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 $A = Df(\mathbf{x}) = \nabla f(\mathbf{x})^\top$, so then We can write stuff like $f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x})$

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 open interval $X \subset \mathbf{dom}(f)$, and suppose that $\forall \mathbf{x} \in X \implies |f'(\mathbf{x})| \leq B$, then We can say

$$|f(\mathbf{y}) - f(\mathbf{x})| = |f'(c)(\mathbf{y} - \mathbf{x}) \leq B|\mathbf{y} - \mathbf{x}|$$

Which, recall, $f'(c)$ is just the slope of the the line between $f(\mathbf{y}) - f(\mathbf{x})$, and We know it exists by the Mean Value Theorem.

So, really, (sort of by definition almost really), any function whose derivative is bounded by B is B-Lipschitz over that whatever interval X.

So We went from mean theorem to Lipschitz, sort of. Ugh the notes do this:

$$|f'(c)| \leq \lim_{\delta \to 0} \left| \frac{f(c + \delta) - f(c)}{\sigma} \right| \leq B$$

Which, like, yeah, the value of the derivative is bounded in a Lipchitz function. Shocking. I feel like this is a trivial observation but off We go to prove it (ah it's sorta nontrivial since We are not in the univariate case and have to deal with the Spectral norm):

Theorem 2.3. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a continuous differentiable function over some $X \subseteq \mathbf{dom}(f), B \in \mathbb{R}_+$, where X is open and convex, then the following two statements are equivalent:*

$$\begin{aligned} &i)\ \|f(\mathbf{y}) - f(\mathbf{x})\| \leq B \cdot \|\mathbf{y} - \mathbf{x}\|, \qquad \forall \mathbf{x}, \mathbf{y} \in X. \\ &ii)\ |Df(\mathbf{x})\| \leq B, \qquad \forall \mathbf{x} \in X. \end{aligned}$$

Proof. First, let's assume $i)$ and prove $ii)$:

$$\|f(\mathbf{y}) - f(\mathbf{x})\| \leq B \cdot \|\mathbf{y} - \mathbf{x}\| \tag{1}$$
$$\tag{2}$$

Recall that We have

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}) \\ f(\mathbf{y}) - f(\mathbf{x}) &= \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}) \end{aligned}$$

Using the above

$$\begin{aligned}
\|f(\mathbf{y}) - f(\mathbf{x})| &\leq B \cdot \|\mathbf{y} - \mathbf{x}\| \\
\|\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x})\| &\leq B \cdot \|\mathbf{y} - \mathbf{x}\| \\
\|\nabla f(\mathbf{x})^\top (\mathbf{x} + \mathbf{v} - \mathbf{x}) + r(\mathbf{x} + \mathbf{v} - \mathbf{x})\| &\leq B \cdot \|\mathbf{x} + \mathbf{v} - \mathbf{x}\|, \ (\mathbf{y} = \mathbf{x} + \mathbf{v}) \\
\|\nabla f(\mathbf{x})^\top (\mathbf{v}) + r(\mathbf{v})\| &\leq B \cdot \|\mathbf{v}\| \\
\|\nabla f(\mathbf{x})^\top (\mathbf{v}) + r(\mathbf{v})\| \leq \|\nabla f(\mathbf{x})^\top (\mathbf{v})\| + \|r(\mathbf{v})\| &\leq B \cdot \|\mathbf{v}\| \\
\|\nabla f(\mathbf{x})^\top (\mathbf{v})\| + \|r(\mathbf{v})\| &\leq B \cdot \|\mathbf{v}\| \\
\frac{\|\nabla f(\mathbf{x})^\top (\mathbf{v})\|}{\|\mathbf{v}\|} + \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} &\leq \frac{B \cdot \|\mathbf{v}\|}{\|\mathbf{v}\|} \\
\frac{\|\nabla f(\mathbf{x})^\top (\mathbf{v})\|}{\|\mathbf{v}\|} &\leq B - \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|}
\end{aligned}$$

And now comes some shenanigans. Let $\mathbf{v} = t \cdot \mathbf{v}^*$ such that $\mathbf{v}^*$ is a unit vector and

$$\frac{\|\nabla f(\mathbf{x})^\top (\mathbf{v}^*)\|}{\|\mathbf{v}^*\|} = \|\nabla f(\mathbf{x})\|$$

Then We just do

$$\begin{aligned}
\frac{\|\nabla f(\mathbf{x})^\top (\mathbf{v})\|}{\|\mathbf{v}\|} &\leq B - \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} \\
\frac{\|\nabla f(\mathbf{x})^\top t(\mathbf{v}^*)\|}{\|t\mathbf{v}^*\|} &\leq B - \frac{\|r(t\mathbf{v}^*)\|}{\|t\mathbf{v}^*\|} \\
\frac{|t|\|\nabla f(\mathbf{x})^\top (\mathbf{v}^*)\|}{|t|\|\mathbf{v}^*\|} &\leq B - \frac{\|r(t\mathbf{v}^*)\|}{\|t\mathbf{v}^*\|} \\
\frac{\|\nabla f(\mathbf{x})^\top (\mathbf{v}^*)\|}{\|\mathbf{v}^*\|} &\leq B - \frac{\|r(t\mathbf{v}^*)\|}{\|t\mathbf{v}^*\|} \\
\|\nabla f(\mathbf{x})\| &\leq B - \frac{\|r(t\mathbf{v}^*)\|}{\|t\mathbf{v}^*\|}
\end{aligned}$$

and as $t\mathbf{v}^* \to \mathbf{0}$, We get our result.

