Recitation 1: Gradients and Directional Derivatives

Intro Question

1. 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 $\mathcal{F} = \{f(x) = w^T x \mid w \in \mathbb{R}^d\}$ 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?
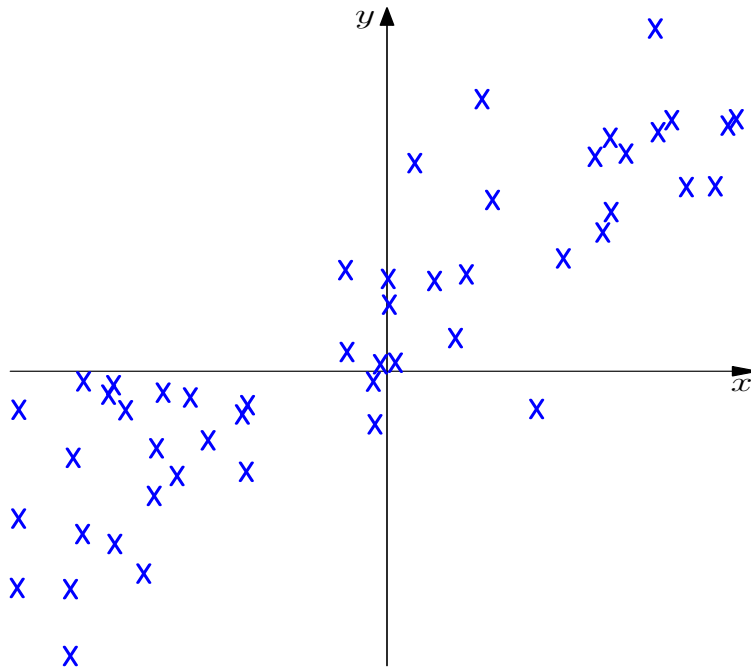

Figure 1: Data Set With $d = 1$

Multivariable Differentiation

Differential calculus allows us to convert non-linear problems into local linear problems, to which we can apply the well-developed techniques of linear algebra. Here we will review some of the important concepts needed in the rest of the course.

Single Variable Differentiation

To gain intuition, we first recall the single variable case. Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable. The derivative

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

gives us a local linear approximation of f near x. This is seen more clearly in the following form:

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

where $o(h)$ represents a function $g(h)$ with $g(h)/h \to 0$ as $h \to 0$. This can be used to show that if x is a local extremum of f then $f'(x) = 0$. Points with $f'(x) = 0$ are called *critical points*.
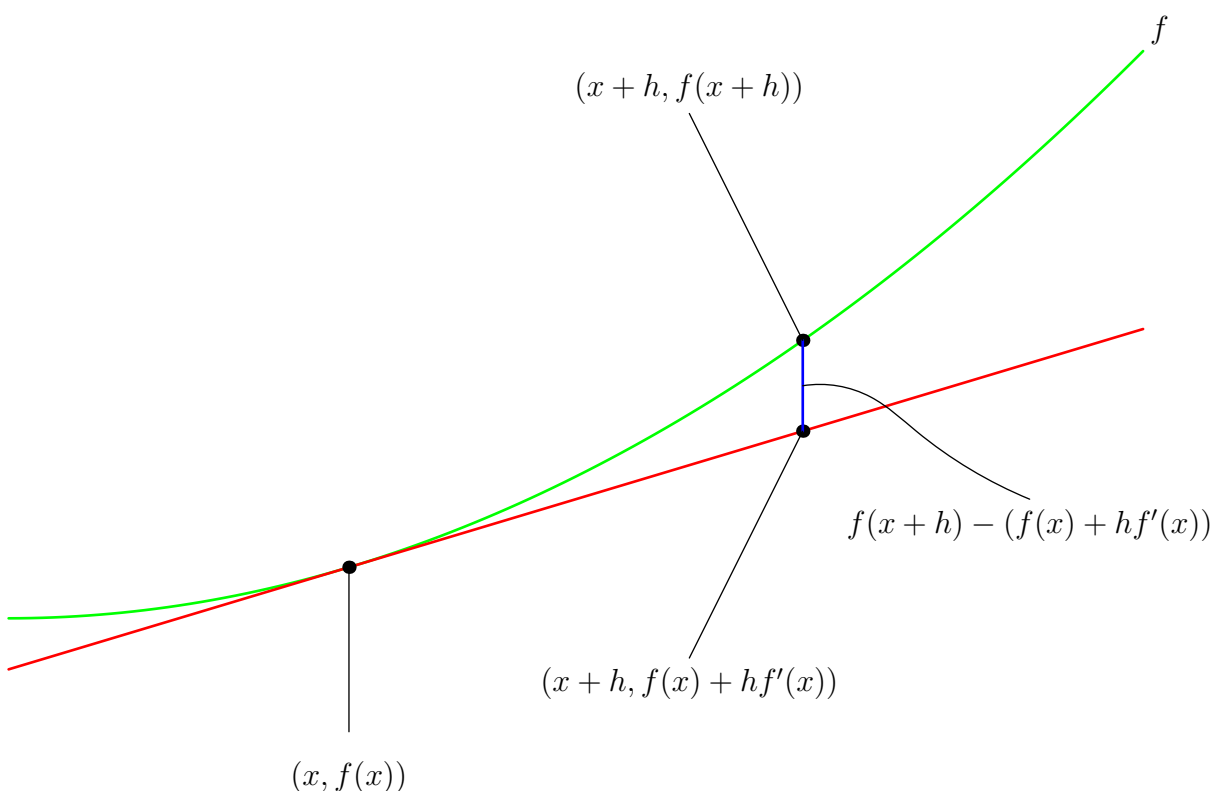

