EE 364M

Taught by John Duchi
Notes by Chris Fifty

Winter 2024

Contents

1 January 10

1.1 Background on Convex Sets

Definition 1 (Line)**.** *Let x, y be two points, then a line between x, y are all points of the form $\theta x + (1-\theta)y$ for $\theta \in \mathbb{R}$.*

Definition 2 (Affine Set)**.** *Consider a set $S = \{x_1, x_2, ...\}$. Then S is **affine** if it contains all lines through any two points in S.*

Definition 3 (Line Segment)**.** *A **line**, but with $\theta \in [0, 1]$: the line does not extend beyond its endpoints x, y.*

Definition 4 (Convex Set)**.** *A set $S = \{x_1, x_2, ...\}$ such that for any $x, y \in S$, S contains all points on the line segment between x and y.*

1.2 Background on Analysis and Linear Algebra

Definition 5 (Metric Space)**.** *A set S equipped with a distance measure. The distance measure is called a metric and it must satisfy three properties:*

1. $d(x, x) = 0$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Definition 6 (Complete Metric Space)**.** *Let M be a metric space. Then M is complete if every Cauchy sequence converges to a limit that is also in M.*

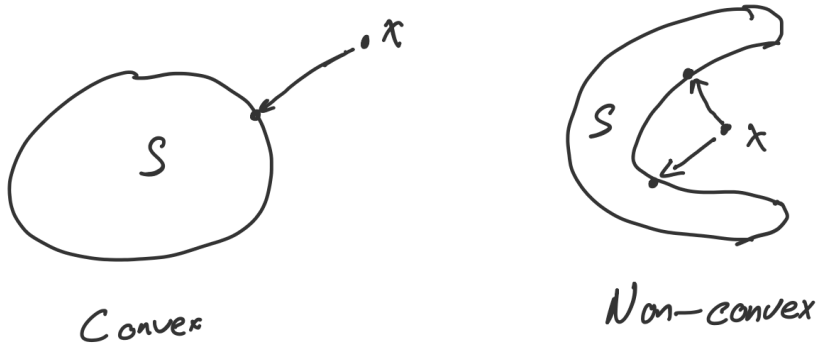

Figure 1: Left: example of a projection of a point x to a convex set S. Right: example of a projection of a point x to a non-convex set S.

Definition 7 (Inner Product)**.** *Given a vector space V over a field F, we can equip this vector space with an inner product. The inner product is simply a function that takes an ordered pair $(u, v) \in V$ to a value $\langle u\ ,\ v\rangle \in F$. This function must satisfy the following properties:*

1. *Positivity: $\langle v\ ,\ v\rangle \geq 0$*
2. *Definiteness: $\langle v\ ,\ v\rangle = 0 \iff v = 0$*
3. *Additivity in the first position: $\langle u + a\ ,\ v\rangle = \langle u\ ,\ v\rangle + \langle a\ ,\ v\rangle$.*
4. *Homogenity in the first position: $\langle \alpha u\ ,\ v\rangle = \alpha\langle u\ ,\ v\rangle$ for $\alpha \in F$.*
5. *Conjugate symmetry: $\langle u\ ,\ v\rangle = \overline{\langle v\ ,\ u\rangle}$*

Definition 8 (Norm)**.** *An inner product induces a norm (a notion of "size") on the vector space: $\|u\| = \sqrt{\langle u\ ,\ u\rangle}$*

Remark 1 (Remark: Norms and Metrics)**.** *A vector space with a norm is called a **normed space**. The norm actually induces a metric space by defining distance to be the norm of the difference between two vectors: $d(x, y) = \|x - y\|$*

Definition 9 (Hilbert Space)**.** *A Hilbert space is simply a complete inner product space. It is a vector space equipped with an inner product that induces a norm that induces a metric so that all Cauchy Sequences converge to a limit under this metric.*

1.3 Existence and Uniqueness of Projections onto a Convex Set

Intuition: You have a convex set and a point outside of this set. Does there always exist a projection of that point to the set, and if so, is that projection unique?

Let X be a Hilbet Space: a complete vector space with metric induced by an inner product: $\|x\|^2 = \langle x\ ,\ x\rangle$. Let $C \subseteq X$ be a closed convex set.

Definition 10 (Projection to a Convex Set)**.** *The projection of a point x onto C is defined as $\pi_c(x) := \arg\min_{y \in C} \|x - y\|^2 = \arg\min_{y \in C} \langle x - y \ , \ x - y \rangle$.*

Theorem 1. *Such a projection exists and is unique: $\exists y \in C$ so that $\|x - y\|^2$ is minimized.*

Proof. Sketch: We shift our space so that $x = 0$. Then the point $y \in C$ with lowest norm satisfies the definition of the Projection to a Convex Set. However, does such a point exist and is it unique?

The set $\{||y|| \mid y \in C\}$ is lower-bounded by 0, so there exists a subsequence $\{||y_n||\}_{n=1}^{\infty}$ converging to the infimum of this set. We use this property, along with the parallelogram identity, to prove our $\{y_n\}_{n=1}^{\infty}$ is Cauchy. As we assume a Hilbert space, Cauchy sequences converge to points within C. So $\exists$ a point with lowest norm in C, thus proving the projection of x onto C exists and is unique.

If $x \in C$, then $\pi_c(x) = x$, so clearly, this value exists and is unique.

Now suppose $x \notin C$. WLOG, we recenter the Hilbert space so $x = 0$ by applying the translation $-x$ to all points. Let $M = \inf_{y \in C} \|y\|$ and further, $\exists$ a sequence $\{y_n\}_{n=1}^{\infty}$ so that $\{\|y_n\|\}_{n=1}^{\infty} \to M$ (i.e. the norm of a sequence of points in C converges to the infimum).

Then, for any indices $n, m \in \mathbb{N}$ in this sequence converging to the infimum:

$$\begin{aligned}
2\|y_n\|^2 + 2\|y_m\|^2 &= \|y_n - y_m\|^2 + \|y_n + y_m\|^2 \\
&= \|y_n - y_m\|^2 + \langle y_n + y_m \ , \ y_n + y_m \rangle \\
&= \|y_n - y_m\|^2 + 2\langle \frac{y_n + y_m}{2} \ , \ y_n + y_m \rangle \\
&= \|y_n - y_m\|^2 + 4\langle \frac{y_n + y_m}{2} \ , \ \frac{y_n + y_m}{2} \rangle \\
&= \|y_n - y_m\|^2 + 4\|\frac{y_n + y_m}{2}\|^2
\end{aligned}$$

Definition 11 (Parallelogram Identity)**.**

$$\begin{aligned}
\|y - z\|^2 + \|y + z\|^2 &= \langle y - z \ , \ y - z \rangle + \langle y + z \ , \ y + z \rangle \\
&= \langle y \ , \ y - z \rangle - \langle z \ , \ y - z \rangle + \langle y \ , \ y + z \rangle + \langle z \ , \ y + z \rangle \\
&= \langle y - z \ , \ y \rangle - y - zz + \langle y + z \ , \ y \rangle + \langle y + z \ , \ z \rangle \\
&= \langle y \ , \ y \rangle - 2\langle z \ , \ y \rangle + \langle z \ , \ z \rangle + \langle y \ , \ y \rangle + 2\langle y \ , \ z \rangle + \langle z \ , \ z \rangle \\
&= 2\langle y \ , \ y \rangle + 2\langle z \ , \ z \rangle \\
&= 2\|y\|^2 + 2\|z\|^2
\end{aligned}$$

□

As C is a convex set, $\frac{y_n+y_m}{2} \in C$, so $\|\frac{y_n+y_m}{2}\| \geq M$ and as $n, m \to \infty$ (i.e. as we go sufficiently far in the sequence $\{y_n\}_{n=1}^{\infty}$):

$$\lim_{n,m\to\infty} 2\|y_n\|^2 + 2\|y_m\|^2 = \lim_{n,m\to\infty} \|y_n - y_m\|^2 + 4\|\frac{y_n + y_m}{2}\|^2$$
$$2M^2 + 2M^2 \leq^* \lim_{n,m\to\infty} \|y_n - y_m\|^2 + 4M^2$$
$$0 \leq \lim_{n,m\to\infty} \|y_n - y_m\|^2$$
$$0 =^{**} \lim_{n,m\to\infty} \|y_n - y_m\|^2$$

$*$: We use $4M^2 \leq 4 \lim_{n,m\to\infty} \|\frac{y_n+y_m}{2}\|^2$, so the RHS is now less than or equal to the LHS.
$**$: The norm is non-negative, so it must be equal to 0.

As $\lim_{n,m\to 0} d(y_n, y_m) = 0$, with the distance metric induced by the norm, $\{y_n\}_{n=1}^{\infty}$ is Cauchy. As we reside in a Hilbert space, Cauchy sequences converge **to points within our set**, so $\lim_{n\to 0} \{y_n\}_{n=1}^{\infty} = \pi_c(x) \in C$. Moreover, limits are unique, so $\pi_c(x)$ is unique.

Returning to our high-level picture; we used the fact that $\inf_{y\in C} \|y\| = M$ to construct a sequence converging to M: $\{\|y_n\|\}_{n=1}^{\infty} \to M$. With some basic manipulation, we get that $0 = \lim_{n,m\to\infty} \|y_n - y_m\|^2$, so this sequence is actually Cauchy. As we reside in a Hilbert space, Cauchy sequences converge to elements in C, so $\exists$ a unique point with lowest norm in our convex set C.

Theorem 2. *$\pi_c(x)$ is completely characterized by $\langle x - \pi_c(x)\ ,\ y - \pi_c(x)\rangle \leq 0$ for all $y \in C$.*

[Aside]: a property characterizes an object if having this property is equivalent to an object satisfying its definition. For example, an equilateral triangle is defined by having three equal sides, but it can be completely characterized by two 60° angles.