Now for the other direction, let's create the following function:

$$g(t) = \mathbf{f}(\mathbf{x}) + \mathbf{t}(\mathbf{y} - \mathbf{x})$$

This simply gives us smooth means of traversal through the function, which will allow us to use the bounded differentials. We then define $h(t) = f(g(t))$, so We are just wrapping our traversal function with, well, the function of interest f. This allows us to say

$$
\begin{aligned}
\|f(\mathbf{y}) - f\mathbf{x})\| &= \|h(1) - h(0)\| \\
&= \left\| \int_{t=0}^{1} h'(t)dt \right\| \\
&= \left\| \int_{t=0}^{1} D(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})dt \right\| \\
&\leq \int_{t=0}^{1} \|D(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})\|dt \quad \text{Jensen's Inequality} \\
&\leq \int_{t=0}^{1} \|D(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\|\|(\mathbf{y} - \mathbf{x})\|dt \quad \text{Spectral Norm} \\
&\leq \int_{t=0}^{1} B\|(\mathbf{y} - \mathbf{x})\|dt \quad \text{Bounded differentials} \\
&\leq B\|(\mathbf{y} - \mathbf{x})\|
\end{aligned}
$$

□

3 Convex Functions

Definition 3.1 (Convex Function)**.** *A function* $f : \mathbf{dom}(f) \to \mathbb{R}$ *is a* convex function *if i)* $\mathbf{dom}(f)$ *is a convex set and ii) if* $\forall\ \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ *it is true that*

$$f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})$$

for $\lambda \in [0, 1]$.

A useful concept to go along with the convex function is the epigraph:

$$\mathbf{epi}(f) = \{(\mathbf{x}, \alpha) \in \mathbb{R}^{d+1} : \mathbf{x} \in \mathbf{dom}(f), \alpha \geq f(\mathbf{x})\}.$$

So just the set of all points above the graph of the function f.

Observation 3.2. *Function* $f : \mathbf{dom}(f) \to \mathbb{R}$ *is convex if and only if* $\mathbf{epi}(f)$ *is a convex set.*

Proof. First direction, We can assume convex f:

$$f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})$$

We need to prove:

$$\forall(\mathbf{x}, \alpha_x),\ (\mathbf{y}, \alpha_y) \in \mathbf{epi}(f) : \lambda(\mathbf{x}, \alpha_x) + (1 - \lambda)(\mathbf{y}, \alpha_y) \in \mathbf{epi}(f)$$

So We have the point $(\underbrace{\lambda\mathbf{x}+(1-\lambda)\mathbf{y}}_{q}, \underbrace{\lambda\alpha_x+(1-\lambda)\alpha_y}_{w})$. Now We just need to assert that $q \in \mathbf{dom}(f), w \geq f(q)$ to satisfy the epigraph conditions. We know that $q \in \mathbf{dom}(f)$ by convexity of the domain of f, and for the second bit

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda\alpha_x+(1-\lambda)\alpha_y$$
$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$$

Which is true by convexity of f. Wee.

Now for the other side - We can assume that $\mathbf{epi}(f)$ is a convex set, and We wish to prove that f is convex, so We want

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$$

This in words is quite trivial - take the two endpoints, they by definition are in the epigraph, then consider all points in between and observe that since they too are in the epigraph since the epigraph is convex they must be lie above the graph, concluding the proof. □

Lemma 3.3 (Jensen's Inequality)**.** *For a convex function $f : \mathbf{dom}(f) \to \mathbb{R}$ and a random variable X such that X has a mean and support$(X) \in \mathbf{dom}(f)$ it is true that*

$$E(f(X)) \leq f(E(X))$$

Proof. Let $\mu = E(X)$, and construct a tangent $a+bX$ to the convex function X at μ such that $a+b\mu = f(\mu)$. By convexity of f We have

$$\begin{aligned} f(X) &\geq a+bX \\ E(f(X) &\geq E(a+bX) \\ E(f(X) &\geq a+bE(X) \\ E(f(X) &\geq a+b\mu \\ E(f(X)) &\geq f(\mu) \\ E(f(X)) &\geq f(E(X)) \end{aligned}$$

□

3.1 First Order Characterization of Convexity

Lemma 3.4. *Suppose f is differentiable, then f is also convex if and only if $\forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$*

$$f(\mathbf{y}) \geq f(\mathbf{x})+\nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})$$

And intuitively this just means that the graph of f is above the tangent to f at any point.

