Optimization for Data Science

Bernd Gärtner, Martin Jaggi

1 Theory of Convex Functions

1.1 Mathematical Notation

Bold lower case symbols are used for vectors, and normal font is used for their coordinates e.g. $\mathbf{x} = (x_1, x_2 \dots x_d) \in \mathbb{R}^d$. Vectors are assumed to be column vectors unless transposed, so $\mathbf{x}$ is a column and $\mathbf{x}^\top$ is a row. Revolutionary. $\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, i.e. the damn dot product.

$\|\mathbf{x}\|$ is the Euclidian norm i.e. the length of a vector and the squared Euclidian norm is $\|\mathbf{x}\|^2 = \mathbf{x}^\top \mathbf{x}$.

Conventions here include that $\mathbb{N} = \{1, 2, \dots\}$, so natural numbers do not include 0 and $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$, so the positive real numbers do include 0.

1.2 Cauchy-Schwarz Inequality

Lemma 1.1 (Cauchy-Schwarz Inequality)**.** *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$*, then*

$$|\mathbf{u}^\top \mathbf{v}| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

Proof. The core of the dot product, to me anyhow, is that given that θ is the acute angle between vector $\mathbf{u}, \mathbf{v}$, We have:

$$\|\mathbf{u}^\top \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| \cos(\theta)$$

Thankfully, We can assume the above. For an intuition for the above fact, see here.

The proof is then simply an observation that $\cos(\theta)$ is at most 1 (and at least -1), and so the inequality is satisfied in all cases. □

1.3 Spectral Norm

Spectral norm is a norm on matrices. Without further ado let's define it:

Definition 1.2 (Spectral Norm)**.** *For any matrix* $A \in \mathbb{R}^{m \times d}$ *let the spectral norm of* A *be*

$$\|A\| = \max_{\mathbf{v} \in \mathbb{R}^d, \mathbf{v} \neq 0} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\| = 1} \|A\mathbf{v}\|$$

So, what does this mean - there is some direction that A particularly likes, and vectors aligned with that direction get scaled by a whole lot. The spectral norm is an upper bound on how much the transformation by A can change the length of $\mathbf{v}$.

Other note is that spectral norm is indeed a norm. What is a norm again?

A norm is some operator that

1. Satisfies the *triangle inequality*. In our case this would be, for two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times d}$, it must be true that $\|AB\| \leq \|A\| + \|B\|$. This is pretty easy to understand in our case - remember that AB can be thought of as sequential matrix multiplication instead of some complicated megamatrix. In the sequential case, the first matrix application can scale the input vector by at most $\|B\|$, and the second transformation by at most $\|A\|$, so the result of the sequential application cannot stretch the input vector by more than $\|A\| + \|B\|$, and We're done.

2. A norm must be *absolutely homogeneous*, which is to say $\|\lambda A\| \leq |\lambda|\|A\|$ for some constant λ and in our case some matrix A. Absolute as in absolute value of λ. This is obviously true for our spectral norm.

3. Norm is only equal to 0 in case that the input is 0. In our case, $\|A\| = 0 \iff A = 0$. Again, makes sense, since We are looking at max in our definition, and for any non-zero matrix A there will be some vector $\mathbf{v}$ such that the transformation will result in a non-zero vector.

1.4 Mean Value Theorem

Theorem 1.3 (Mean Value Theorem)**.** *Let a, b be real numbers such that $a < b$, and let $h : [a, b] \to \mathbb{R}$ be a continuous differentiable function on (a, b) and let h' be it's derivative, then there exists $c \in (a, b)$ such that*

$$h'(c) = \frac{h(b) - h(a)}{b - a}$$

To see this is true, think about drawing a line between $h(b)$ and $h(a)$. Now drag this line up and down on the cartesian plane. At some point as You are dragging it up and down, it will be *just* tangent to the function h. At this point, which We called c, the slope of the function h will be equivalent to the slope between the two points $(b, h(b)), (a, h(a))$.