Proof. We just proved that $\pi_c(x)$ is unique, but now suppose $\pi \in C$ satisfies $\langle x - \pi_c(x)\ ,\ y - \pi_c(x)\rangle \leq 0$ for all $y \in C$. (i.e. the angle is oblique). Then

$$\langle x - \pi\ ,\ y - \pi\rangle^{(*)} = \frac{1}{2}\|x - \pi\|^2 + \frac{1}{2}\|y - \pi\|^2 - \frac{1}{2}\|x - y\|^2 \leq 0$$

$(*)$: This is a polarization identity for a vector space over $\mathbb{R}$.

$$\frac{1}{2}\|x - \pi\|^2 + \frac{1}{2}\|y - \pi\|^2 \leq \frac{1}{2}\|x - y\|^2$$

$\forall y \in C$. As $\frac{1}{2}\|x - y\|^2 \geq \frac{1}{2}\|x - \pi\|^2 + \frac{1}{2}\|y - \pi\|^2$, and norms are non-zero, this implies $\|x - y\| \geq \|x - \pi\|$. Therefore, π is as close to x as any $y \in C$, so it satisfies the definition of the orthogonal projection.

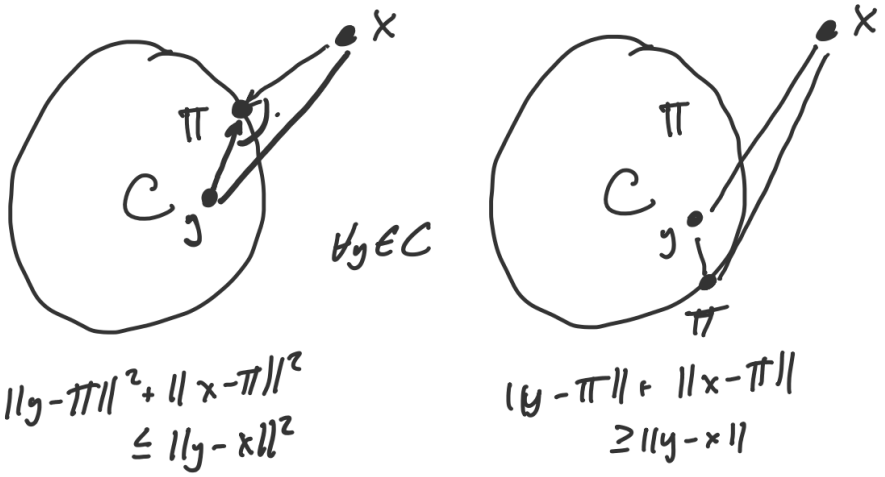

Figure 2: Left: The projection is uniquely characterized by an oblique angle of $\langle x - \pi_c(x)\ ,\ y - \pi_c(x)\rangle \leq 0\ \forall y \in C$. Right: a non-oblique angle implies there's point in C closer to x.

Now we'll show the converse: Let $\pi = \pi_c(x)$ be the projection. We want to show $\langle x - \pi\ ,\ y - \pi\rangle < 0$, or equivalently, there is an obtuse angle between $(x - \pi)$ and $(y - \pi)\ \forall y \in C$.

Let $y_1 \in C$ and $y_\lambda = \pi + \lambda(y_1 - \pi)$ for $\lambda \in [0, 1]$ so y_λ is the line segment between y_1 and π. Then

$$\begin{aligned}
\frac{1}{2}\|x - \pi\|^2 &\leq \frac{1}{2}\|x - y_\lambda\|^2 \\
&= \frac{1}{2}\|x - \pi + \lambda(y_1 - \pi)\|^2 \\
\text{[Polarization ID]}\|x - u\|^2 &= \|x - w\|^2 + 2\langle x - w\ ,\ w - u\rangle + \|w - u\|^2 \\
&= \frac{1}{2}\|x - \pi\|^2 - \lambda\langle x - \pi\ ,\ y_1 - \pi\rangle + \frac{\lambda^2}{2}\|y_1 - \pi\|^2 \\
0 &\leq -\lambda\langle x - \pi\ ,\ y_1 - \pi\rangle + \frac{\lambda^2}{2}\|y_1 - \pi\|^2 \\
\lambda\langle x - \pi\ ,\ y_1 - \pi\rangle &\leq \frac{\lambda^2}{2}\|y_1 - \pi\|^2
\end{aligned}$$

And for $\lambda \in (0, 1]$,

$$\begin{aligned}
\langle x - \pi\ ,\ y_1 - \pi\rangle &\leq \frac{\lambda}{2}\|y_1 - \pi\|^2 \\
&\leq \lim_{\lambda \to 0} \frac{\lambda}{2}\|y_1 - \pi\|^2 \\
&\leq 0
\end{aligned}$$

And when $\lambda = 0$, $y_\lambda = \pi$, so $\langle x - \pi \ , \ 0 \rangle = 0$. $\square$

We can now go on to develop a notion of separation between convex sets in addition to separating and supporting hyperplanes.

1.4 Separating and Supporting Hyperplanes

Corollary 1. *For a closed, convex set C, and any point $x \notin C$, $\exists$ some $\vec{v} \neq \vec{0}$ such that $\langle v \ , \ x \rangle > \sup_{y \in C} \langle v \ , \ y \rangle$. That is, there is strict separation between that point and the convex set C.*

Proof. Idea: Simply look at the projection of $\pi(x)$ onto C, and that gives us our separating hyperplane.

Let $v = x - \pi(x)$ which is non-zero because $x \notin C$. Then for any $y \in C$,

$$\begin{aligned} \langle x - \pi(x) \ , \ y - \pi(x) \rangle &\leq 0 \\ \langle v \ , \ y - \pi + x - x \rangle &\leq 0 \\ \langle v \ , \ y \rangle - \langle v \ , \ x \rangle + \langle v \ , \ x - \pi \rangle &\leq 0 \\ \langle v \ , \ y \rangle + \|v\|^2 &\leq \langle v \ , \ x \rangle \end{aligned}$$

As $\|v\|^2$ is non-zero, this gives us strict separation between C and x. Intuitively, $\|v\|^2$ is the margin between the point x and the closest point $\pi(x)$ in C. $\square$

Corollary 2. *Let C_1 be a closed, convex set and C_2 be a compact, convex set with $C_1 \cap C_2 = \emptyset$. Then $\exists$ $\vec{v} \neq \vec{0}$ such that $\inf_{x \in C_1} \langle v \ , \ x \rangle > \sup_{x \in C_2} \langle v \ , \ x \rangle$.*

Proof. Consider $C_1 - C_2 = \{y - x | y \in C_1, x \in C_2\}$. $C_1 - C_2$ is convex because it is the combination of two convex sets. As C_2 is compact, it is also closed [this can be showed with sequential compactness, i.e., every sequence has a convergent subsequence]. We also have that $0 \notin C_1 - C_2$ because $C_1 \cap C_2 = 0$. From our Corollary showing separation between a convex set and a point, there is separation between $C_1 - C_2$ and 0:

$\exists v \neq 0$ such that

$$\begin{aligned} \langle v \ , \ 0 \rangle = 0 > & \sup_{y \in C_1, x \in C_2} \langle v \ , \ y - x \rangle \\ &= \sup_{y \in C_1, x \in C_2} \langle v \ , \ y \rangle - \langle v \ , \ x \rangle \\ &= \sup_{y \in C_1} \langle v \ , \ y \rangle - \inf_{x \in C_2} \langle v \ , \ x \rangle \\ \inf_{x \in C_2} \langle v \ , \ x \rangle > & \sup_{y \in C_1} \langle v \ , \ y \rangle \end{aligned}$$

Where WLOG we suppose that $C_1 - C_2$ is maximized when we take the minimum element of C_2 and the maximum element from C_1. $\square$

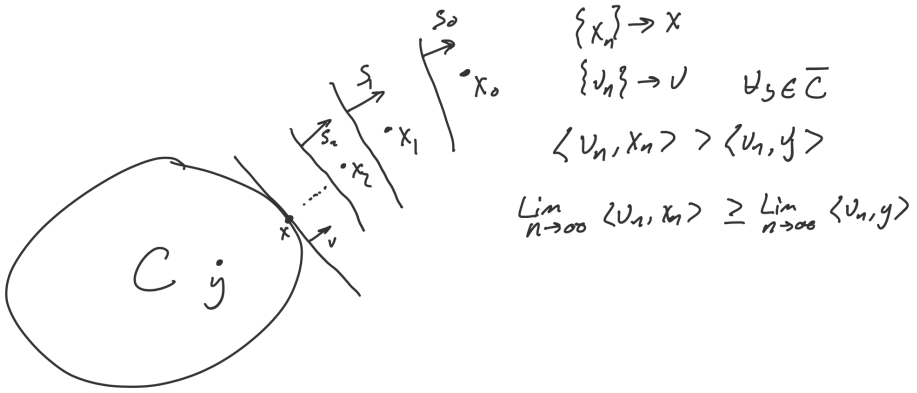

Figure 3: "Proof by picture" of the existence of supporting hyperplanes.

[Aside] Compactness on the second set guarantees that one set cannot approach the other arbitrarily closely: the difference between C_1 and C_2 cannot become arbitrarily small and converge to 0. This allows us to state that $0 \notin C_1 - C_2$.

Specifically, we get into trouble when both sets converge to each other. In Corollary 1, we have separation between a point and a convex set, because while the convex set may become arbitrarily close to the point, the point does not become arbitrarily close to the convex set. However, if both sets converge to the same point v, then $\sup_{x \in C_2} \langle v \ , \ x \rangle = \inf_{x \in C_1} \langle v \ , \ x \rangle$.

We now show that supporting hyperplanes exist. Supporting hyperplanes touch the boundary of the set; while separating hyperplanes are strictly between the set and a point. For simplicity, assume $X = \mathbb{R}^n$ is a finite dimensional space. This is not necessary with the Banach-Alaoglu; however, it simplifies the proof.

Theorem 3 (Supporting Hyperplane Theorem)**.** *If C is convex and $x \in \partial C$ (i.e. the boundary of C), then $\exists v \neq 0$ that supports X: $\langle v \ , \ x \rangle \geq \langle v \ , \ y \rangle \ \forall y \in C$.*

Recall that $\partial C = \overline{C} - int(C)$: the closure of a set minus the interior of the set.

Intuitively, we have a convex set C, and a point x on the boundary of C, then you can draw a hyperplane through x with v tangent to the set.

Proof. Let $\{x_n\}_{n=1}^{\infty} \to x$ such that $\forall n \in N$, $x_n \notin \overline{C}$. By Corollary 1, $\exists$ strict separation between x_n and $\overline{C}$ where s_n is a separating hyperplane: $\langle s_n \ , \ x_n \rangle > \sup_{y \in C} \langle s_n \ , \ y \rangle$.