Proof. First direction relies on rearranging convexity:

$$\begin{aligned}
f(\lambda\mathbf{x} + (1-\lambda)(\mathbf{y})) &\leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \\
f(\lambda\mathbf{x} + \mathbf{y} - \lambda\mathbf{y}) &\leq \lambda f(\mathbf{x}) + f(\mathbf{y}) - \lambda f(\mathbf{y}) \\
f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) &\leq f(\mathbf{y}) + \lambda(f(\mathbf{x}) - f(\mathbf{y})) \\
f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) &\leq \lambda(f(\mathbf{x}) - f(\mathbf{y})) \\
f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) &\leq \lambda(f(\mathbf{x}) - f(\mathbf{y})) \\
\frac{f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{\lambda} &\leq f(\mathbf{x}) - f(\mathbf{y}) \\
\frac{f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{\lambda} + f(\mathbf{y}) &\leq f(\mathbf{x}) \\
\frac{\nabla f(\mathbf{x})^\top \lambda(\mathbf{x} - \mathbf{y}) + r(\lambda(\mathbf{x} - \mathbf{y}))}{\lambda} + f(\mathbf{y}) &\leq f(\mathbf{x}) \\
\frac{\nabla f(\mathbf{x})^\top \lambda(\mathbf{x} - \mathbf{y})}{\lambda} + \frac{r(\lambda(\mathbf{x} - \mathbf{y}))}{\lambda} + f(\mathbf{y}) &\leq f(\mathbf{x}) \\
\nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{y}) + f(\mathbf{y}) \leq f(\mathbf{x}) \quad \lim_{t \to 0}
\end{aligned}$$

So two pieces to this one - reframing convexity as having one "fixed" point (in this case $\mathbf{y}$, and observing that

$$f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) = \nabla f(\mathbf{y})^\top \lambda(\mathbf{x} - \mathbf{y}) + r(\lambda(\mathbf{x} - \mathbf{y}))$$

Since We are calculating the difference from $f(\mathbf{y})$ when our intput is shifted by $\lambda(\mathbf{x} - \mathbf{y})$. □

Proof. Now for the other direction, i.e. first order characterization implies convexity.

First We establish some point $\mathbf{z}$ in between $\mathbf{x}, \mathbf{y}$:

$$\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$$

By first order characterization We then have

$$\begin{aligned}
f(\mathbf{x}) &\geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) \\
f(\mathbf{y}) &\geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z})
\end{aligned}$$

Then We start constructing our conclusion by multiplying the first line by λ and the second by $(1-\lambda)$:

$$\lambda f(\mathbf{x}) \geq \lambda f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z})$$
$$(1-\lambda) f(\mathbf{y}) \geq (1-\lambda) f(\mathbf{z}) + (1-\lambda) \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z})$$

And then We add those two and simplify

$$\begin{aligned}
&\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq \\
&\lambda f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) + (1-\lambda) f(\mathbf{z}) + (1-\lambda) \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) \\
&\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq \\
&\lambda f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) + f(\mathbf{z}) - \lambda f(\mathbf{z}) + (1-\lambda) \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) \\
&\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq \\
&f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}) + (1-\lambda) \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) \\
&\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq \\
&f(\mathbf{z}) + \nabla f(\mathbf{z})^\top \Big(\lambda(\mathbf{x} - \mathbf{z}) + (1-\lambda)(\mathbf{y} - \mathbf{z}) \Big)
\end{aligned}$$

Now We just focus on those gradient terms:

$$\begin{aligned}
\lambda(\mathbf{x} - \mathbf{z}) + (1-\lambda)(\mathbf{y} - \mathbf{z}) &= \lambda \mathbf{x} - \lambda \mathbf{z} + \mathbf{y} - \mathbf{z} - \lambda \mathbf{y} + \lambda \mathbf{z} \\
&= \lambda \mathbf{x} + (-\lambda - 1 + \lambda)\mathbf{z} + \mathbf{y} - \lambda \mathbf{y} \\
&= \lambda \mathbf{x} + (1-\lambda)\mathbf{y} - \mathbf{z} \\
&= 0
\end{aligned}$$

So there's really one bit to to this proof - take the tangent at the middle point $\mathbf{z}$ and compare it with the endpoints $f(\mathbf{x}), f(\mathbf{y})$, and from there it's addition and simplification. □

Only really useful if You have a derivative. I think in later chapters this is generalized with subgradients. Furthermore there are obviously convex functions like the absolute value operator that are not differentiable.

Lemma 3.5. *An alternative first-order characterization of convexity is:*

$$\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \geq 0$$

A loose intuition for this is that whatever direction You go in ($\mathbf{y} - \mathbf{x}$, the gradient will also point in that same direction.

Proof. By convexity We have these two statements

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$$

Adding those two and cancelling out $f(\mathbf{y}), f(\mathbf{x})$ gives

$$\begin{aligned}
0 &\geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \\
0 &\geq -\nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \\
0 &\geq (\mathbf{x} - \mathbf{y})(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top \\
0 &\leq (\mathbf{y} - \mathbf{x})(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top
\end{aligned}$$

Which concludes this direction. All We needed to do was "look" at each point from the direction of the other point and combine the perspectives.