Figure 2: 1D Linear Approximation By Derivative

Multivariate Differentiation

More generally, we will look at functions $f : \mathbb{R}^n \to \mathbb{R}$. In the single-variable case, the derivative was just a number that signified how much the function increased when we moved in the positive x-direction. In the multivariable case, we have many possible directions we can move along from a given point $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.
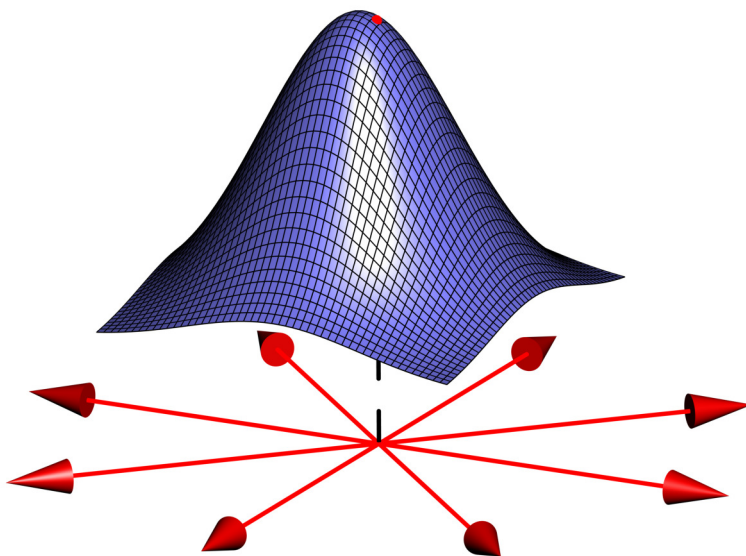
Figure 3: Multiple Possible Directions for $f : \mathbb{R}^2 \to \mathbb{R}$

If we fix a direction u we can compute the directional derivative $f'(x; u)$ as

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

This allows us to turn our multidimensional problem into a 1-dimensional computation. For instance,

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

mimicking our earlier 1-d formula. This says that nearby x we can get a good approximation of $f(x+hu)$ using the linear approximation $f(x)+hf'(x; u)$. In particular, if $f'(x; u) < 0$ (such a u is called a *descent direction*) then for sufficiently small $h > 0$ we have $f(x + hu) < f(x)$.
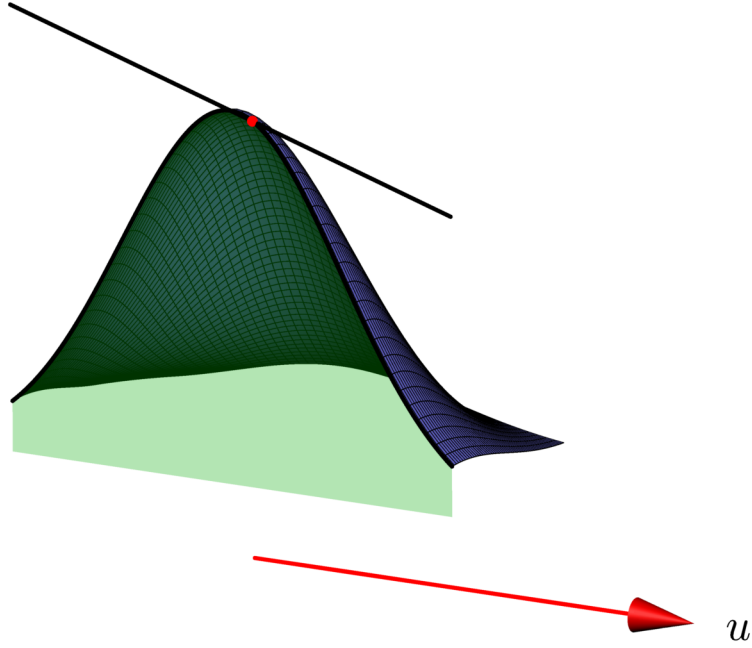
Figure 4: Directional Derivative as a Slope of a Slice

