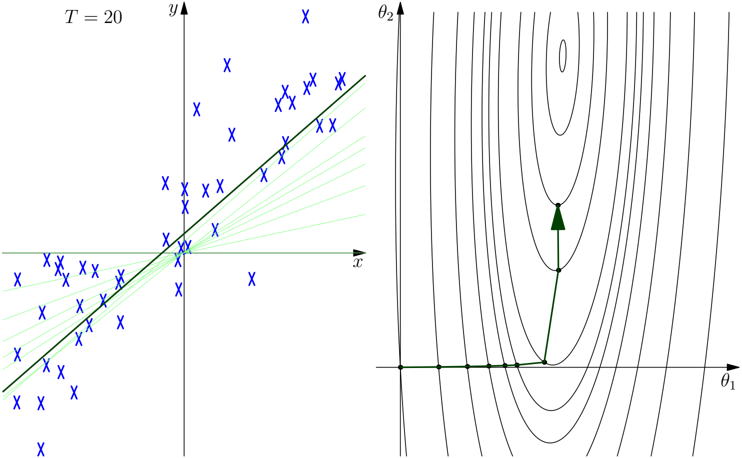
Negative Gradient Steps



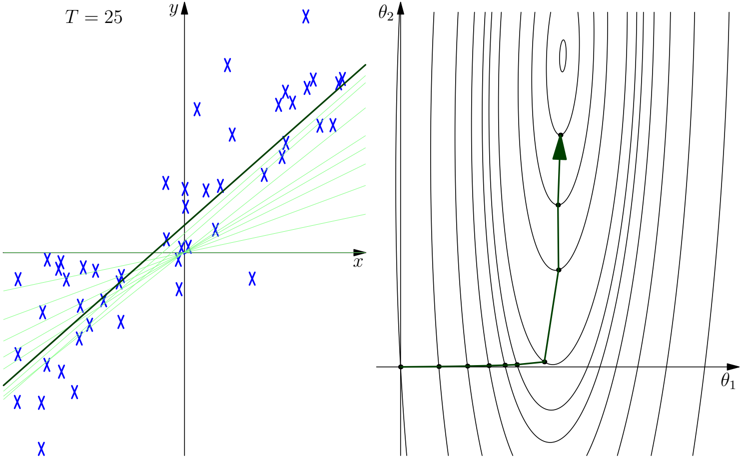
Negative Gradient Steps



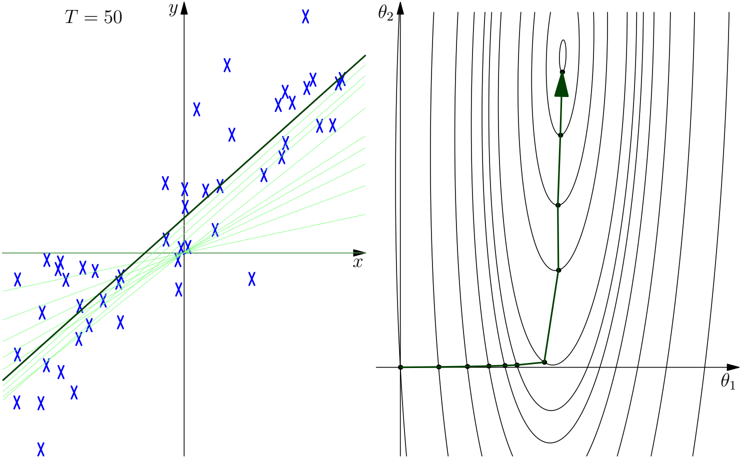
Negative Gradient Steps



Negative Gradient Steps



Negative Gradient Steps



Single Variable Differentiation

- Calculus lets us turn non-linear problems into linear algebra.
- For $f : \mathbb{R} \to \mathbb{R}$ differentiable, the derivative is given by