Now for the other direction, We have

$$\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \geq 0$$

Let $\mathbf{y} = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$ to yield (note that the above holds for $t \in [0, 1]$)

$$\begin{aligned}
\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top (\mathbf{x} + t(\mathbf{y} - \mathbf{x}) - \mathbf{x}) &\geq 0 \\
\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top t(\mathbf{y} - \mathbf{x}) &\geq 0 \\
\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) &\geq 0 \\
\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) &\geq 0 \\
\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) &\geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})
\end{aligned}$$

Now We define

$$h(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

and therefore

$$h'(t) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x})$$

Which We can use to rewrite

$$\begin{aligned}
\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \top (\mathbf{y} - \mathbf{x}) &\geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\
h'(t) &\geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})
\end{aligned}$$

And this still holds for all $t \in [0, 1]$.

And now comes just random stuff plucked from the air: observe that by the Mean Value Theorem, it is true that there exists a c such that $h'(c) = h(1) - h(0)$.

$$f(\mathbf{y}) = h(1) = h(0) + h(1) - h(0) = h(0) + h'(c)$$
$$f(\mathbf{y}) \geq h(0) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

Honestly no fun and seemingly unmotivated proof. □

3.2 Second-order characterization of convexity

Lemma 3.6. *If f is twice differentiable on it's domain, namely*

$$\nabla^2 f(\mathbf{x})_{i.j} = \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}$$

exists and $\nabla^2 f(\mathbf{x})$ is positive definite (or positive semidefinite), then f is convex.

Proof. Is omitted, at least for now. □

3.3 Operations that preserve convexity

Lemma 3.7. *i) Multiplying by a positive constant, addition of convex functions while taking the intersection of their domains preserves convexity.*

ii) Let g be a linear function and f be a convex function, then $f \circ g$ is also convex.

4 Minimizing Convex Functions

Definition 4.1. *A function $f : \mathbf{dom}(f) \to \mathbb{R}$ has a local minimum $\mathbf{x} \in \mathbf{dom}(f)$ if there exists $\varepsilon > 0$ such that*

$$f(\mathbf{y}) \geq f(\mathbf{x}) \qquad \forall \mathbf{y} \in \mathbf{dom}(f),\ \|\mathbf{y} - \mathbf{x}\| < \varepsilon$$

In this case thing of ε as a sort of radius for the local minimum. All points in the domain within that radius are greater than $f(\mathbf{x})$, so We have a neat little local dent (or plateau).

Lemma 4.2. *Let $\mathbf{x}^*$ be a local minimum to a convex function $f : \mathbf{dom}(f) \to \mathbb{R}$, then $\mathbf{x}^*$ is a global minimum meaning*

$$f(\mathbf{x}^*) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f).$$

Proof. Is simple - let $\mathbf{y} \in \mathbf{dom}(f)$ be such that $f(\mathbf{y}) < f(\mathbf{x}^*)$. Let $\mathbf{y}' = \lambda \mathbf{y} + (1 - \lambda)\mathbf{x}^*$. Then for $\lambda \in (0, 1)$ by convexity We have $f(\mathbf{y}') < f(\mathbf{x}^*)$. Choosing λ small enough to violate $\|\mathbf{y}' - \mathbf{x}^*\| < \varepsilon$ contradicts local minima assumption of $\mathbf{x}^*$. □

Then there's a bunch of caveats. Convex functions need not have a global minimum - $f(x) = x$ etc. Then even if it is bounded from below, asymptotic functions won't have a minimum e.g. $f(x) = e^x$.

Lemma 4.3. *Suppose $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable, and let $\mathbf{x} \in \mathbf{dom}(f)$. If $\nabla f(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is a global minimum.*

Proof. Assuming $\nabla f(\mathbf{x}) = \mathbf{0}$, with first order characterization We have

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\ &\geq f(\mathbf{x}) + \mathbf{0} \\ &\geq f(\mathbf{x}) \end{aligned}$$

□

Lemma 4.4. *Suppose f is convex and differentiable and let $\mathbf{x} \in \mathbf{dom}(f)$ be a global minimum, then it is true that $\nabla f(\mathbf{x}) = \mathbf{0}$.*

4.1 Strictly Convex Functions

Definition 4.5. *A function $f : \mathbf{dom} \to \mathbb{R}$ is strictly convex if it is true $\forall\, \mathbf{x} \neq \mathbf{y} \in \mathbf{dom}(f),\ \lambda \in (0,1)$ that*

$$f(\lambda(\mathbf{x}) + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Lemma 4.6. *The second derivative matrix a.k.a. Hessian of a strictly convex function is positive definite (if it exists).*

Lemma 4.7. *Suppose f is strictly convex, then f has a unique global minimum.*

Proof. Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a strictly convex function. Suppose there are two global minima $\mathbf{x} \neq \mathbf{y} \in \mathbf{dom}(f)$. Then by taking a linear combination of those two minima, by strict convexity, We obtain an even lower point, violating the assumption that $\mathbf{x}, \mathbf{y}$ were minima. □

4.2 Constrained Optimization

Definition 4.8. *Let $f : \mathbf{x} \to \mathbb{R}$ be some function, $X \subseteq \mathbf{dom}(f)$ and $\mathbf{x} \in X$. If*

$$f(\mathbf{y} \geq f(\mathbf{x}) \quad \forall \mathbf{y} \in X$$

Then $\mathbf{x}$ is a minimizer of f over X.