Let $e_i = (\overbrace{0, 0, \ldots, 0}^{i-1}, 1, 0, \ldots, 0)$ be the ith standard basis vector. The directional derivative in the direction e_i is called the ith partial derivative and can be written in several ways:

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

We say that f is differentiable at x if

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

for some $g \in \mathbb{R}^n$ (note that the limit for v is taken in $\mathbb{R}^n$). This g is uniquely determined, and is called the gradient of f at x denoted by $\nabla f(x)$. It can be shown that the gradient is the vector of partial derivatives:

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

The kth entry of the gradient (i.e., the kth partial derivative) is the approximate change in f due to a small positive change in x_k.

Sometimes we will split the variables of f into two parts. For instance, we could write $f(x, w)$ with $x \in \mathbb{R}^p$ and $w \in \mathbb{R}^q$. It is often useful to take the gradient with respect to some of the variables. Here we would write ∇_x or ∇_w to specify which part:

$$\nabla_x f(x, w) := \begin{pmatrix} \partial_{x_1} f(x, w) \\ \vdots \\ \partial_{x_p} f(x, w) \end{pmatrix} \quad \text{and} \quad \nabla_w f(x, w) := \begin{pmatrix} \partial_{w_1} f(x, w) \\ \vdots \\ \partial_{w_q} f(x, w) \end{pmatrix}.$$

Analogous to the univariate case, can express the condition for differentiability in terms of a gradient approximation:

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

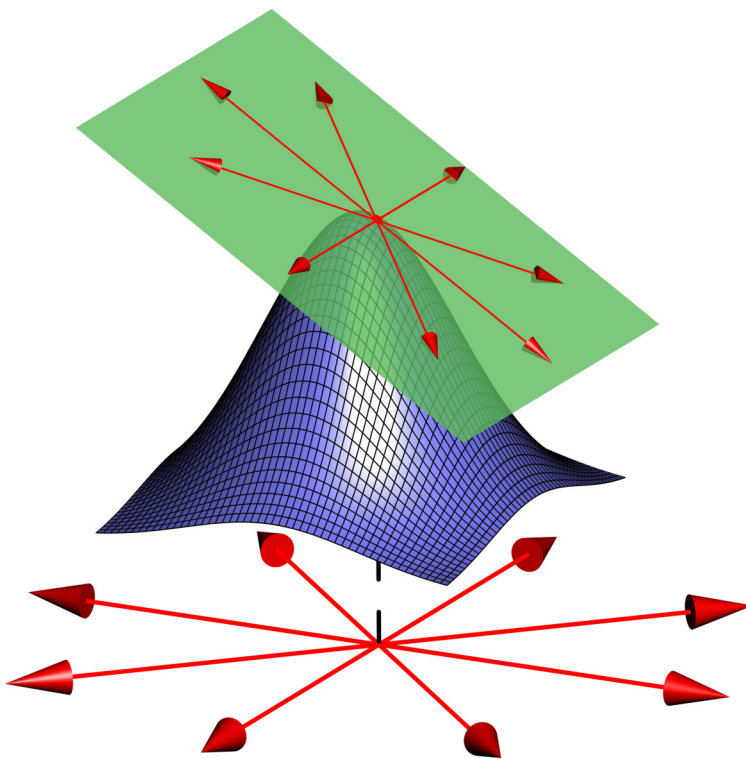The approximation $f(x + v) \approx f(x) + \nabla f(x)^T v$ gives a tangent plane at the point x as we let v vary.

Figure 5: Tangent Plane for $f : \mathbb{R}^2 \to \mathbb{R}$

If f is differentiable, we can use the gradient to compute an arbitrary directional derivative:

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

From this expression we can prove that (assuming $\nabla f(x) \neq 0$)

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

In words, we say that the gradient points in the direction of steepest ascent, and the negative gradient points in the direction of steepest descent.