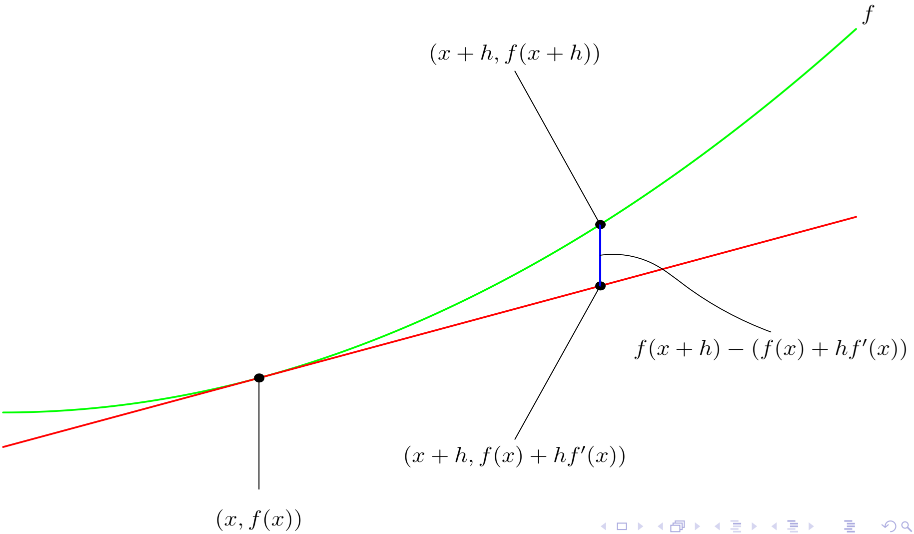
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

- Can also be written as

$$f(x+h) = f(x) + hf'(x) + o(h) \quad \text{as } h \to 0,$$

  where $o(h)$ denotes a function $g(h)$ with $g(h)/h \to 0$ as $h \to 0$.
- Points with $f'(x) = 0$ are called *critical points*.

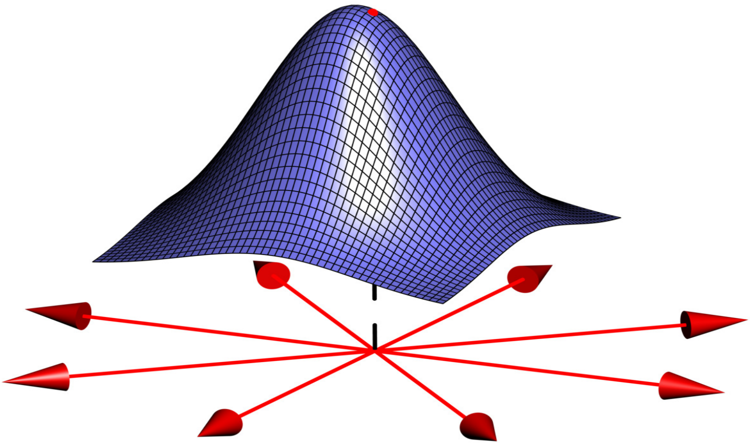
1D Linear Approximation By Derivative



Multivariable Differentiation

- Consider now a function $f : \mathbb{R}^n \to \mathbb{R}$ with inputs of the form $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.
- Unlike the 1-dimensional case, we cannot assign a single number to the slope at a point since there are many directions we can move in.

Multiple Possible Directions for $f : \mathbb{R}^2 \to \mathbb{R}$



Multivariable Differentiation

1. We will look at two (related) methods for understanding multivarible differentiation:
    1. Directional Derivatives: Derivative computed along a single direction
    2. Gradient: Gives multidimensional linear approximation and the steepest ascent direction

Directional Derivative

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$. The directional derivative $f'(x; u)$ of f at $x \in \mathbb{R}^n$ in the direction $u \in \mathbb{R}^n$ is given by