Lemma 4.9. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a convex differentiable function, $X \subseteq \mathbf{dom}(f)$, then $\mathbf{x}^* \in X$ is a minimizer of f over X if and only if*

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0 \quad \forall \mathbf{x} \in X.$$

Basically no direction exists such that the gradient is negative from our optimal point. Makes sense - if the above was not true, We could take a step against the direction of the gradient and obtain a smaller minimum.

5 Existence of a minimizer

We care about this since most algorithms assume the existence of a minimizer. Here are some ways of going about proving such a minimizer exists:

5.1 Sublevel sets and Weierstrass Theorem

Definition 5.1. *Let $f : \mathbb{R}^d \to \mathbb{R}, \alpha \in \mathbb{R}$, the set*

$$f^{\leq \alpha} = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x} \leq \alpha\}$$

is called the α-sublevel set of f.

So just the closed (due to the $\leq$ condition) subset of the domain of f. Cool.

Theorem 5.2. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function with a bounded nonempty sublevel set $f^{\leq \alpha}$, then f has a global minimum.*

Proof. We know that the sublevel set is continuous since it is a subset of the convex domain of f. Furthermore We know it is bounded, so somewhere along that continuous line a minimum is achieved, and since all points outside of the sublevel set are larger, We've got a global minimum. □

5.2 Recession cone and lineality space

The idea is that there is probably some "direction of recession" in our convex set C, and that is the reason the sublevel set is not bounded. Think of the case $f(x) = x$ - the line with slope -1 is a direction of recession. Formally:

Definition 5.3. *Let C be a convex set, then $\mathbf{y} \in \mathbf{R}^d$ is a direction of recession if for some $\mathbf{x} \in C$ and $\lambda \in \mathbb{R}_+$ it holds that $\mathbf{x} + \lambda\mathbf{y} \in C$.*

Basically the set extends infinitely in that direction.

Lemma 5.4. *For a closed convex set C with $\mathbf{y} \in C$, then the following statements are equivalent:*

$$i)\ \exists \mathbf{x} \in C : \mathbf{x} + \lambda\mathbf{y} \in C,\ \forall \lambda \geq 0$$
$$ii)\ \forall \mathbf{x} \in C : \mathbf{x} + \lambda\mathbf{y} \in C,\ \forall \lambda \geq 0$$

Proof. We only need to prove $i) \to ii)$. First assume

$$\exists \mathbf{x} \in C : \mathbf{x} + \lambda\mathbf{y} \in C,\ \forall \lambda \geq 0$$

Now define $w_k = \mathbf{x} + k\lambda\mathbf{y}, k \in \mathbb{N}$. $w_k \in C$ by our assumption.

Now, We need to show that $\mathbf{x}' + \lambda\mathbf{y} \in C$ for any $\mathbf{x}'$. To this end suppose We have

$$\frac{1}{k}w_k + (1 - \frac{1}{k})\mathbf{x}'$$

And then let k go to infinity.

So what did We do? We took the direction of recession, stretched it out infinitely by some factor k, and then formulated a line between the infinite point and any other point $\mathbf{x}'$ using the same factor We used to stretch out the line. This way, when taking limits, the factors cancel and convexity gives us membership in the closed set. □

Lemma 5.5. *Let C be a closed convex set and let $\mathbf{y}_1, \mathbf{y}_2 \in C$ be directions of recession of C, then $\mathbf{y} = \lambda_1\mathbf{y}_1 + \lambda_2\mathbf{y}_2$, $lambda_1\lambda_2 \in \mathbb{R}_+$ is also a direction of recession in C.*

Proof. Scale both $\mathbf{y}$ by $1/(\lambda_1 + \lambda_2)$, which allows us to assume without loss of generality that $\lambda_1 + \lambda_2 = 1$. Anyway, now the idea is to simply

$$\begin{aligned}\mathbf{x} + \lambda\mathbf{y} = \mathbf{x} + \lambda(\lambda_1\mathbf{y}_1 + \lambda_2\mathbf{y}_2) &= \mathbf{x} + \lambda\lambda_1\mathbf{y}_1 + \lambda\lambda_2\mathbf{y}_2 \\ &= \lambda_1\mathbf{x} + \lambda_2\mathbf{x} + \lambda\lambda_1\mathbf{y}_1 + \lambda\lambda_2\mathbf{y}_2 \\ &= \lambda_1(\mathbf{x} + \lambda\mathbf{y}_1) + \lambda_2(\mathbf{x} + \lambda\mathbf{y}_2)\end{aligned}$$

Since both endpoints are in the set, We are done. So the idea there was to split the new direction of recession into it's original components and since the result will be between the two original directions, by convexity the new point is also in C. The scaling thing is there to make it neat.

