

Recitation 1

Gradients and Directional Derivatives

Brett Bernstein

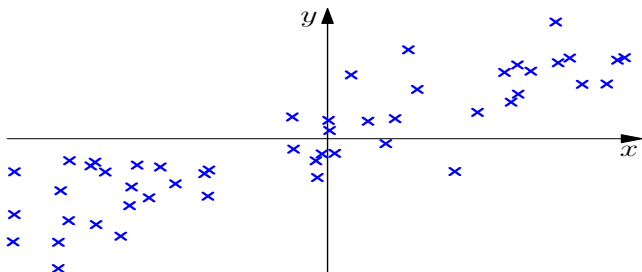
CDS at NYU

January 22, 2019

Intro Question

Question

We are given the data set $(x_1, y_1), \dots, (x_n, y_n)$ 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 $\mathcal{F} = \{f(x) = w^T x \mid w \in \mathbb{R}^d\}$ and the loss function $\ell(a, y) = (a - y)^2$. Given an initial guess \tilde{w} for the empirical risk minimizing parameter vector, how could we improve our guess?



Intro Solution

Solution

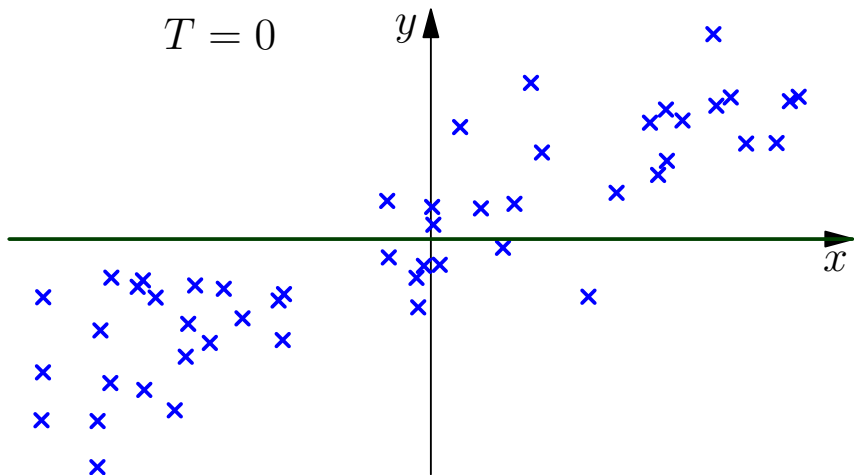
- The empirical risk is given by

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) = \frac{1}{n} \sum_{i=1}^n (w^T x_i - y_i)^2 = \frac{1}{n} \|Xw - y\|_2^2,$$

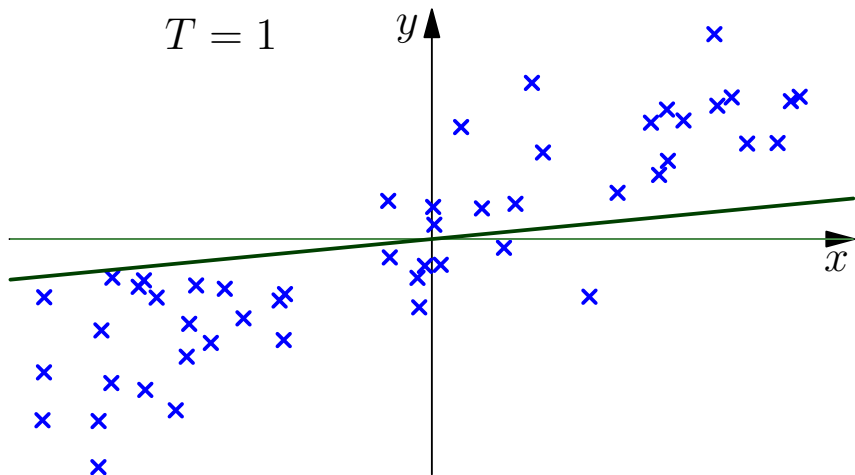
where $X \in \mathbb{R}^{n \times d}$ is the matrix whose i th row is given by x_i .

- Can improve a non-optimal guess \tilde{w} by taking a small step in the direction of the negative gradient.