Let $v_n = \frac{s_n}{\|s_n\|}$ be the normalized vector for this separating hyperplane. Then $\|v_n\| = 1$, and the sequence $\{v_n\}_{n=1}^{\infty}$ has a converging subsequence as each v_n lives on the unit hypersphere, and this set is compact in $\mathbb{R}^n$. This is where we use finite dimensionality: the fact that a norm ball is compact only holds in finite dimensions.

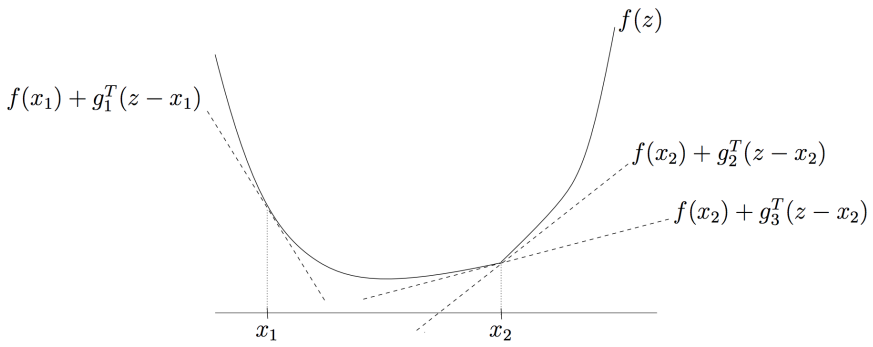

Figure 4: Three different subgradients: one at point x_1 and the other two at x_2.

WLOG considering the sequence, there is a subsequence $\{v_n\}_{n=1}^{\infty} \to v$ with $\|v\| = 1$. Recall that for any $y \in C$,

$$\begin{aligned} \langle v_n \ , \ x_n \rangle &> \langle v_n \ , \ y \rangle \\ \lim_{n\to\infty} \langle v_n \ , \ x_n \rangle &\geq \lim_{n\to\infty} \langle v_n \ , \ y \rangle \\ \langle v \ , \ x \rangle &\geq \langle v \ , \ y \rangle \end{aligned}$$

since as $n \to \infty$, $v_n \to v$, $x_n \to x$, so $\langle v_n \ , \ y \rangle \to \langle v \ , \ y \rangle$. □

For infinite dimensions, we need Banach-Alaoglu to state the norm ball $B = \{x | \|x\| \leq 1\}$ is compact in the weak-$*$ topology. So inner products themselves have convergent subsequences.

2 January 17

2.1 Closed Convex Functions and Subdifferentials

Definition 12 (Epigraph of a function)**.** *The epigraph (epi is latin for above) of a function is the set of points lying above the graph of a function:*

$$\{(x, t) | f(x) \leq t\}$$

Lemma 1. *A function is convex* $\iff$ *the epigraph is a convex set.*

Definition 13 (Closed Convex Function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is closed if its epigraph is a closed set.*

Definition 14 (Subgradient)**.** *A vector $s \in \mathbb{R}^n$ is a subgradient of f at a point x if $f(y) \geq f(x) + s^T(y - x)$. We will denote this property as $(*)$*

Definition 15 (Subdifferential)**.** *The subdifferential is the closed convex set of all subgradients of the convex function f:*

$$\partial f(x) = \{s | f(y) \geq f(x) + s^T(y - x) \ \forall y \in C\}$$

Theorem 4. *If f is a convex function and $x \in dom(f)^\circ$ [recall $^\circ$ denotes the interior of a set], then $\partial f(x) \neq \emptyset$.*

Proof. By the existence of supporting hyperplanes, $\exists v = \begin{bmatrix} a & b \end{bmatrix} \neq 0 \in \mathbb{R}^{n+1}$ at point $z = \begin{bmatrix} x & f(x) \end{bmatrix}$ such that $\forall z' \in epi(f)$ $v^T z \geq v^T z'$. We can now set $v = -v$ so that $v^T z \leq v^T z'$. Multiplying this out for $z' = \begin{bmatrix} y & t \end{bmatrix}$,

$$a^T y + bt \geq a^T x + bf(x)$$

and taking the limit as $t \to \infty$ implies that $b \geq 0$. If $b > 0$, then

$$t \geq f(x) + (-\frac{a}{b})^T (y - x)$$

and since $t \geq f(y)$ $\forall (y, t) \in epi(f)$, $-\frac{a}{b}$ is the subgradient at x.

Now suppose $b = 0$. Then $\exists \epsilon > 0$ such that $x + u\epsilon \in dom(f)$ for $\|u\| = 1$ as $x \in dom(f)^\circ$. Intuitively, we can "wiggle" x to the left and right without escaping the domain as x is in the interior of $dom(f)$. Then by using the opposite sign of the supporting hyperplane $a = -a$:

$$a^T (x + u\epsilon) \geq a^T x$$

and we get that $\epsilon a^T u \geq 0$. However, we can set $u = \frac{-a}{\|a\|}$ so that $a^T u < 0$. Therefore, we have a contradiction and $b > 0$. □

In summary, we have a non-empty subdifferential at all points in the interior of $dom(f)$. "Don't go near the boundary, things get weird."

2.2 Outer Representations of a Convex Function

Theorem 5. *If f is a closed, convex function, then*

$$f(x) = \sup_h \{h(x)\}$$

where h is affine and $h(x) \leq f(x)$ $\forall x \in dom(f)$. Simply, we can write f as the supremum of all affine global under estimators of f by taking the largest estimator at each $x \in dom(f)$ (point-wise supremum).

Proof. This proof gets a bit messy because we need "close to" vertical hyperplanes to match f if it goes vertical and vertical lines are not functions.

We know epi$f = \cap_h \{\text{epi}f \subset h\}$ where h is a half-space, so the epigraph can be written as an intersection of half-spaces. We will show that we can take the supremum of these affine functions (i.e. the half-spaces) to get the function itself.

We denote the halfspace $H_{abc} = \{(x,t)|a^T x + bt \geq c\}$. H_{abc} contains epi$f$, and by taking $t \to \infty$ (recall the epigraph of f is a set that contains all points above $f(x)$ and epi$f \subset H_{abc}$), it must be the case that $b \geq 0$.

$$a^T x + bt \geq c$$
$$\text{If } b > 0, \text{ then..} \ \frac{a}{b}^T x + t \geq \frac{c}{b}$$
$$\text{And if } b = 0, \text{ then...} \ a^T x \geq c$$

We reduce to the case that $b \in \{0, 1\}$: either a vertical hyperplane $(a, 0)$ or a non-vertical hyperplane $(a, 1)$. Define $H_b = \{(a, c)|H_{abc} \supset \text{epi}f\}$ for $b \in \{0, 1\}$. We want to show:

$$\bigcap_{H_1} H_{a1c} = (\bigcap_{H_0} H_{a0c}) \cap (\bigcap_{H_1} H_{a1c})$$

so we can ignore all vertical hyperplanes: we only want to deal with hyperplanes that have $(a, 1)$ as opposed to the ones that have $(a, 0)$. Showing the above equality allows us to only consider the non-vertical ones.

Fix $(a_0, c_0) \in H_0$ and $(a_1, c_1) \in H_1$. THen define $v_0 = (a_0, 0, c_0)$ and $v_1 = (a_1, 1, c_1)$. Interpolating between v_1 and v_0, we have:

$$v(t) = (a_1 t a_0, 1c_1 + tc_0)$$

Define the interpolated halfspace as:

$$H(t) = \{(x, r)|(a_1 + ta_0)^T x + r \geq c_1 + tc_0\}$$

as one of the halfspaces interpolating between the non-vertical and vertical halfspaces. We want to show that this interpolation lower-bounds epi$f$ without "going full vertical". Each interpolated halfspace must contain epi$f$ since both H_0 and H_1 half-spaces contain epi$f$, so the interpolation of half spaces between them also contain epi$f$.

Note $H(t)$ is of the form $H_{a,1,c}$ for $a = a_1 + ta_0$ and $c = c_1 + tc_0$, so in particular, $H(t) \in H_1$. Taking $t \to \infty$

$$a_0^T x \geq c_0 \iff ta_0^T x \geq tc_0$$

for all $t \geq 0$. Then $(x, r) \in H(t) \ \forall t \geq 0$. But notice the inequality $a_0^T x \geq c_0$ is the constraint for $(x, r) \in H_{a0c}$. Since $(x, r) \in H(t) \ \forall t > 0$, the intersection of all $H(t)$ will also have this point.

We then have:

$$H_{a_0,0,c_0} \subseteq \bigcap_{t \geq 0} H(t)$$

for an arbitrary point (x, r). Moreover, as $H(t) \in H_1$, we have shown

$$\bigcap_{H_1} H_{a1c} = (\bigcap_{H_0} H_{a0c}) \cap (\bigcap_{H_1} H_{a1c})$$

The conjugate function

the **conjugate** of a function f is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x))$$

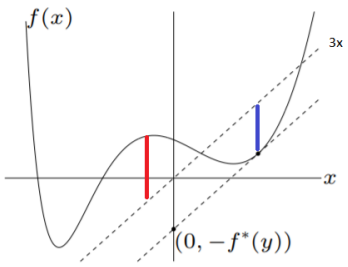

Figure 5: Example of the convex conjugate of a function: we can plot the graph $(x, -f^*(y))\ \forall y$. For example, $f^*(-3)$ will give $\sup_x\{3x - f(x)\}$ which has supremum $3 - f'(x) = 0 \Rightarrow f'(x) = 3$.

Intuitively, we can interpolate from non-vertical half-spaces to vertical half-spaces without going "full vertical" so that an arbitrary point covered by the set of vertical half-spaces is contained in the interpolation. □

2.3 Fenchel/Convex Conjugate

Definition 16 (Fenchel/Convex Conjugate)**.** *The convex (or Fenchel) conjugate is*

$$f^*(s) := \sup_x\{s^T x - f(x)\}$$

is the supremum of the difference between $s^T x$ and $f(x)\ \forall x$ in $dom(f)$.

Note: since the convex conjugate is the supremum of affine functions in s (for a fixed x), it is always closed and convex.

Definition 17 (Double Convex Conjugate)**.** *Define the double convex conjugate as:*

$$f^{**}(x) = \sup_x\{h(x) | h \leq f\}$$

and h is affine.