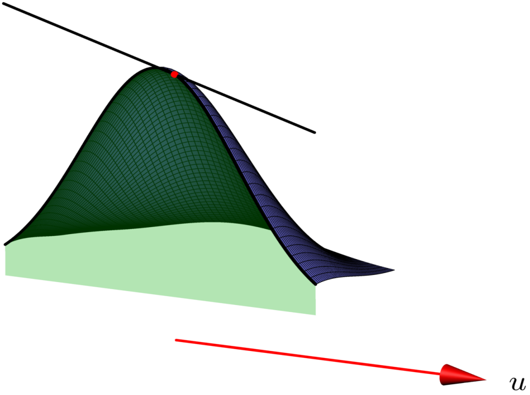
$$f'(x; u) = \lim_{h \to 0} \frac{f(x + hu) - f(x)}{h}.$$

- By fixing a direction u we turned our multidimensional problem into a 1-dimensional problem.
- Similar to 1-d we have

$$f(x + hu) = f(x) + hf'(x; u) + o(h).$$

- We say that u is a *descent direction* of f at x if $f'(x; u) < 0$.
- Taking a small enough step in a descent direction causes the function value decreases.

Directional Derivative as a Slope of a Slice



Partial Derivative

- Let $e_i = (\overbrace{0, 0, \ldots, 0}^{i-1}, 1, 0, \ldots, 0)$ denote the ith standard basis vector.
- The ith *partial derivative* is defined to be the directional derivative along e_i.
- It can be written many ways:

$$f'(x; e_i) = \frac{\partial}{\partial x_i} f(x) = \partial_{x_i} f(x) = \partial_i f(x).$$

- What is the intuitive meaning of $\partial_{x_i} f(x)$? For example, what does a large value of $\partial_{x_3} f(x)$ imply?

Differentiability and Gradients

- We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable* at $x \in \mathbb{R}^n$ if

$$\lim_{v \to 0} \frac{f(x+v) - f(x) - g^T v}{\|v\|_2} = 0,$$

  for some $g \in \mathbb{R}^n$.

- If it exists, this g is unique and is called the *gradient* of f at x with notation

$$g = \nabla f(x).$$

- It can be shown that

$$\nabla f(x) = \begin{pmatrix} \partial_{x_1} f(x) \\ \vdots \\ \partial_{x_n} f(x) \end{pmatrix}.$$

Useful Convention

- Consider $f : \mathbb{R}^{p+q} \to \mathbb{R}$.
- Split the input $x \in \mathbb{R}^{p+q}$ into parts $w \in \mathbb{R}^p$ and $z \in \mathbb{R}^q$ so that $x = (w, z)$.
- Define the partial gradients

$$\nabla_w f(w,z) := \begin{pmatrix} \partial_{w_1} f(w,z) \\ \vdots \\ \partial_{w_p} f(w,z) \end{pmatrix} \quad \text{and} \quad \nabla_z f(w,z) := \begin{pmatrix} \partial_{z_1} f(w,z) \\ \vdots \\ \partial_{z_q} f(w,z) \end{pmatrix}.$$

Linear Approximation and Tangent Plane

- Gradient gives us a linear approximation of f near the point x:

$$f(x+v) \approx f(x) + \nabla f(x)^T v.$$

- Analogous to the 1-d case we can express differentiability as

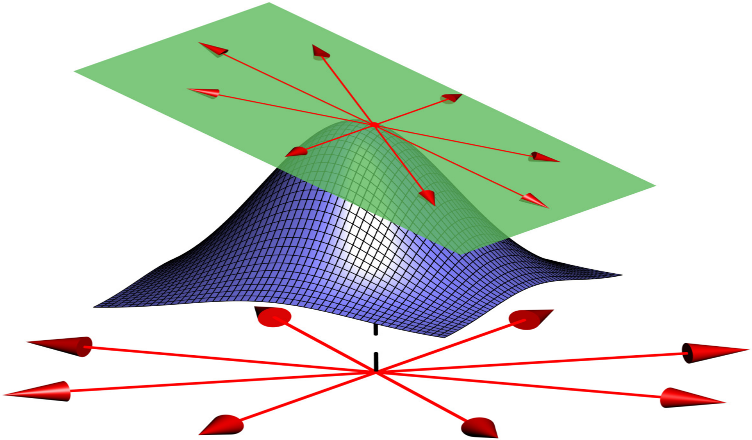
$$f(x+v) = f(x) + \nabla f(x)^T v + o(\|v\|_2).$$

- The gradient approximation can be seen as a tangent plane given by

$$P = \{(x+v, f(x) + \nabla f(x)^T v) \mid v \in \mathbb{R}^n\} \subseteq \mathbb{R}^{n+1}.$$

- Methods like gradient descent approximate a function locally by its tangent plane, and then take a step accordingly.