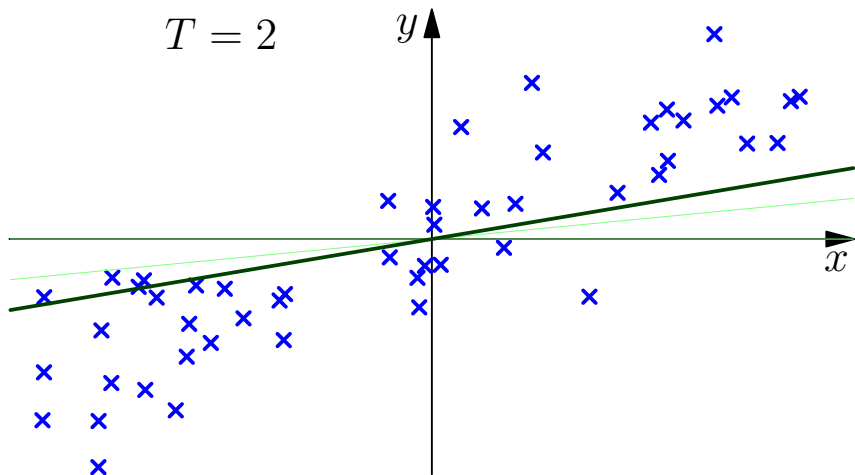
Negative Gradient Steps



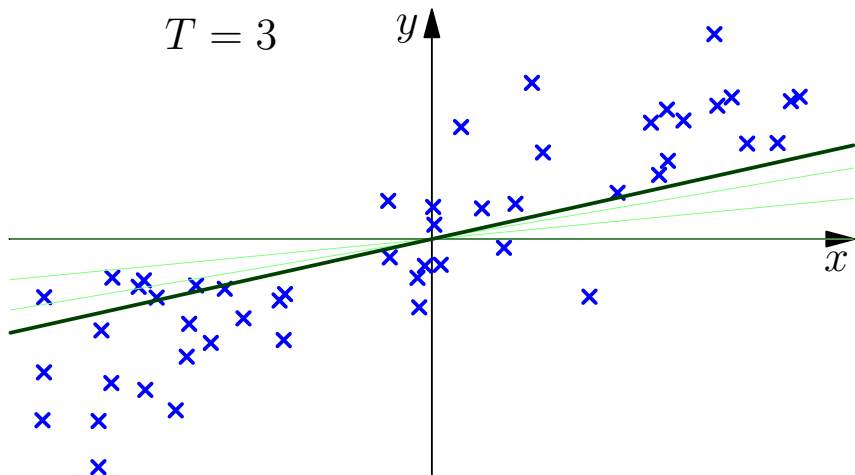
Negative Gradient Steps



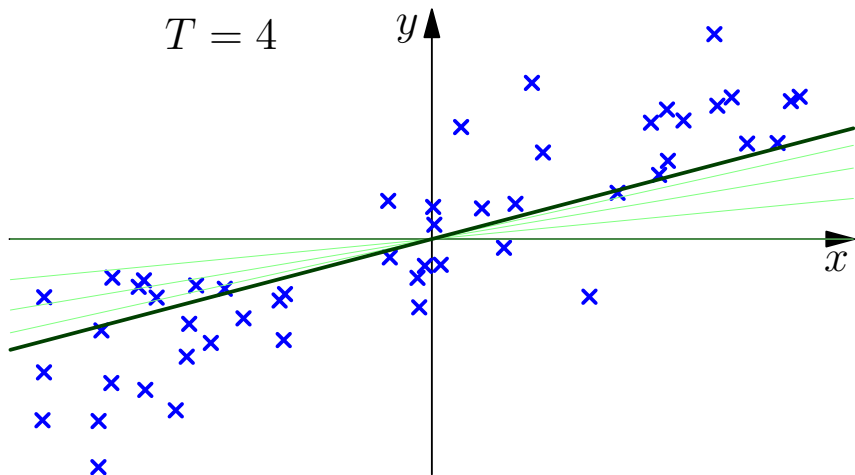
Negative Gradient Steps



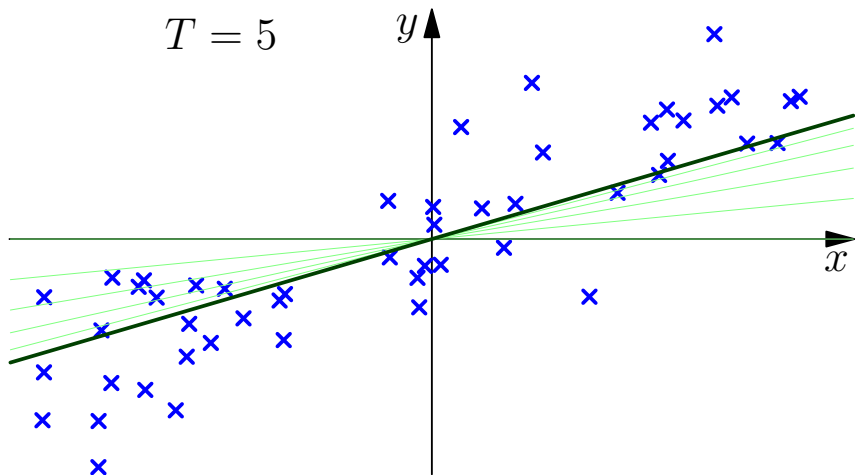
Negative Gradient Steps



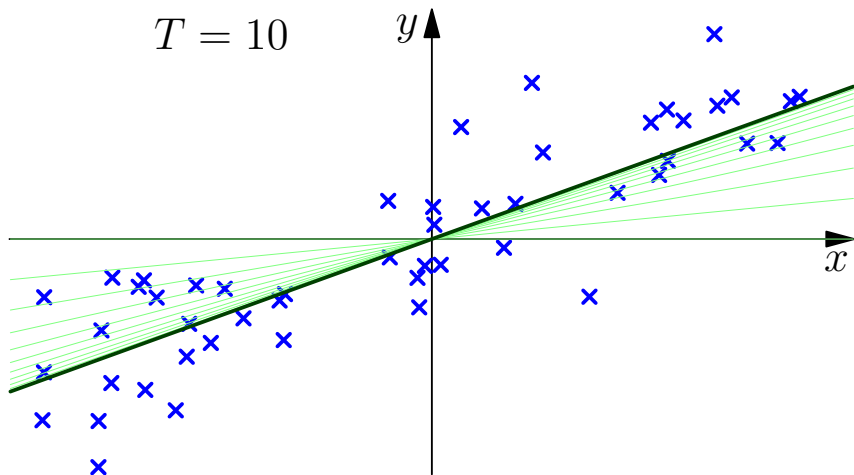
Negative Gradient Steps