The bi-conjugate (or double conjugate) looks at all the affine functions that underestimate f and then take the point-wise maximum (supremum).

Corollary 3. *If f is a closed convex function, then $f^{**} = f$.*

Proof. Say h is affine and $h \leq f$. Then $h(x) = s^T x + b$ and $s^T x + b \leq f(x)$ so

$$b \leq f(x) - s^T x$$

by taking the infimum over all x of $f(x) - s^T x$

$$\inf_x b \leq \inf_x \{f(x) - s^T x\}$$

$$b \leq -f^*(s)$$

returning to our original equation:

$$h(x) = s^T x + b$$

$$h(x) \leq s^T x - f^*(s)$$

$$h(x) \leq s^T x - f^*(s) \leq s^T x - (s^T x - f(x)) = f(x)$$

So the supremum of affine functions recover the function f at each $x \in dom(f)$. Since the bi-conjugate of a function is the supremum of all affine functions below it, we've shown $f^{**}(x) = f(x)\ \forall x \in dom(f)$. □

Corollary 4 (Fenchel-Young Inequality)**.** *Suppose f is closed convex. For any s, x such that*

$$s^T x \leq f(x) + f^*(s)$$

Equality holds:

$$s^T x = f(x) + f^*(s)$$

if and only if s is a subdifferential of f at x: $s \in \partial f(x)$.

Finally, we have:

$$s \in \partial f(x) \iff x \in \partial f^*(s)$$

i.e., $(\partial f)^{-1} = \partial f^*$

Proof.

$$s^T x = f(x) + s^T x - f(x)$$

$$s^T x - f(x) \leq f^*(s)$$

$$(\Rightarrow) s^T x \leq f(x) + f^*(s)$$

if equality holds, then

$$\begin{aligned} s^T x &= f(x) + f^*(s) \\ &= f(x) + \sup_y \{y^T s - f(y) \\ &\geq f(x) + s^T y - f(y) \end{aligned}$$

for any fixed y. This inequality holds $\iff$ $f(y) \geq f(x) + s^T(y-x)$, which is exactly the definition of the subgradient. Therefore $s \in \partial f(x)$.

Conversely, if $s \in \partial f(x)$, then $\forall y$,

$$f(y) \geq f(x) + s^T(y-x)$$
$$s^T x - f(x) \geq s^T y - f(y)$$

$\forall y$. We can then take the supremum over all y:

$$\sup_y s^T x - f(x) = \sup_y \{s^T y - f(y)\}$$
$$s^T x - f(x) = f^*(s)$$

The second $\iff$ in the proof is immediate by swapping f with f^*. $\square$

3 January 24

3.1 Convex Optimization Problems

Definition 18 (Convex Optimization Problem)**.** *A convex optimization problem minimizes a closed, convex function $f(x)$ such that $x \in C$ where C is a closed convex set.*

If you can formulate a problem as a convex optimization problem, morally it's solved.

Definition 19 (Discrete Optimization Problem)**.** *A discrete optimization problem occurs when x is discrete rather than continuous:*

Minimize $f(x)$ such that $x \in \{\pm 1\}^n$.

Discrete optimization problems often manifest as combinatorial problems where x is discrete rather than continuous. Clearly, this problem is not convex because the domain is not continuous.

Example 1 (Goemans-Williamson Maximum Cut Problem)**.** *You're given a weighted graph: $w_{i,j}$ for each edge $e_{i,j}$ of the graph. The goal is to assign values $x_i \in \{\pm 1\}$ to each node i so that you maximize the cut value $\frac{1}{2}\sum_{i \leq j} w_{i,j}(1 - x_i x_j)$. Intuitively, nodes with different signs should be connected.*

3.2 Relaxations of Hard Problems

General idea: relax the problem into a convex problem, solve the convex problem, and "hope" you can convert it back to a good solution in the original

problem.

Observe: $x_j \in \{\pm 1\} \iff x_j^2 = 1$. Note that if we define $X = xx^T$, then X is a positive semi-definite (psd) matrix ($X \geq 0$) with 1 along the diagonal ($diag(X) = 1$). Note that $\begin{bmatrix}1\\x\end{bmatrix}\begin{bmatrix}1\\x\end{bmatrix}^T = \begin{bmatrix}1 & x^T\\x & xx^T\end{bmatrix} = M$ and $diag(X) = 1$, $rank(M) = 1$.

We can now formulate an equivalent problem: minimize $f(x)$ such that the constraints

1. $\begin{bmatrix}1 & x^T\\x & X\end{bmatrix} \geq 0$
2. diag(X) = 1
3. $rank(\begin{bmatrix}1 & x^T\\x & xx^T\end{bmatrix}) = 1$

are satisfied. This formulation is still non-convex; however if we drop the rank constraint, then the constraints become convex: $\begin{bmatrix}1 & x^T\\x & xx^T\end{bmatrix} \geq 0$ and $diag(X) = 1$ is a convex (semidefinite) constraint.

Claim:

$$\begin{bmatrix}1 & x^T\\x & xx^T\end{bmatrix} \geq 0$$

is equivalent to

$$X \geq xx^T$$

Proof.

$$\begin{aligned}\inf_{a\in\mathbb{R}} \begin{bmatrix}a\\b\end{bmatrix}^T \begin{bmatrix}1 & x^T\\x & xx^T\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix} &= \inf_{a\in\mathbb{R}}\{a^2 + 2ab^Tx + b^TXb\}\\ &= b^TXb - (b^Tx)^2\\ &= b^T(x - xx^T)b\end{aligned}$$

and

$$\inf_b b^T(x - xx^T)b = -\infty$$

if $X \geq xx^T$ but 0 otherwise. □

A final remark: if $\hat{x} = \{\pm 1\}$, then the optimal value of relaxation certifies gap.

3.3 Max Cut Problems and Randomized Rounding

We can relax the max cut problem to:
maximize $\frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij})$
such that $X \geq 0$ and $diag(X) = 1$.

This is equivalent to the problem of max $tr(WX)$ such that $X \geq 0$ and $diag(X) = 1$ where $W = [w_{ij}]$ (matrix of w_{ij}s) This is called the max cut semi-definite program.

Challenge: how to transform non-exact solution into one that you can actually use in the original problem. In other words, how do you transform the solution to the relaxed problem into a solution for the original discrete problem?

Common Solution: Randomized Rounding. Note $X \geq 0$, then $X = v^T v$ where $v = \begin{bmatrix} v_1 & ... & v_n \end{bmatrix}$ and $X_{ij} = v_i^T v_j$ for vectors $v_1, v_2, ..., v_n$. If $diag(X) = 1$, then $\|v_i\|_2 = 1$ for each i. We're supposed to have that $X_{ij} = x_i x_j$ for some $x \in \{\pm 1\}^n$.

Intuition: think of $v_i^T v_j$ as "correlation" (since this is in between -1 and 1) between values of x_i and x_j. We draw

$$U \sim Uni(S^{n-1})$$

where $S^{n-1} = \{u \in \mathbb{R}^n | \|u\|_2 = 1\}$ is the surface of a sphere with unit norm radius. We sample a single U and then set $x_i = sign(v_i^T U)$ for all x_is.

Note: if $rank(X) = 1$, then you know $v_i = \pm v\}$ for some $v \in \mathbb{R}^n$ (rank 1 matrix must have same columns up to sign + same norm), then $x_i x_j = v_i^T v_j = \pm\|v\|_2^2 = \pm 1$. So randomized rounding will not kill a correct solution to the original problem.

Theorem 6 (Goemans-Williamson)**.** *If $M = \frac{1}{2}\sum_{i<j} w_{ij}(1 - x_i x_j)$ where $x_i x_j$ are the random draws then*

$$\mathbb{E}[M] \geq \alpha M^* \geq \alpha \cdot \textit{[the value optimal discrete solution]}$$

where $\alpha = \inf_{\theta \in [0, 2\pi]} \frac{2}{\pi} \frac{\theta}{1 - \cos\theta} > 0.87856$ and M^* is the value of the relaxed problem. In other words, we're within a factor of α of the optimal discrete solution.

Proof.

$$\frac{1}{2}\mathbb{E}[w_{ij}(1 - x_i x_j)]$$

will be 2 or 0 where $x_i x_j$ are the randomized roundings. Therefore,

$$\frac{1}{2}\mathbb{E}[w_{ij}(1 - x_i x_j)] = w_{ij} P(\langle v_i \ , \ U\rangle \langle v_j \ , \ U\rangle \leq 0)$$

and further by symmetry

$$w_{ij} P(\langle v_i \ , \ U\rangle \langle v_j \ , \ U\rangle \leq 0) = 2w_{ij} P(\langle v_i \ , \ u\rangle \geq 0, \langle v_j \ , \ u\rangle \leq 0)$$

because we can flip the signs and this would then be equal probability – one must be positive and the other must be negative.

Let $v_1^T v_2 = \cos\theta$ be the angle between v_1 and v_2 on the surface of the unit sphere: $\theta = \cos^{-1}(v_1^T v_2)$. Then we have

$$\begin{aligned} P(\langle v_i \ , \ u\rangle \geq 0, \langle v_j \ , \ u\rangle \leq 0) &= \frac{\theta}{2\pi} \\ &= \frac{\cos^{-1}\langle v_i \ , \ v_j\rangle}{2\pi} \end{aligned}$$

Combining with $2w_{ij} P(\langle v_i \ , \ u\rangle \geq 0, \langle v_j \ , \ u\rangle \leq 0)$, we get

$$\mathbb{E}[M] = \frac{1}{\pi} \sum_{i<j} w_{ij} \cos^{-1}\langle v_i \ , \ v_j\rangle$$

Lemma 2. *For $t \in [-1, 1]$, $\frac{\cos^{-1}(t)}{\pi} \geq \frac{\alpha}{2}(1 - t)$*

Proof. Set $t = \cos\theta$. Then $\cos^{-1}(t)/\pi = \frac{\theta}{\pi} = \frac{2\theta}{\pi(1-\cos(\theta))} \cdot \frac{1-\cos\theta}{2}$ where $1 - \cos\theta = 1 - t \geq \alpha \cdot \frac{1-t}{2}$ □

Lemma 3. $\inf_\theta \frac{2\theta}{\pi(1-\cos\theta)} \geq 0.87856$

Proof. By computer or tedious computation. $\cos(\theta) \approx 1 - \theta^2$ for θ near 0. Then $\frac{2\theta}{1-\theta^2} \geq 1$... □

The theorem then follows by observing

$$\begin{aligned} \frac{\cos^{-1}(\langle v_i \ , \ v_j\rangle}{\pi}) &\geq \frac{\alpha}{2}(1 - \langle v_i \ , \ v_j\rangle) \\ &= \frac{\alpha}{2}(1 - x_{ij}) \end{aligned}$$

□

In summary, we can take a hard discrete problem and expand the constraint set so that the constraints are convex to get a "relaxed" problem. We can then solve the relaxed problem and prove it is within a factor of the optimal solution to the original problem.