Tangent Plane for $f : \mathbb{R}^2 \to \mathbb{R}$



Directional Derivatives from Gradients

- If f is differentiable we obtain a formula for any directional derivative in terms of the gradient

$$f'(x;u) = \nabla f(x)^T u.$$

- This means a direction is a descent direction if and only if it makes an acute angle with the negative gradient.
- If $\nabla f(x) \neq 0$ applying Cauchy-Schwarz gives

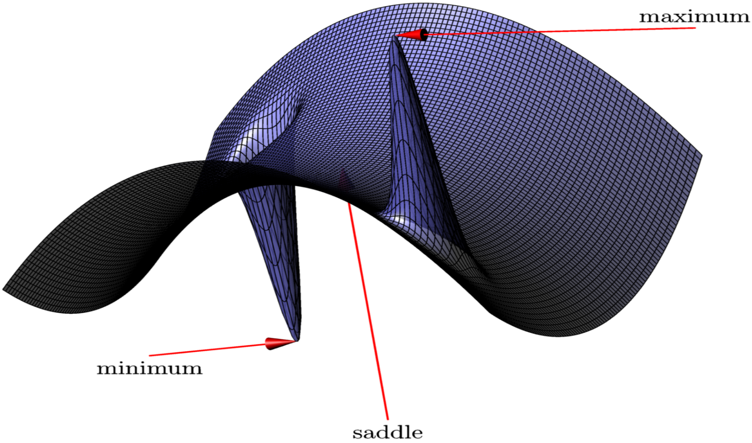
$$\arg\max_{\|u\|_2=1} f'(x;u) = \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \quad \text{and} \quad \arg\min_{\|u\|_2=1} f'(x;u) = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

- The gradient points in the direction of steepest ascent.
- The negative gradient points in the direction of steepest descent.

Critical Points

- Analogous to 1-d, if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and x is a local extremum then we must have $\nabla f(x) = 0$.
- Points with $\nabla f(x) = 0$ are called *critical points*.
- Later in the course we will see that for a convex differentiable function, x is a critical point if and only if it is a global minimizer.

Critical Points of $f : \mathbb{R}^2 \to \mathbb{R}$



Recap

- To find a good decision function we will minimize the empirical loss on the training data.
- Having fixed a hypothesis space parameterized by θ, finding a good decision function means finding a good θ.
- Given a current guess for θ, we will use the gradient of the empirical loss (w.r.t. θ) to get a local linear approximation.
- If the gradient is non-zero, taking a small step in the direction of the negative gradient is guaranteed to decrease the empirical loss.
- This motivates the minimization algorithm called gradient descent.

Computing Gradients

Question

For questions 1 and 2, compute the gradient of the given function.

1. $J : \mathbb{R}^3 \to \mathbb{R}$ is given by
$$J(\theta_1, \theta_2, \theta_3) = \log(1 + e^{\theta_1 + 2\theta_2 + 3\theta_3}).$$
2. $J : \mathbb{R}^n \to \mathbb{R}$ is given by
$$J(\theta) = \|X\theta - y\|_2^2 = (X\theta - y)^T(X\theta - y) = \theta^T X^T X\theta - 2y^T X\theta + y^T y,$$
for some $X \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$.
3. Assume X in the previous question has full column rank. What is the critical point of J?