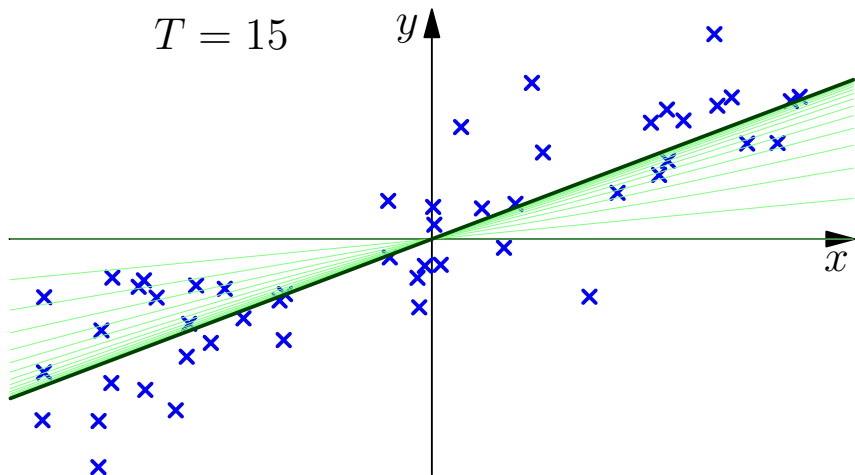
Negative Gradient Steps



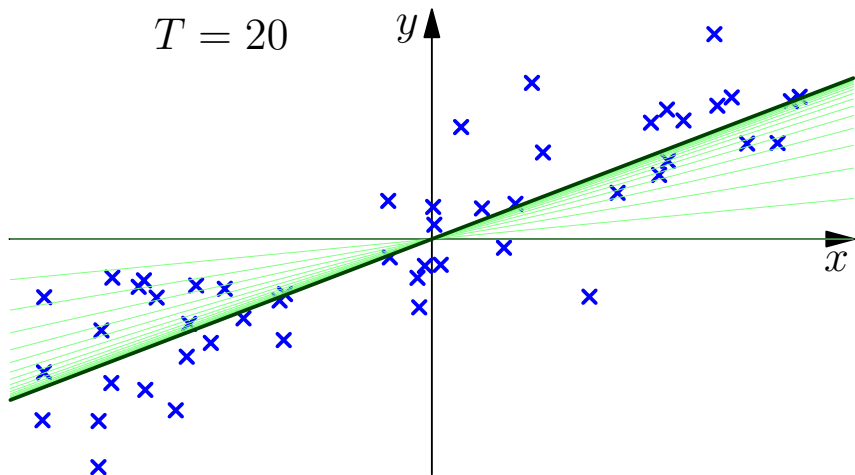
Negative Gradient Steps



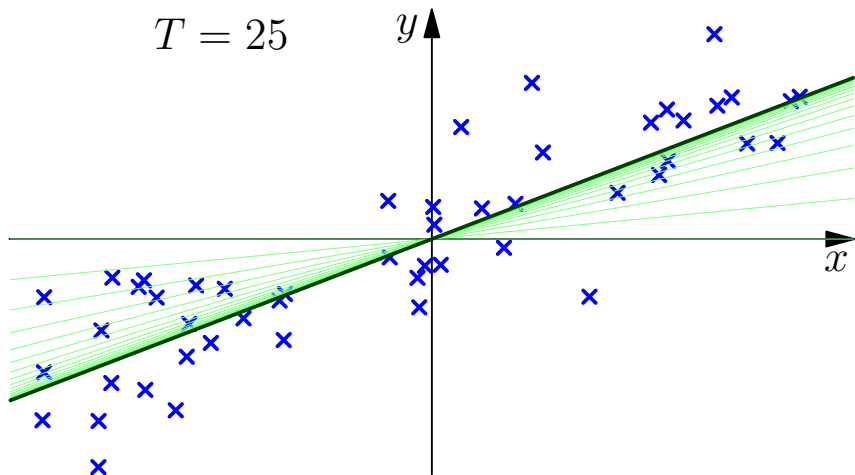
Negative Gradient Steps



Negative Gradient Steps



Negative Gradient Steps



Single Variable Differentiation

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