But we now need a way to associated the solution of the relaxed problem with a solution to the original problem. We need to identify the solution to the relaxed problem with a sphere, random bits, random coin flips, etc., and then prove in expectation that this randomized process yields a solution to the original problem within a factor of the optimal solution in expectation.

4 January 31

4.1 Dual Problems

Define (P) to be the problem: Minimize $f(x)$ subject to $G(x) \leq 0$ and $H(x) = 0$ where

1. $G = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}$ constraints of convex functions.
2. H is affine.

Let w^* be the optimal value of the problem (P). The Lagrangian of (P) is $L(X, \lambda, \nu) = f(x) + \lambda^T G(x) + \nu^T H(X)$.

Definition 20 (Dual Problem)**.** *Any optimization problem has a dual of the form:*

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\ &= \inf_{x \in \mathcal{D}} f(x) + \lambda^T G(x) + \nu^T H(X) \end{aligned}$$

Note that any feasible x guarantees $G(x) \leq 0$, so $\lambda^T G(x) \leq 0$ for feasible solutions. So moreover,

$$w^* \geq g(\lambda, \nu)$$

The dual problem (D) is to maximize $d(\lambda, \nu)$ subject to $\lambda \geq 0$ [element-wise non-negative]. This will get us close to a solution for our optimization problem as $w^* \geq g(\lambda, \nu)$. For today, we will ignore equalities (i.e. $\nu^T H(x)$).

Definition 21 (Strong Duality)**.** *Strong duality means $w^* = g(\lambda, \nu)$ for some choice of λ, ν. Specifically:*

$$w^* = \sup_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

Question: When does strong duality hold?

4.2 Dual Problems

Define the optimal value function

$$w(u) := \inf_x \{f(x) \text{ s.t } G(x) \leq u\}$$

so that $w(0) = w^*$ is the optimal value for (P).

Question: As we perturb u, how does the problem change?
Answer: Compute sub-differentials of the optimal function w.

As u increases, the objective is only going to decrease (more freedom in the value $f(x)$). Notice $w(u)$ is always convex as it is a partial minimization of a convex function. A non-vertical sub-gradient will support everything above it: $\begin{bmatrix}1\\ \lambda\end{bmatrix}$ where 1 is the "λ" multiplier on $f(x)$.

Recall from Lecture 2 that a subdifferential [i.e. all the subgradients at a point x on the domain] is non-empty on the interior of the domain of w. Therefore, we want a way to guarantee that 0 is in the interior of the domain of w. This immediately motivates Slater's condition: $0 \in dom(w)^\circ$, i.e., $\exists x_0$ such that $G(x_0) < 0$, $f(x_0) < \infty$. Simply, this condition guarantees that 0 is not on the boundary of the domain of w. If it is on the boundary, we'd have vertical supporting hyperplanes (failure condition).

Theorem 7. *Let d^* = optimal value of the (D). If (SC) (slater's condition) holds then $w^* = d^*$ and if $w^* > -\infty$, then there exists $\lambda^* \geq 0$ such that $w^* = d(\lambda^*)$.*

Proof. Use the following separating hyperplane theorem: If A, B are convex, and A has interior, and int $A \cap B = \emptyset$ (B contains no points in the interior of A), then $\exists v \neq 0$ such that $\inf_{a \in A} \langle v\ ,\ a\rangle \geq \sup_{b \in B} \langle v\ ,\ b\rangle$: we have a non-trivial separating hyperplane.

Define $A = \{(u,t) | f(x) \leq t, G(x) \leq u, \text{ for some x}\}$ $B = \{(u,t) | u \leq 0, t \leq w^*\}$.

Note B has an interior as long as $w^* > -\infty$. As B° is contained in $\{(u,t) | t < w^*, u \leq 0\}$, the interior of B is completely disjoint from A.

So $B^\circ \neq \emptyset$ and $(B \cap A)^\circ = \emptyset$ [flipped A and B]. Thus exists $\begin{bmatrix}\lambda\\ r\end{bmatrix} \neq 0$ such that

$$\lambda^T u + rt \geq \lambda^T u_0 + r t_0$$

for all $(u,t) \in A$, $(u_0, t_0) \in B$. Need to show this is non-vertical, i.e. $r > 0$.

Take $t \to \infty$ to see that $r \geq 0$ [otherwise, this will fail]. Take $u \to \infty$ (elementwise) to see that $\lambda \geq 0$ (elementwise).

If $r > 0$, we're done because can divide by r:

$$\frac{\lambda}{r}^T u + t \geq \frac{\lambda}{r}^T u_0 + t_0$$

take $u_0 = 0$, $t_0 = w^*$:

$$\begin{aligned}\inf\{t + \frac{\lambda}{r}^T u | f(x) \leq t, G(x) \leq u\} &= \inf_x \{f(x) + \frac{\lambda}{r}^T G(x)\} \\ &= d(\frac{\lambda}{r}) \\ &\geq w^*\end{aligned}$$

So primal = dual.

If r = 0, then our supporting hyperplane is entirely vertical. Say (for contradiction) that $r = 0$. Then we know $u = 0$ is in A, and as $\exists x_0$ such that $G(x_0) < 0$, there's an $\epsilon > 0$ such that $\|u\| \leq \epsilon$ implies there are $(t, u) \in A$ [take an ϵ-neighborhood of 0 and still have this remain in A].

Substitute in $\lambda^T u + rt \geq \lambda^T u_0 + rt_0$ with $r = 0$ and we see that $\lambda^T u \geq \lambda^T u_0$ for all $\|u\| \leq \epsilon$. Take $u_0 = 0$, then $-\epsilon\|\lambda\| \geq 0$. But certainly, this cannot occur as $\lambda \neq 0$ [separating hyperplane is non-zero]. □

4.3 Saddle Points

For a function $L(x, \lambda)$ a pair $(\overline{x}, \overline{\lambda})$ is a saddle point if $\sup_\lambda L(\overline{x}, \lambda) \leq L(\overline{x}, \overline{\lambda}) \leq \inf_x L(x, \overline{\lambda})$. Note if such a pair exists,

$$L^* := \inf_x \sup_\lambda L(x, \lambda)$$

satisfies

$$L^* = L(\overline{x}, \overline{\lambda}) = \sup_\lambda \sup_x L(x, \lambda)$$

[can switch the order of inf and sup]. Why does this hold? We always have: $\sup_\lambda \inf_x L(x, \lambda) \leq \inf_x \sup_\lambda L(x, \lambda)$ [the basic max-min inequality].

But given if $(\overline{x}, \overline{y})$ is a saddle, we have $\inf_x \sup_\lambda L(x, \lambda) \leq \sup_\lambda L(\overline{x}, \lambda) \leq L(\overline{x}, \overline{\lambda}) \leq \inf L(x, \overline{\lambda}) \leq \sup_\lambda \inf_x L(x, y)$ (so everything is an equality). This is sometimes called min-max duality.

Proposition 1. *If $L(x, \lambda) = f(x) + \lambda^T G(x)$ and Slater holds and optimum x^* is attained in (P), then (x^*, λ^*) is a saddle point for the Lagrangian.*

Proof. Exercise. □

4.4 Generalized KKT Conditions

Corollary 5 (Karush-Kuhn-Tucker KKT Conditions)**.** *KKT conditions are the following*

1. $0 \in \partial f(x) + \sum_{i=1}^{n} \lambda_i \partial g_i(x)$
2. $G(x) \leq 0$ *[is feasible]*
3. $\lambda \geq 0$
4. $\langle \lambda \ , \ G(x) \rangle = 0$ *[complementary slackness].*

Corollary 6. (x^*, λ^*) *is primal-dual optimal* $\iff$ *KKT conditions hold for convex problems.*

Proof. Say (x^*, λ^*) is optimal (is a saddle point). Then by the definition of optimality,

$$\begin{aligned} d(\lambda^*) &= f(x^*) \\ &\geq f(x^*) + \langle \lambda^* \ , \ G(x^*) \rangle \\ &\geq d(\lambda^*) \end{aligned}$$

That is,

$$L(x^*, \lambda^*) = \inf_x L(x, \lambda^*)$$

Then we have $0 \in \partial_x [f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*)] = [\partial f(x^*) + \sum_{i=1}^{m} \lambda_i^* \partial g_i(x^*)]$ which is equal to the KKT condition. Note that we cheated a bit here: showing the subdifferential can be put inside is non-trivial.

Now we simply need to check that the complementary slackness holds; however, because of the earlier equality $d(\lambda^*) = f(x^*)$ [definition of optimality]. And further:

$$\begin{aligned} f(x^*) &\geq f(x^*) + \langle \lambda^* \ , \ G(x^*) \rangle \\ &\geq d(\lambda^*) \end{aligned}$$

so we have that complementary slackness holds.

If the pair (x^*, λ^*) satisfies KKT, then $0 \in d_x L(x^*, \lambda^*)$, i.e., $L(x^*, \lambda^*) = \inf_x L(x, \lambda^*) = d(\lambda^*)$ and x^* is feasible, so that

$$f(x^*) = f(x^*) + \langle \lambda^* \ , \ G(x^*) \rangle = \sup_{\lambda \geq 0} L(x^*, \lambda)$$

So we have strong duality and are now done. In summary,

$$\begin{aligned} L(x^*, \lambda^*) &= \inf_x \sup_{\lambda \geq 0} L(x, \lambda) \\ &= sup_{\lambda > 0} \inf_x L(x, \lambda) \\ &= w^* \\ &= d^* \end{aligned}$$

□

5 February 7

5.1 Convex Geometry

Big Picture: deep connections between convex geometry and convex bodies (compact, convex set with an interior) and many areas of mathematics. Many modern information theory tools follow from what we do today. Sampling theory from markov chain convergence results.