□

Definition 5.6. *Let $C \subseteq R^d$ be a closed convex set, then $\mathbf{y} \in C$ is a* direction of constancy *if both $\mathbf{y}, -\mathbf{y}$ are directions of recession.*

So, a direction of constancy has with it associated a linear subspace called *lineality space L(C)*. The set shape here is something like cylinder for example. Anyway, lineality spaces are sort of straight lines, where as recession stuff is more like a cone, since it only deals with one direction (especially in convex cases, You're getting a cone). Remember that We are still just dealing with sets, though.

Lemma 5.7. *Let $C \subseteq \mathbb{R}^d$ be a closed convex set and $\mathbf{y}_1, \mathbf{y}_2$ be directions of constancy, then $\mathbf{y} = \lambda_1\mathbf{y}_1 + \lambda_2\mathbf{y}_2, \lambda_1, \lambda_2 \in \mathbb{R}$ is also a direction of constancy.*

Proof. Simply use above lemma and deal with positive and negative directions one at a time. □

So how do We use this stuff for functions? Well, We have our sublevel sets. If the sublevel sets are well behaved, then We have a global minimum.

Lemma 5.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, then any two non-empty sublevel sets $f^{\leq\alpha}, f^{\leq\alpha'}, \alpha \neq \alpha'$ share the same directions of recession.*

Proof. We first assume that a sublevel set $f^{\leq\alpha}$ has a direction of recession $\mathbf{y}$ giving us

$$f(\mathbf{x} + \lambda\mathbf{y}) \leq \alpha$$

The claim is that We can actually state a stronger version of the above:

$$f(\mathbf{x} + \lambda\mathbf{y}) \leq f(\mathbf{x})$$

To prove this We let $\mathbf{z} = \lambda\mathbf{y}$, $\mathbf{w}_k = \mathbf{x} + k\lambda\mathbf{y} \in f^{\leq\alpha}$ We get

$$\mathbf{x} + \mathbf{z} = \left(1 - \frac{1}{k}\right)\mathbf{x} + \frac{1}{k}\mathbf{w}_k$$

Which We can then use to say

$$\begin{aligned} f(\mathbf{x} + \mathbf{z}) = f\left(\left(1 - \frac{1}{k}\right)\mathbf{x} + \frac{1}{k}\mathbf{w}_k\right) &\leq \left(1 - \frac{1}{k}\right)f(\mathbf{x}) + \frac{1}{k}f(\mathbf{w}_k) \\ &\leq \left(1 - \frac{1}{k}\right)f(\mathbf{x}) + \frac{1}{k}f(\mathbf{w}_k) \\ &\leq \left(1 - \frac{1}{k}\right)f(\mathbf{x}) + \frac{1}{k}\alpha \\ &\leq f(\mathbf{x}) \quad \left(\lim_{k\to\inf}\right) \end{aligned}$$

The trick with directions of recession seems to be to allow an endpoint extend to infinity, and this allows for terms to vanish. In this case, We chose the endpoints such that We could use our definition of convexity.

Anywho, once We have the non-increasing term, it's simple. Since they non-empty sublevel sets, they must have some point in their intersection $\mathbf{x}'$, and by our non-increasing fact We can say $f(\mathbf{x}' + \lambda\mathbf{y}) \leq f(\mathbf{x}')$, and so it is in any sublevel set We'd like. □

Definition 5.9. *R(f) is the set of directions of recession, L(f) is the set of f.*

Having all of the above, We can make the following statement:

Lemma 5.10. *Let $f\mathbb{R}^d \to \mathbb{R}$ be a convex function, then the following are equivalent:*

i) $\mathbf{y} \in \mathbb{R}^d$ *is a direction of recession of f.*

ii) $f(\mathbf{x} + \lambda\mathbf{y}) \leq f(\mathbf{x}), \forall\mathbf{x} \in \mathbb{R}^d, \lambda \geq 0.$

iii) $(\mathbf{y}, 0) \in \mathbb{R}^{d+1}$ *is a "horizontal" direction of recession of the convex set* $\mathbf{epi}(f)$.

The first two are clear.

To understand the third one, first think of the epigraph. In our usual case of f having a single output dimension, We got our graph on the cartesian plane, and the epigraph was everything above the graph. No big deal.

Now, observe that f in this case was one dimensional, and the epigraph is two dimensional since We need to add another dimension to the points to account for the output of the function f.

Now think about directions *in the epigraph.* We have a direction in the input space - $\mathbf{y}$. No worries there. In our 2D example, that's either left or right lol. By creating the vector $(\mathbf{y}, 0) \in \mathbb{R}^{d+1}$, We are saying that as You move along $\mathbf{y}$ in the input space, the output value does not change. In the epigraph, it is "horizontal".