1.5 Fundamental Theorem of Calculus

Theorem 1.4 (Fundamental Theorem of Calculus)**.** *Let a, b be real numbers such that $a < b$ and let $h : \mathbf{dom}(h) \to \mathbb{R}$ be a differentiable function in the interval (a, b), let h' be that derivative and let h' be continuous on $[a, b]$, then*

$$h(b) - h(a) = \int_b^a h'(t)dt.$$

Tale as old as time.

1.6 Differentiability

Definition 1.5 (Differentiability)**.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}^m, \mathbf{dom}(f) \subset \mathbb{R}^d$, then f is called* differentiable *at $\mathbf{x}$ in the interior of $\mathbf{dom}(f)$ if there exists a matrix $A \in \mathbb{R}^{m \times d}$ and an error function $r : \mathbb{R}^d \to \mathbb{R}^m$ in some neighbourhood of $\mathbf{0} \in \mathbb{R}^d$ such that $\forall\ \mathbf{y}$ in some neighbourhood of $\mathbf{x}$:*

$$f(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x})$$

where

$$\lim_{\mathbf{v} \to 0} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|}$$

It then also follows that A is unique and We call A the differential *or* Jacobian *of f at $\mathbf{x}$, and more accurately A is a matrix of partial derivatives such that*

$$Df(\mathbf{x})_{i,j} = \frac{\partial f(\mathbf{x})_i}{\partial \mathbf{x}_j}$$

Finally, f is called differentiable *if it is differentiable at all points in it's domain.*

So, remarks: the idea is that f is differentiable if it is approximated arbitrarily well by some linear function ($f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x})$) and that the error of this approximation is sublinear, i.e. as size of the difference $\mathbf{y} - \mathbf{x}$ decreases linearly, the error term $r(\mathbf{y} - \mathbf{x})$ decreases even faster.

There are also some notation to be aware of - $\nabla f(\mathbf{x})$ is the gradient vector and it is a*column*, and $Df(\mathbf{x}) = \nabla f(\mathbf{x})^\top$, so then We can write stuff like $f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{y} - \mathbf{x})^\top + r(\mathbf{y} - \mathbf{x})$

Lemma 1.6 (Chain Rule)**.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}^m, \mathbf{dom}(f) \in \mathbb{R}^d$ and $g : \mathbf{dom}(g) \to \mathbb{R}^d$, g be differentiable at $\mathbf{x} \in \mathbf{dom}(g)$ and f be differentiable at $g(\mathbf{x})$, then*

$$Df(g(\mathbf{x})) = D(f \circ g)(\mathbf{x}) = Df(g(\mathbf{x}))Dg(\mathbf{x})$$

It kind of can't be a lot of other things. Don't think about how the information flows through the matrix - it *does*, but 1 variable of g affects all variables in f, and it's kind of a pain.

2 Convex Sets

Definition 2.1 (Convex Set)**.** *A set $C \in \mathbb{R}^d$ is* convex *if $\forall\ \mathbf{x}, \mathbf{y} \in C$ it is true that $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in C, \lambda \in [0, 1]$.*

Not incredible. Idea is that a set is convex if for all two points in that set, the line between those two points is also in the set, so the shape of the set has no dents etc.

Observation 2.2. *Let $C_i, i \in I$ be convex sets and I be a potentially infinite index set, then $C = \cap_{i \in I} C_i$ is a convex set.*

Proof. $\forall\ \mathbf{x}, \mathbf{y} \in C$ that must mean that $\mathbf{x}, \mathbf{y} \in \cap_{i \in I} C_i$, and by virtue of convexity the line between those points must also be in the intersection. □

2.1 Mean Value Inequality

Is a relaxation of the mean value theorem. Really it's Lipschitz but from the perspective of mean value theorem:

Suppose $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable over the interval $X \subset \mathbf{dom}(f)$.