Example 2. *What shape do soap bubbles have to take? Minimum energy configuration is a sphere. The thing that minimizes surface area subject to a volume constraint is a ball.*

Theorem 8 (Brunn-Minkowski (BM))**.** *Let A, B be compact [closed and bounded] subsets of $\mathbb{R}^n$. Then*

$$Vol(\lambda A + (1-\lambda)B)^{1/n} \geq \lambda Vol(A)^{\frac{1}{n}} + (1-\lambda)Vol(B)^{1/n}$$

where $A + B = \{a + b | a \in A, b \in B\}$. That is $Vol^{1/n}$ is concave w.r.t. set sums with $\lambda \in [0, 1]$

Remark 2 (Log-Concave Brunn-Minkowski (LBM))**.** *Equivalent formulation is Log-Concave Brunn-Minkowski (will allow us to do a simple integral argument to prove its concave).*

Let V_n = n-dimensional volume. Then

$$V_n(\lambda A + (1-\lambda)B) \geq V_n(A)^{\lambda} V_n(B)^{1-\lambda}$$

for all $A, B \subset \mathbb{R}^n$, $\lambda \in [0, 1]$

Proof. $BM(\Rightarrow)LBM$:
Recall AM-GM inequality $\lambda x + (1-\lambda)y \geq x^{\lambda} y^{1-\lambda}$ for all $x, y \geq 0$, $\lambda \in (0, 1)$ [so we're done.].

$BM(\Leftarrow)LBM$.
Make particular choices of A, B, λ. Set $\tilde{A} = A/V_n(A)^{1/n}$ and $\tilde{B} = B/V_n(B)^{1/n}$ so their volumes are exactly 1: $V_n(\tilde{A}) = V_n(\tilde{B}) = 1$. Choose $\lambda = \frac{V_n(A^{1/n})}{V_n(A)^{1/n} + V_n(B)^{1/n}}$.

Then by LBM, $V_n(\lambda\tilde{A} + (1-\lambda)\tilde{B}) \geq 1$ from

$$\begin{aligned} V(\tilde{A})^{\lambda} V(\tilde{B})^{1-\lambda} &\leq V_n(\lambda\tilde{A} + (1-\lambda)\tilde{B}) \\ 1^{\lambda} 1^{1-\lambda} &\leq V_n(\lambda\tilde{A} + (1-\lambda)\tilde{B}) \\ 1 &\leq \frac{A}{V_n(A)^{1/n} + V_n(B)^{1/n}} + \frac{B}{V_n(A)^{1/n} + V_n(B)^{1/n}} \end{aligned}$$

So $V_n(A+B) \geq ((V_n(A)^{1/n} + V_n(B)^{1/n})^n$ as $V_n(cA) = c^n V_n(A)$.

i.e., $V_n(A+B)^{1/n} \geq V_n(A)^{1/n} + V_n(B)^{1/n}$. So for any $s \geq 0$, $t \geq 0$, $V_n(sA) = s^n V_n(A)$, similarly $V(tB) = t^n V_n(B)$ therefore,

$$V_n(sA+tB)^{1/n} \geq sV_n(A)^{1/n} + tV_n(B)^{1/n}$$

for $s, t \geq 0$. We call the above (SDBM).

□

Theorem 9 (Isoperimetric Inequality)**.** *Among all bodies (any set with interior) $A \subset R^n$ of the same volume, the l_2-ball minimizes the surface area.*

Proof. Surface area is the $n-1$ dimensional volume of the boundary of the set C.

Note ∂ represents the boundary of a set [see Stoke's Thm for why]. Also SA_n is the n-dimensional surface area.

$$\begin{aligned} SA_n(C) &= V_{n-1}(\partial C) \\ &= \lim_{\epsilon \to 0} \frac{V_n(C + \epsilon B_2^n) - V_n(C)}{\epsilon} \end{aligned}$$

where $\mathbb{B}_2^n$ is the unit l2-epsilon ball in n dimensions.

Use (SDBM) to see

$$V_n(C + \epsilon \mathbb{B}_2^n)^{1/n} \geq V_n(C)^{1/n} + \epsilon V_n(\mathbb{B}_2^n)^{1/n}$$

Raising this to the power n, we have:

$$\begin{aligned} V_n(C + \epsilon \mathbb{B}_2^n) &\geq (V_n(C)^{1/n} + \epsilon V_n(\mathbb{B}_2^n)^{1/n})^n \\ &\geq V_n(C) + n\epsilon V_n(C)^{\frac{n-1}{n}} V_n(\mathbb{B}_2^n)^{1/n} \end{aligned}$$

With the final inequality attained by ignoring all higher-order terms in the binomial expansion.

So we have

$$\begin{aligned} V_{n-1}(\partial C) &\geq \lim_{\epsilon \to 0} n \frac{\epsilon V_n(C)^{\frac{n-1}{n}} V_n(\mathbb{B}_2^n)^{1/n}}{\epsilon} \\ &= n V_n(C)^{\frac{n-1}{n}} V_n(\mathbb{B}_2^n)^{1/n} \end{aligned}$$

WLOG assume $V_n(C) = V_n(\mathbb{B}_2^n)$, then ..

$$nV_n(C)^{\frac{n-1}{n}} V_n(\mathbb{B}_2^n)^{1/n} = nV_n(\mathbb{B}_2^n)$$

If we had $C = \mathbb{B}_2^n$, then $C + \epsilon \mathbb{B}_2^n = \mathbb{B}_2^n + \epsilon \mathbb{B}_2^n = (1+\epsilon)\mathbb{B}_2^n$, and

$$\begin{aligned} SA(\mathbb{B}_2^n) &= \lim_{\epsilon \to 0} \frac{((1+\epsilon)^n - 1) V_n(\mathbb{B}_2^n)}{\epsilon} \\ &= n V_n(\mathbb{B}_2^n) \end{aligned}$$

□

5.2 Prékopa-Leindler Inequality

We'll prove the (LBM) as a consequence of a more powerful integration inequality. Let $f = 1_A$, $g = 1_B$ be the (0-1 indicator) function:

$$f(x) = \begin{cases} 1 \text{ if x is in A} \\ 0 \text{ otherwise} \end{cases}$$

Define $m = 1_{\lambda A+(1-\lambda)B}$. Then (LBM) is same as $\int m \geq (\int f)^\lambda (\int g)^{1-\lambda}$ because $(\int f)^\lambda = V(A)$, $(\int g)^{1-\lambda} = V(B)$ and $\int m = V(\lambda A + (1-\lambda)B)$.

Note that $m(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda}$ as if $x \in A$, $y \in B$, then both sides are 1. Otherwise, RHS is 0.

Theorem 10 (Prékopa-Leindler)**.** *For any $f, g, m : \mathbb{R}^n \to \mathbb{R}_+$, if*

$$m(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda}$$

then

$$\int m \geq (\int f)^\lambda (\int g)^{1-\lambda}$$

Proof. Idea: use level sets of f, g to prove for $n = 1$, and then use induction to go to the higher dimension cases by integrating one variable at a time. Induction part is the HW for this week.

1-dimensional Brunn-Minkowski [all we need for the general case]:

$$V_1(\lambda A + (1-\lambda)B) \geq \lambda V_1(A) + (1-\lambda)V_1(B)$$

WLOG let's say A,B are compact (we can approximate any set by compact sets). Shift leftmost endpoint of A and rightmost endpoint of B so $\inf B = 0$ and $\sup A = 0$. I did not change $V_1(A)$ or $V_1(B)$ to do this. However; we made the convex combination of $V_1(\lambda A + (1-\lambda)B)$ only bigger. Then $\lambda A + (1-\lambda)B \supset \{\lambda A\} \cup \{(1-\lambda)B\}$ as λA and $(1-\lambda)B$ overlap only at 0, $V_1(\lambda A + (1-\lambda)B) \geq \lambda V_1(A) + (1-\lambda)V_1(B)$. This gives use 1-d Brunn-Minkowski.

Assume WLOG $\sup f = \sup g = 1$ [so we have bounded functions]. Then

$$\begin{aligned} \int_{\mathbb{R}} f(x)dx &= \int \int_0^1 1(t \leq f(x))dtdx \\ &= \int_0^1 \int 1(t \leq f(x))dxdt \\ &= \int_0^1 V_1(\{x | f(x) \geq t\}dt \\ &= \int_0^1 V_1(f \geq t)dt \end{aligned}$$

because $\int 1(t \leq f(x))dx = V_1(\{x|f(x) \geq t\})$

If $f(x) \geq t$, $g(y) \geq t$, then $m(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda} \geq t$. So

$$\begin{aligned}\{Z|m(z) \geq t\} &= (m \geq t) \\ &= \lambda(f \geq t) + (1-\lambda)(g \geq t)\end{aligned}$$

All sets in $\mathbb{R}$. Then

$$V_1(m \geq t) \geq \lambda V_1(f \geq t) + (1-\lambda)V_1(g \geq t)$$

Then we observe that

$$\begin{aligned}\int m(x)dx &= \int_0^1 V_1(m \geq t)dt \\ &\geq \lambda \int_0^1 (f \geq t)dt + (1-\lambda)\int_0^1 V_1(g \geq t)dt \\ &= \lambda \int_{\mathbb{R}} f(x)dx + (1-\lambda)\int_{\mathbb{R}} g(x)dx \\ &\geq (\int f)^\lambda (\int g)^{1-\lambda}\end{aligned}$$

[final inequality do to GM-IM].

This is the n=1 dimensional version. The arbitrary n dimensional is induction: assume 1-dimensional case, and then integrate across the last coordinate. Will do this by iterating integrals. □

6 February 14

6.1 Centers of Gravity of Convex Bodies

Definition 22. *$K \subset \mathbb{R}^n$ is a convex body if it is compact, convex, and has non-empty interior.*

Side note: The letter K typically represents compact and convex as the German word for convex is "konvex" and the German word for compact is "kompakt".