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

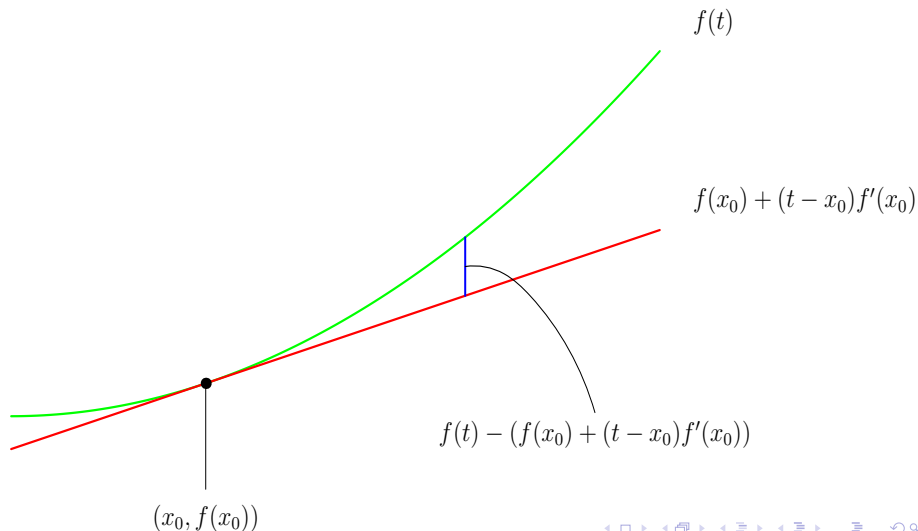
- Can also be written as

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

where $o(h)$ denotes a function $g(h)$ with $g(h)/h \rightarrow 0$ as $h \rightarrow 0$.

- Points with $f'(x) = 0$ are called *critical points*.