5.3 Coercive functions

Aren't very interesting in that they are too nice:

Definition 5.11. *A $f\mathbb{R}^d \to \mathbb{R}$ convex function is* coercive *if it's recession cone is trivial, i.e. $R(f) = \mathbf{0}$.*

So no matter which way You go, the function is increasing.

Lemma 5.12. *Let $f : \mathbb{R}^d$ be a coercive function, then every nonempty sublevel set $f^{\leq \alpha}$ is bounded.*

Proof. The proof is a nightmare so We are just going to skip it. Probably not a nightmare if You've done real analysis though. □

The point is that We claim that coercive functions have bounded sublevel sets, and We know that bounded sublevel sets obviously have a global minimum, so:

Theorem 5.13. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a coercive function, then f has a global minimum.*

5.4 Weakly Coercive Functions

Are simply functions such that their recession cone and lineality space are the same set. We then come back to having coercive functions by taking the compliment of the lineality space of a weakly coercive function f.

6 Exercises

Q2

Let $f : \mathbf{dom}(f) \to \mathbb{R}$, $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ be a convex function, our goal is to show that

$$f\Big(\sum_{i=1}^{N} w_i \mathbf{x}_i\Big) \leq \sum_{i=1}^{N} w_i f(\mathbf{x}_i)$$

Where $\sum_{i=1}^{N} w_i = 1, w_i \geq 0$ and $\mathbf{x}_i \in \mathbf{dom}(f)$.

Now construct a random variable X such that $P(X = \mathbf{x}_i) = w_i$, so $E(X) = \sum_{i=1}^{N} \mathbf{x}_i w_i = \mu$.

By convexity of f We have:

$$f(\mathbf{y}) \geq f(\mu) + \mathbf{g}_\mu^\top (\mathbf{y} - \mu) \; \forall \mathbf{y} \in \mathbf{dom}(f)$$

Where, $\mathbf{g}_m u \in \partial f(\mu)$, and $\partial f(\mu)$ is of course the subgradient set of f at μ.

Now let us construct a new random variable $Y = f(\mu) + \mathbf{g}_\mu^\top (X - \mu)$. We can then say. By the previous inequality and the fact that X, Y are constructed from the same random variable:

$$\begin{aligned}
E(f(X)) &\geq E(Y) && (1)\\
&\geq E(f(\mu) + \mathbf{g}_\mu (X - \mu)) && (2)\\
&\geq f(\mu) + \mathbf{g}_\mu (E(X) - \mu) && (3)\\
&\geq f(\mu) + \mathbf{g}_\mu (\mu - \mu) && (4)\\
&\geq f(\mu) && (5)\\
&\geq f(E(X)) && (6)
\end{aligned}$$

We can probably do this without using probability stuff.

$$\begin{aligned}
f(\mathbf{y}) &\geq f(\mu) + \mathbf{g}_\mu^\top (\mathbf{y} - \mu) \; \forall \mathbf{y} \in \mathbf{dom}(f) && (7)\\
f(\mathbf{x}_i) &\geq f(\mu) + \mathbf{g}_\mu^\top (\mathbf{x}_i - \mu) \; \forall i \in \{1 \dots N\} && (8)\\
w_i \cdot f(\mathbf{x}_i) &\geq w_i \cdot \big(f(\mu) + \mathbf{g}_\mu^\top (\mathbf{x}_i - \mu)\big) \; \forall i \in \{1 \dots N\} && (9)\\
\sum_{i=1}^{N} w_i \cdot f(\mathbf{x}_i) &\geq \sum_{i=1}^{N} w_i \cdot \big(f(\mu) + \mathbf{g}_\mu^\top (\mathbf{x}_i - \mu)\big) && (10)\\
\sum_{i=1}^{N} w_i \cdot f(\mathbf{x}_i) &\geq f(\mu) + \sum_{i=1}^{N} w_i \cdot \big(\mathbf{g}_\mu^\top (\mathbf{x}_i - \mu)\big) && (11)
\end{aligned}$$

Focusing on the last term:

$$\sum_{i=1}^{N} w_i \cdot \left(\mathbf{g}_\mu^\top (\mathbf{x}_i - \mu)\right) = \mathbf{g}_\mu^\top \sum_{i=1}^{N} w_i \cdot (\mathbf{x}_i - \mu) \tag{12}$$

$$= \mathbf{g}_\mu^\top \left(\left(\sum_{i=1}^{N} w_i \cdot \mathbf{x}_i \right) - \mu \right) \tag{13}$$

$$= \mathbf{g}_\mu^\top (\mu - \mu) \tag{14}$$

$$= 0 \tag{15}$$

Therefore

$$\sum_{i=1}^{N} w_i \cdot f(\mathbf{x}_i) \geq f(\mu) + \sum_{i=1}^{N} w_i \cdot \left(\mathbf{g}_\mu^\top (\mathbf{x}_i - \mu)\right) \tag{16}$$

$$\sum_{i=1}^{N} w_i \cdot f(\mathbf{x}_i) \geq f(\mu) + 0 \tag{17}$$

$$\sum_{i=1}^{N} w_i \cdot f(\mathbf{x}_i) \geq f(\mu) \tag{18}$$

$$\sum_{i=1}^{N} w_i \cdot f(\mathbf{x}_i) \geq f(\sum_{i=1}^{N} \mathbf{x}_i w_i) \tag{19}$$

Q4

Alright We are trying to show that $f(\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is a strictly convex function.