$J(\theta_1, \theta_2, \theta_3) = \log(1 + e^{\theta_1+2\theta_2+3\theta_3})$ Solution 1

We can compute the partial derivatives directly:

$$\begin{aligned}
\partial_{\theta_1} J(\theta_1, \theta_2, \theta_3) &= \frac{e^{\theta_1+2\theta_2+3\theta_3}}{1 + e^{\theta_1+2\theta_2+3\theta_3}} \\
\partial_{\theta_2} J(\theta_1, \theta_2, \theta_3) &= \frac{2e^{\theta_1+2\theta_2+3\theta_3}}{1 + e^{\theta_1+2\theta_2+3\theta_3}} \\
\partial_{\theta_3} J(\theta_1, \theta_2, \theta_3) &= \frac{3e^{\theta_1+2\theta_2+3\theta_3}}{1 + e^{\theta_1+2\theta_2+3\theta_3}}
\end{aligned}$$

and obtain

$$\nabla J(\theta_1, \theta_2, \theta_3) = \begin{pmatrix} \frac{e^{\theta_1+2\theta_2+3\theta_3}}{1 + e^{\theta_1+2\theta_2+3\theta_3}} \\ \frac{2e^{\theta_1+2\theta_2+3\theta_3}}{1 + e^{\theta_1+2\theta_2+3\theta_3}} \\ \frac{3e^{\theta_1+2\theta_2+3\theta_3}}{1 + e^{\theta_1+2\theta_2+3\theta_3}} \end{pmatrix}.$$

$J(\theta_1, \theta_2, \theta_3) = \log(1 + e^{\theta_1 + 2\theta_2 + 3\theta_3})$ Solution 2

- Spot the linear algebra!
- Let $w = (1, 2, 3)^T$.
- Write $J(\theta) = \log(1 + e^{w^T\theta})$.
- Apply a version of the chain rule:

$$\nabla J(\theta) = \frac{e^{w^T\theta}}{1 + e^{w^T\theta}} w.$$

Theorem (Chain Rule)

If $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are differentiable then

$$\nabla(g \circ h)(x) = g'(h(x)) \nabla h(x).$$

$J(\theta) = \|X\theta - y\|_2^2$ Solution

- We could use techniques similar to the previous problem, but instead we show a different method using directional derivatives.
- For arbitrary $t \in \mathbb{R}$ and $\theta, v \in \mathbb{R}^n$ we have

$$\begin{aligned}
&J(\theta + tv)\\
&= (\theta + tv)^T X^T X(\theta + tv) - 2y^T X(\theta + tv) + y^T y\\
&= \underline{\theta^T X^T X\theta} + t^2 v^T X^T Xv + 2t\theta^T X^T Xv - \underline{2y^T X\theta} - 2ty^T Xv + \underline{y^T y}\\
&= \underline{J(\theta)} + t(2\theta^T X^T X - 2y^T X)v + t^2 v^T X^T Xv.
\end{aligned}$$

- This gives

$$J'(\theta; v) = \lim_{t \to 0} \frac{J(\theta + tv) - J(\theta)}{t} = (2\theta^T X^T X - 2y^T X)v = \nabla J(\theta)^T v$$

- Thus $\nabla J(\theta) = 2(X^T X\theta - X^T y) = 2X^T(X\theta - y)$.
- Data science interpretation of $\nabla J(\theta)$ (assuming columns of X are centered)?

Critical Points of $J(\theta) = \|X\theta - y\|_2^2$

- Need $\nabla J(\theta) = 2X^TX\theta - 2X^Ty = 0$.
- Since X is assumed to have full column rank, we see that X^TX is invertible.
- Thus we have $\theta = (X^TX)^{-1}X^Ty$.
- As we will see later, this function is strictly convex (Hessian $\nabla^2 J(\theta) = 2X^TX$ is positive definite).
- Thus we have found the unique minimizer (least squares solution).

Technical Aside: Differentiability

- When computing the gradients above we assumed the functions were differentiable.
- Can use the following theorem to be completely rigorous.

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose $\partial_i f : \mathbb{R}^n \to \mathbb{R}$ is continuous for $i = 1, \ldots, n$. Then f is differentiable.