Definition 23. *The center of gravity of a set $K \subset \mathbb{R}^n$ is*

$$cg(k) := \frac{1}{V_n(K)} \int_K x dx$$

with volume as defined the previous week: $V_n(K) = Vol_n(K) = \int_K dx$

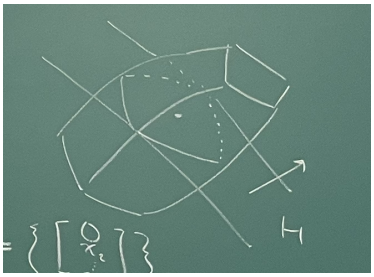

Figure 6: Artistic masterpiece by Dr. Duchi depicting the shape of a slice of a compact, convex body K.

Theorem 11 (Grünbaum's Theorem)**.** *Let $K \subset \mathbb{R}^n$ be a convex body, H a hyperplane containing $cg(K)$, and H^- represent one halfspace of H. Then*

$$\frac{1}{e} \leq (\frac{n}{n+1})^n \leq \frac{V_n(K \cap H^-)}{V_n(K)}$$
$$\leq 1 - \frac{n}{n+1}^n \leq 1 - \frac{1}{e}$$

Intuitively, cutting a convex body by putting a hyperplane through a convex body effectively cuts it in half (bounded below by 0.367 and above by 1- 0.367).

Proof. This proof will consist of three main parts. Part 1 is the spherical symmetrization of K: convex bodies are hard to work with; however, approximating a convex body as infinitesimal-width circles makes it more tractable. Part 2 then approximates a spherical convex body wiht a cone. Finally, Part 3 shows the center of gravity for a cone satisfies the inequality (worst case body), so our initial K does as well.

Part 1: Spherical Symmetrization of K. Intuitively, take slices of the convex body and turn these slices into circles.

WLOG assume that H is axis-aligned, $H = \{x | e_1^T x = 0\} = \{ \begin{bmatrix} 0 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$. Also define $H_t = \{x | x_1 = t\}$ as a sequence of half-spaces going from left to right. WLOG let $\inf\{x_1 | x \in K\} = 0$ and $\sup\{x_1 | x \in K\} = 1$ So we have a sequence of half-spaces $\{H_0, ..., H_t, ..., H_1\}$ all moving in the e_1 direction. See Figure 7.

Define

$$r(t) := V_{n-1}(K \cap H_t)^{\frac{1}{n-1}}$$

as the analogue to the radius of my set: what is the square root of the area of my cut. This is the "radius" of the slice $K \cap H_t$.

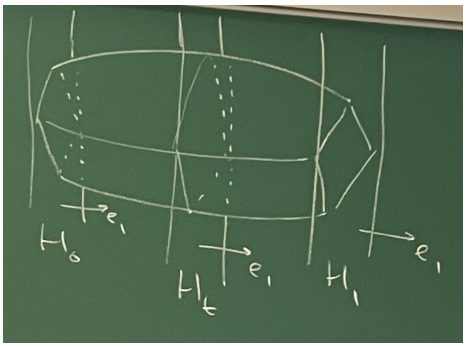

Figure 7: A second artistic masterpiece by Dr. Duchi depicting a sequence of half-spaces aligned in the e_1 direction slicing K.

By convexity, $K \cap H_t \supset (1-t)K \cap H_0 + tK \cap H_1$. Now we're going to use Brunn-Minkowski (says volume functional is concave). By Brunn-Minkowski,

$$\begin{aligned} r(t) &= V_{n-1}(K \cap H_t)^{\frac{1}{n-1}} \\ &\geq tV_{n-1}(K \cap H_1)^{1/(n-1)} + (1-t)V_{n-1}(K \cap H_0)^{\frac{1}{n-1}} \\ &= t \cdot r(1) + (1-t)r(0) \end{aligned}$$

Define the spherical version of these slices:

$$S_t = \{\begin{bmatrix} t \\ x \end{bmatrix} \mid \|x\| \leq r(t), r \in \mathbb{R}^{n-1}\}$$

so that S_t is the spherical version of the slice $K \cap H_t$: S_t has the same volume as H_t; however, they are not circular/spherical.

Lemma 4. *$S := \cup_t S_t$ is a convex set: the union of all spherical slices is a convex set.*

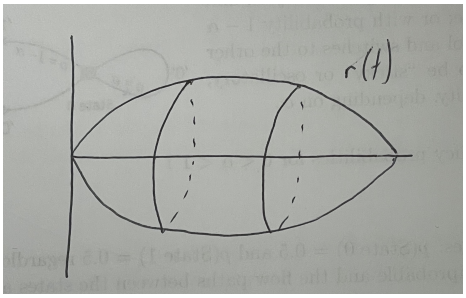

Figure 8: Amateur artist's reconstruction of a third artistic masterpiece by Dr. Duchi depicting the convexity of S.

Proof. Let $\begin{bmatrix} t_0 \\ x_0 \end{bmatrix} \in S_{t_0}$ and $\begin{bmatrix} t_1 \\ x_1 \end{bmatrix} \in S_{t_1}$. Then for any $\lambda \in [0,1]$,

$$\begin{aligned} \|\lambda x_0 + (1-\lambda)x_1\| &\leq \lambda\|x_0\| + (1-\lambda)\|x_1\| \\ &\leq \lambda r(t_0) + (1-\lambda) r(t_1) \\ &\leq r(\lambda t_0 + (1-\lambda)t_1) \end{aligned}$$

by concavity. The point is that if I take any concave function, and then define the a body as the spherical slices of that concave function, then you get a convex body. So $\lambda \begin{bmatrix} t_0 \\ x_0 \end{bmatrix} + (1-\lambda) \begin{bmatrix} t_1 \\ x_1 \end{bmatrix} \in S$ and S is called the spherical symmetrization of the slices. □

Now without loss of generality, let K be a convex body so that $K \cap H_t$ is a ball in $\mathbb{R}^{n-1}$ [assume K is its own spherical symmetrization].

Part 2: Conification. WLOG, move our set so $cg(K) = 0$.

Construct a conic section [cone with its top chopped off] C with

$$C \cap H_0 = C_0 = K \cap H_0$$

"Pull" origin point of C^- until $V_n(C^-) = V_n(K^-)$ where $C^- = C \cap \{x | x_1 \leq 0\}$. $C^+ = C \cap \{x | x_1 \geq 0\}$ Take C^+ to be the "right side" such that $V_n(C^+) = V_n(K^+)$.

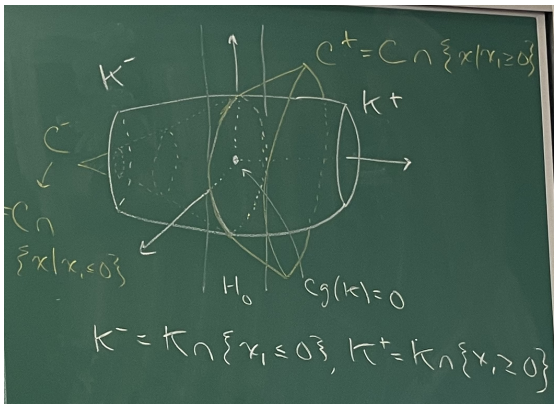

Figure 9: Artistic masterpiece by Dr. Duchi extending a spherical slice C_0 to a cone C so that $V_n(C^-) = V_n(K^-)$ and $V_n(C^+) = V_n(K^+)$. Match $V_n(K^-)$ with $V_n(C^-)$ and then similarly extend the base-side of the cone to match $V_n(K^+)$.

Obviously, $\frac{V_n(C^+)}{V_n(C)} = \frac{V_n(K^+)}{V_n(K)}$. Need to make the argument that the cone's center of gravity is to the left of the center of gravity of the body

Lemma 5. *$cg(C) = \alpha e_1$ for some $\alpha \leq 0$. Simply, the center of gravity of C is shifted to the left relative to the cenger of gravity of K.*

Proof. It suffices to show that both $cg(C^+)$ and $cg(C^-)$ have moved left.

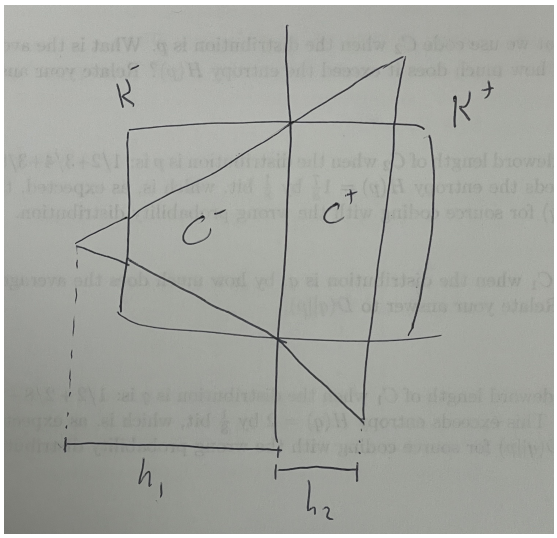

Figure 10: Want to show center of gravity of C is shifted left relative to center of gravity of K.

Define $r(t) = \text{radius}(K \cap H_t) = V_{n-1}(K \cap H_t)^{\frac{1}{n-1}}$. $r(t)$ is concave (Brunn-Minkowski) [on its domain].

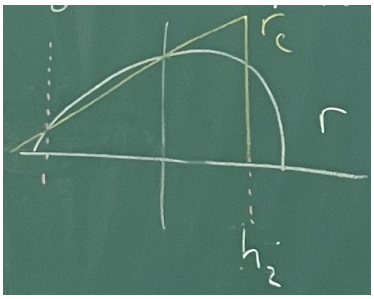

Figure 11: $r(t)$ is a concave function of t and r_C is a linear function of t. See Figure 8 for a similar use of the radius.

Define conic radius $r_C(t) = V_{n-1}(C \cap H_t)^{\frac{1}{n-1}}$ [radius of the cone as we move t] (yellow line Figure 11). So there are two points where r_C intersects with r because r_C is linear and r is concave. Can assume on RHS (i.e. after $t = 0$),

$r_c > r$ until it drops to 0.