We just need to show that it's second derivative is positive definite.

Let's write the function in basic terms We can differentiate:

$$\|\mathbf{x} - \mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \tag{20}$$

$$= \sum_{i=1}^{d} (\mathbf{x}_i - \mathbf{y}_i)^2 \tag{21}$$

Then taking the derivative will yield a vector $\frac{\partial f}{\partial \mathbf{y}}$ such that

$$\frac{\partial f}{\partial \mathbf{x}_i} = \frac{\partial}{\partial \mathbf{x}_i} \sum_{i=1}^{d} (\mathbf{x}_i - \mathbf{y}_i)^2 \tag{22}$$

$$= \frac{\partial}{\partial \mathbf{x}_i} (\mathbf{x}_i - \mathbf{y}_i)^2 \tag{23}$$

$$= \frac{\partial}{\partial \mathbf{x}_i} \left(\mathbf{x}_i^2 - 2\mathbf{x}_i \mathbf{y}_i + \mathbf{y}_i^2\right) \tag{24}$$

$$= 2\mathbf{x}_i - 2\mathbf{y}_i \tag{25}$$

Now We need to differentiate this vector. This will result in a two dimensional square matrix of size $d \times d$. Let's call that matrix H for Hessian. Recall by convention (and convenience) We have

$$H_{i,j} = \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j}$$

With the idea being that if You take the dot product of a vector with a row of the Hessian matrix You'll the rate of change of the rate of change of f w.r.t. $\mathbf{x_i}$, and You need d pieces of information to establish the rate of change of the rate of change for each input. Easy to follow, I know.

So!

$$\begin{aligned} H_{i,j} &= \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{x}_j} && (26)\\ &= \frac{\partial f}{\partial \mathbf{x}_i}\frac{\partial f}{\partial \mathbf{x}_j} && (27)\\ &= \frac{\partial f}{\partial \mathbf{x}_i} 2\mathbf{x}_j - 2\mathbf{y}_j && (28)\\ &= \frac{\partial f}{\partial \mathbf{x}_i} 2\mathbf{x}_j - 2\mathbf{y}_j && (29)\\ &= 2 \text{ if } i = j, \text{ 0 otherwise} && (30) \end{aligned}$$

So We have a diagonal matrix with non-negative entries, which is positive definite. so We're done!

Q5

Let $f_1 \dots f_n$ be convex functions, and let $\lambda_1 \dots \lambda_n \in \mathbb{R}_+$, then

$$f = \sum_{i=1}^{n} \lambda_i \cdot f_i$$

is a convex function on the domain $\cap_{i=1}^{n} \mathbf{dom}(f_i)$

Proof. For f to be convex it's domain has to be convex and We also need

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y})$$

Well the domain is an intersection of n convex sets so it is itself convex, no worries there.

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \quad (31)$$

$$\left(\lambda \sum_{i=1}^{n} \lambda_i \cdot f_i(\mathbf{x})\right) + \left((1-\lambda)\sum_{i=1}^{n} \lambda_i \cdot f_i(\mathbf{y})\right) \geq \sum_{i=1}^{n} \lambda_i \cdot f_i(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \quad (32)$$

Now focus on a particular i:

$$\lambda\lambda_i \cdot f_i(\mathbf{x}) + (1-\lambda)\lambda_i \cdot f_i(\mathbf{y}) \geq \lambda_i \cdot f_i(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \tag{33}$$

So We are just ignoring summation terms that are not associated with f_i here. If the above holds for all i then the summation of n such equations preserves our desired inequality. Divide by the positive λ_i:

$$\lambda \cdot f_i(\mathbf{x}) + (1-\lambda) \cdot f_i(\mathbf{y}) \geq f_i(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \tag{34}$$

And We know that this is true by convexity of f_i, so We're done.

For the second equation, simply let $\mathbf{x}, \mathbf{y}$ equal the linear combination version of the inputs. No problems there. In particular, by convexity of f We have

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})$$

Now simply let $\mathbf{x} = A\mathbf{x} + \mathbf{b}$, $\mathbf{y} = A\mathbf{y} + \mathbf{b}$. Since the inequality held for all $\mathbf{x}, \mathbf{y}$, and $g(\mathbf{x}), g(\mathbf{y}) \in \mathbf{dom}(f)$, We're good to go.

Q7

Alright so We have

$$\sum_{\mathbf{x} \in P} \left(\log \left(\sum_{k=0}^{9} e^{(W\mathbf{x})_k} \right) - (W\mathbf{x})_{d(\mathbf{x})} \right)$$

□