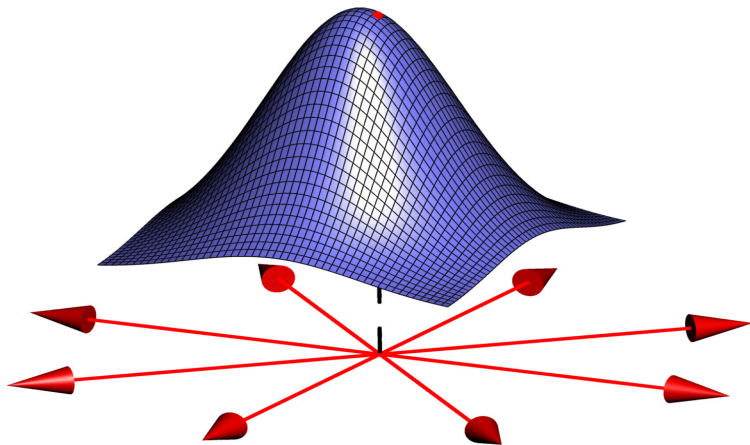
1D Linear Approximation By Derivative



Multivariable Differentiation

- Consider now a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with inputs of the form $x = (x_1, \dots, x_n) \in \mathbb{R}^n$.
- 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 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$



Directional Derivative

Definition

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. 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

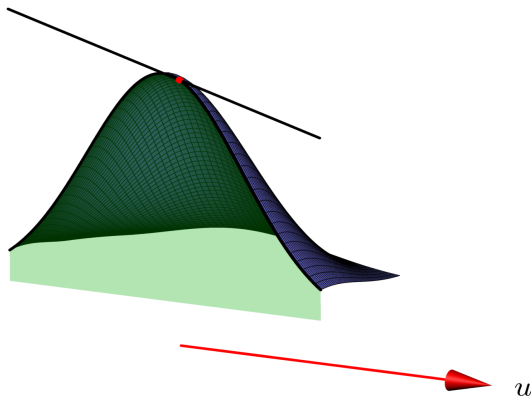
$$f'(x; u) = \lim_{h \rightarrow 0} \frac{f(x + hu) - f(x)}{h}.$$

- 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 $f'(x; u) < 0$.

Directional Derivative as a Slope of a Slice



Partial Derivative

- Let $e_i = (\overbrace{0, 0, \dots, 0}^{i-1}, 1, 0, \dots, 0)$ denote the i th standard basis vector.
- The i th *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

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

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

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 $f : \mathbb{R}^{p+q} \rightarrow \mathbb{R}$.
- 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

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

Tangent Plane

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

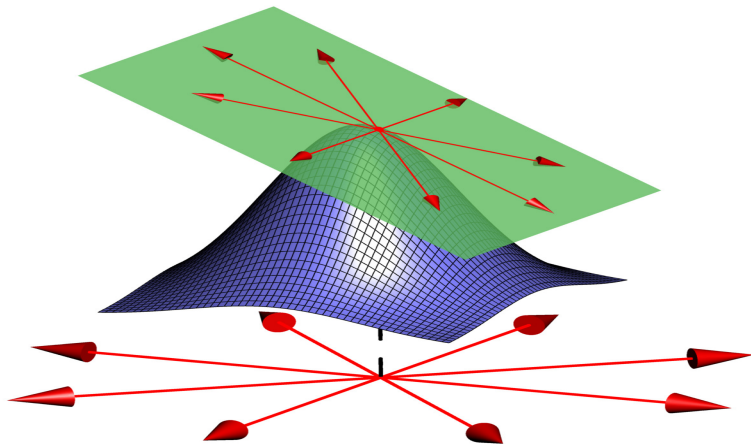
$$f(x + v) = f(x) + \nabla f(x)^T v + o(\|v\|_2).$$

- The approximation $f(x + v) \approx f(x) + \nabla f(x)^T v$ gives a tangent plane at the point x .
- The tangent plane of f at x is 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 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$