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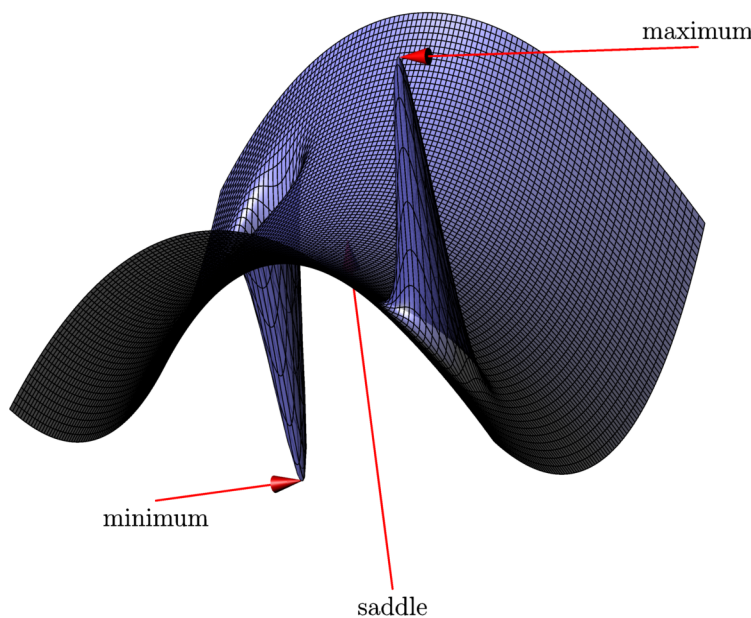

Figure 6: Critical Points of $f : \mathbb{R}^2 \to \mathbb{R}$

Computing Gradients

A simple method to compute the gradient of a function is to compute each partial derivative separately. For example, if $f : \mathbb{R}^3 \to \mathbb{R}$ is given by

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

then we can directly compute

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

and obtain

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

Alternatively, we could let $w = (1, 2, 3)^T$ and write

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

Then we can apply a version of the chain rule which says that if $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are differentiable then

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

Applying the chain rule twice (for log and exp) we obtain

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

where we use the fact that $\nabla_x(w^T x) = w$. This last expression is more concise, and is more amenable to vectorized computation in many languages.

Another useful technique is to compute a general directional derivative and then infer the gradient from the computation. For example, let $f : \mathbb{R}^n \to \mathbb{R}$ be 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 $y \in \mathbb{R}^m$. Assuming f is differentiable (so that $f'(x; v) = \nabla f(x)^T v$) we have

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

Thus we have

$$\frac{f(x + tv) - f(x)}{t} = (2x^T A^T A - 2y^T A)v + t v^T A^T A v.$$

Taking the limit as $t \to 0$ shows

$$f'(x; v) = (2x^T A^T A - 2y^T A)v \implies \nabla f(x) = (2x^T A^T A - 2y^T A)^T = 2A^T A x - 2A^T y.$$

Assume the columns of the data matrix A have been centered (by subtracting their respective means). We can interpret $\nabla f(x) = 2A^T(Ax - y)$ as (up to scaling) the covariance between the features and the residual.

Using the above calculation we can determine the critical points of f. Let's assume here that A has full column rank. Then $A^T A$ is invertible, and the unique critical point is $x = (A^T A)^{-1} A^T y$. As we will see later in the course, this is a global minimum since f is convex (the *Hessian* of f satisfies $\nabla^2 f(x) = 2A^T A \succ 0$).

$(\star)$ Proving Differentiability

With a little extra work we can make the previous technique give a proof of differentiability. Using the computation above, we can rewrite $f(x + v)$ as $f(x)$ plus terms depending on v:

$$f(x + v) = f(x) + (2x^T A^T A - 2y^T A)v + v^T A^T A v.$$

Note that

$$\frac{v^T A^T A v}{\|v\|_2} = \frac{\|Av\|_2^2}{\|v\|_2} \leq \frac{\|A\|_2^2 \|v\|_2^2}{\|v\|_2} = \|A\|_2^2 \|v\|_2 \to 0,$$

as $\|v\|_2 \to 0$. (This section is starred since we used the matrix norm $\|A\|_2$ here.) This shows $f(x + v)$ above has the form

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

This proves that f is differentiable and that

$$\nabla f(x) = 2A^T A x - 2A^T y.$$

Another method we could have used to establish differentiability is to observe that the partial derivatives are all continuous. This relies on the following theorem.

Theorem 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose $\partial_{x_i} f(x) : \mathbb{R}^n \to \mathbb{R}$ is continuous for all $x \in \mathbb{R}^n$ and all $i = 1, \ldots, n$. Then f is differentiable.*