Directional Derivatives from Gradients

- If f is differentiable we have

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

- If $\nabla f(x) \neq 0$ this implies that

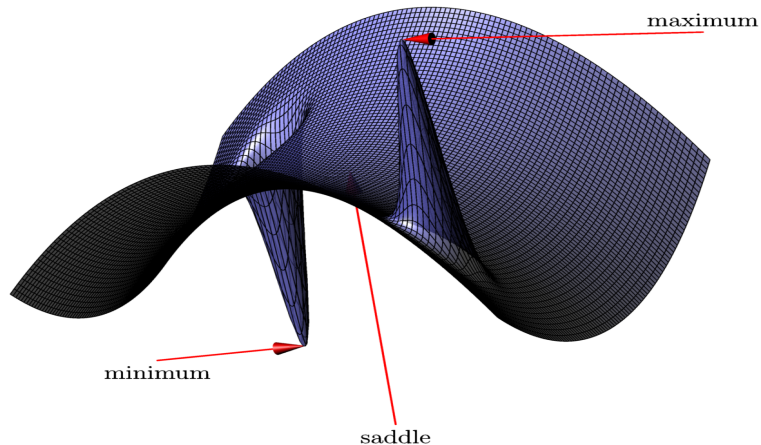
$$\arg \max_{\|u\|_2=1} f'(x; u) = \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \quad \text{and} \quad \arg \min_{\|u\|_2=1} f'(x; u) = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

- 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 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 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$



Computing Gradients

Question

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

- ① $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is given by

$$f(x_1, x_2, x_3) = \log(1 + e^{x_1 + 2x_2 + 3x_3}).$$

- ② $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is given by

$$f(x) = \|Ax - y\|_2^2 = (Ax - y)^T (Ax - y) = x^T A^T A x - 2y^T A x + y^T y,$$

for some $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$.

- ③ Assume A in the previous question has full column rank. What is the critical point of f ?

$f(x_1, x_2, x_3) = \log(1 + e^{x_1+2x_2+3x_3})$ Solution 1

We can compute the partial derivatives directly:

$$\begin{aligned}\partial_{x_1} f(x_1, x_2, x_3) &= \frac{e^{x_1+2x_2+3x_3}}{1 + e^{x_1+2x_2+3x_3}} \\ \partial_{x_2} f(x_1, x_2, x_3) &= \frac{2e^{x_1+2x_2+3x_3}}{1 + e^{x_1+2x_2+3x_3}} \\ \partial_{x_3} f(x_1, x_2, x_3) &= \frac{3e^{x_1+2x_2+3x_3}}{1 + e^{x_1+2x_2+3x_3}}\end{aligned}$$

and obtain

$$\nabla f(x_1, x_2, x_3) = \begin{pmatrix} \frac{e^{x_1+2x_2+3x_3}}{1 + e^{x_1+2x_2+3x_3}} \\ \frac{2e^{x_1+2x_2+3x_3}}{1 + e^{x_1+2x_2+3x_3}} \\ \frac{3e^{x_1+2x_2+3x_3}}{1 + e^{x_1+2x_2+3x_3}} \end{pmatrix}.$$

$f(x_1, x_2, x_3) = \log(1 + e^{x_1 + 2x_2 + 3x_3})$ Solution 2

- Let $w = (1, 2, 3)^T$.
- Write $f(x) = \log(1 + e^{w^T x})$.
- Apply a version of the chain rule:

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

Theorem (Chain Rule)

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

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

$f(x) = \|Ax - y\|_2^2$ 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 $x, v \in \mathbb{R}^n$ we have

$$\begin{aligned} f(x + tv) &= (x + tv)^T A^T A (x + tv) - 2y^T A (x + tv) + y^T y \\ &= x^T A^T A x + t^2 v^T A^T A v + 2tx^T A^T A v - 2y^T A x - 2ty^T A v + y^T y \\ &= f(x) + t(2x^T A^T A - 2y^T A)v + t^2 v^T A^T A v. \end{aligned}$$

- This gives

$$f'(x; v) = \lim_{t \rightarrow 0} \frac{f(x + tv) - f(x)}{t} = (2x^T A^T A - 2y^T A)v = \nabla f(x)^T v$$

- Thus $\nabla f(x) = 2(A^T A x - A^T y) = 2A^T(Ax - y)$.
- Data science interpretation of $\nabla f(x)$?

Critical Points of $f(x) = \|Ax - y\|_2^2$

- Need $\nabla f(x) = 2A^T Ax - 2A^T y = 0$.
- Since A is assumed to have full column rank, we see that $A^T A$ is invertible.
- Thus we have $x = (A^T A)^{-1} A^T y$.
- As we will see later, this function is strictly convex (Hessian $\nabla^2 f(x) = 2A^T A$ 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and suppose $\partial_{x_i} f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous for $i = 1, \dots, n$. Then f is differentiable.