By concavity,

$$\begin{cases} r_C(t) \geq r(t) \text{ for } t \in [0, h_2] \\ r_C(t) = 0 \leq r(t) \text{ for } t \geq h_2 \end{cases}$$

So we have

$$\int_{\mathbb{R}_+} r_C^{n-1} = \int_{\mathbb{R}_+} r^{n-1} = V_n(K^+)$$

for $r_C^{n-1} - r^{n-1} \geq 0$ [i.e. from 0 to h_2 where $r_C > r$]. And similarly

$$\int_0^{h_2} r_c^{n-1} - r^{n-1} = \int_{h_2}^{\infty} r^{n-1} - r_c^{n-1}$$

where $r^{n-1} - r_C^{n-1} \geq 0$ [note: this is the first dashed vertical line in Figure 11 – does not include the tip of the cone].

So

$$\int_0^{h_2} \frac{t}{h_2}(r_c^{n-1} - r^{n-1})dt \leq \int_{h_2}^{\infty} \frac{t}{h_2}(r^{n-1} - r_c^{n-1})dt$$

since $\frac{t}{h_2} \leq 1$ from $[0, h_2]$ and $\frac{t}{h_2} \geq 1$ from $[h_2, \infty]$.

As $cg(C^+) = s \cdot e_1$ [assumed C was axis-aligned on e_1], we see that

$$\frac{1}{V_n(C^+)} \int t r_c^{n-1} \leq \frac{1}{V_n(K^+)} \int t r^{n-1}$$
$$\int t r_c^{n-1} \leq \int t r^{n-1}$$
$$[cg(C^+)]_1 \leq [cg(K^+)]_1$$

as $\int t r^{n-1} = [cg(K^+)]_1$ and $\int t r_c^{n-1} = [cg(C^+)]_1$ because $cg(C) = \int_C x dx$. Note: $[\gamma]_1$ denotes the first entry of a k-dimensional vector. For our purposes, the first entry—e_1 dimension—is all that matters because the spherical symmetrization makes all other entries 0 because the convex body is symmetric about the other axes $e_2, ..., e_n$.

To see why $cg(K^+) = [\int tr^{n-1}]_1$, observe that K^+ is symmetric about e_1 (spherical slices on the e_1 axis) and:

$$\begin{aligned} V_n(K^+)cg(K^+) &= \int_{K^+} x dx \\ &= \int_0^\infty te_1 V_{n-1}(H_\cup K^+) dt \\ &= \int_0^\infty te_1 r^{n-1}(t) dt \\ &= \begin{bmatrix} \int_0^\infty te_1 r^{n-1}(t) dt \\ 0_{n-1} \end{bmatrix} \end{aligned}$$

A similar argument works for the LHS (for the cone). Then $cg(C^-) \leq cg(K^-)$ and $cg(C^+) \leq cg(K^+)$, so the whole center of gravity must move to the left: $cg(C) \leq cg(K)$. □

Part 3: Show the inequality $\frac{1}{e} \leq \frac{V_n(K \cap H^-)}{V_n(K)} \leq 1 - \frac{1}{e}$ holds for cones.

Let $B_2^{n-1} = \{x \in \mathbb{R}^{n-1} | \|x\|_2 \leq 1\}$ and call direction of the cone h [l2 ball at the end]. Let $A_n = V_{n-1}(B_2^{n-1})$ Then

$$\begin{aligned} V_n(C) &= \int_0^h V_{n-1}(\frac{t}{h} B_2^{n-1}) dt \\ &= A_n \cdot \int_0^h \frac{t}{h}^{n-1} dt \\ &= A_n \cdot \frac{h}{n} \end{aligned}$$

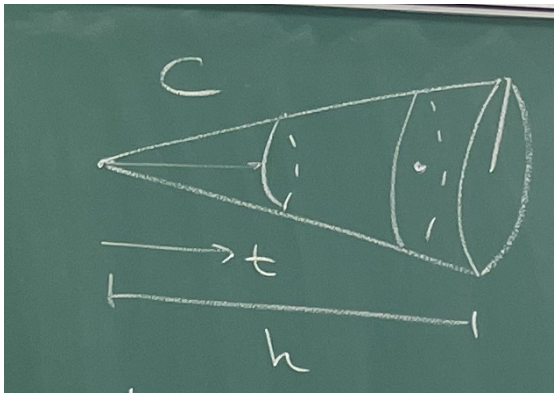

Figure 12: Volume of a cone by integrating hyper-spherical slices (B_2^{n-1}) along h.

The center of gravity will be [ignore everything except first coordinate since symmetric about it]:

$$\begin{aligned}[cg(C)]_1 &= \frac{\int_0^h tV_{n-1}(\frac{t}{h}B_2^{n-1})dt}{V_n(C)} \\ &= \frac{\int_0^h t^n dt}{\int_0^h t^{n-1}dt} \\ &= \frac{n}{n+1}h\end{aligned}$$

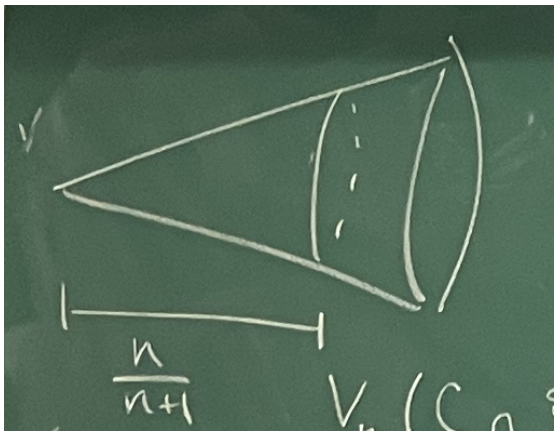

Figure 13: Compute the ratio of the volume of the first $\frac{n}{n+1}$ parts of cone relative to the volume of the entire cone.

Finally, we want to compute the volume ratio $\frac{V_n(K\cap H^-)}{V_n(K)}$ to show it is bounded by $\frac{1}{e}$ and $1-\frac{1}{e}$. Take cone, slice it at point $\frac{n}{n+1}$ along height. Then the volume ratio becomes

$$\begin{aligned}V_n(Cn\{x|x_1 \le \frac{n}{n+1}h\}) &= \int_0^{\frac{n}{n+1}h} V_{n-1}(\frac{t}{h}B_2^{n-1})dt \\ &= A_n \cdot \frac{h}{n} \cdot \frac{n}{n+1}^n \\ &= V_n(C) \cdot (\frac{n}{n+1})^n\end{aligned}$$

We've now proved the volume ratio for cones is the worst case. So for any body, it's bounded by $V_n(C) \cdot (\frac{n}{n+1})^n$.

So $\frac{V_n(K^-)}{V_n(K)} \ge \frac{V_n(C^-)}{V_n(C)} = (\frac{n}{n+1})^n$. Could only have decreased $\frac{V_n(C^-)}{V_n(C)}$ by moving center of gravity to the left. Can do a symmetric argument to get the $1-$ case. □

7 February 21

7.1 Minimizing Quadratics

Motivation:
Trust region methods iteratively approximate f at x_k by

$$\hat{f}(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$$

Then solve

$$x_{k+1} = \arg\min_x \{\hat{f}(x) \text{ s.t. } \|x - x_k\|^2 \leq r_k^2\}$$

where r_k is equal to the trust region radius. This works even when f is non-convex. More generally: we can minimize a quadratic (possibly non-convex) problem:

$$x^T A_1 x 2b_1^T x + c_1$$

subject to

$$x^T A_0 x + 2b_0^T x + c_0 \geq 0$$

We call this problem (Q) for quadratic. Even though $f = x^T A_1 x 2b_1^T x + c_1$ is not necessarily convex, so long as you have a quadratic objective and quadratic inequality constraint, then strong duality holds:

Theorem 12. *If there exists $\bar{x}$ such that $\bar{x}^T A_0 \bar{x} + 2b_0^T \bar{x} + c_0 > 0$ [typical slater-like condition], then strong duality obtains (holds) for (Q).*

Thinking about this some more, let's take the dual of (Q):

$$L(x, \lambda) = x^T(A_1 - \lambda A_0)x + 2(b_1 - \lambda b_0)^T x + c_1 - \lambda c_0$$

is our dual. We then take the inf over all x (so A needs to be positive semi-definite). Note that A^+ denotes the Moore-Penrose (i.e. pseudo) inverse. R is the range of a matrix.

$$\inf_x L(x, \lambda) = \begin{cases} -(b_1 - \lambda b_0)^T (A_1 - \lambda A_0)^+ (b_1 - \lambda b_0) + c_1 - \lambda c_0 \\ \text{if } A_1 - \lambda A_0 \geq 0, \text{ and } b_1 - \lambda b_0 \in R(A_1 - \lambda A_0) \\ 0 \text{ otherwise.} \end{cases}$$

To write the dual problem in a nice way, use Schur-Complements:

Lemma 6. $\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \geq 0 \iff A \geq 0$ *and* $C - B^T A^+ B \geq 0$ *and* $B \in R(A)$ *[B is in the range of A].*

Proof. See BV (Boyd and Vandenberghe). □

(D) Using Schur complement: dual problem is to maximize γ subject to $\lambda \geq 0$ and

$$\begin{bmatrix} A_1 - \lambda A_0 & b_1 - \lambda b_0 \\ (b_1 - \lambda b_0)^T & c_1 - \lambda c_0 - \gamma \end{bmatrix} \geq 0$$

To demonstrate the dual solution is equal to the original optimization problem, we use the S-lemma.

7.2 S-Lemma

Aside: S stands for stability — this lemma originated in control theory where stability is important.

To demonstrate the equality of objectives of (Q) and (D), use S-Lemma:
There are two different versions:

(i) Homogeneous: Assume $\exists \bar{x}$ such that $\bar{x}^T A_0 \bar{x} > 0$. Then $x^T A_0 x \geq 0 \Rightarrow x^T A_1 x \geq 0$ if and only if $A_1 \geq \lambda A_0$ for some $\lambda \geq 0$.

(ii) Inhomogeneous: Assume $\bar{x}^T A_0 \bar{x} + 2b^T \bar{x} + c_0 > 0$ for some $\bar{x}$ [have a quadratic constration qualification], then we have the following implication:

$$x^T A_0 x + 2b_0^T + c_0 \geq 0 \Rightarrow x^T A_1 x + 2b_1^T x + c_1 \geq 0$$

if and only if there exists $\lambda \geq 0$ such that

$$\begin{bmatrix} A_1 - \lambda A_0 & b_1 - \lambda b_0 \\ (b_1 - \lambda b_0)^T & c_1 - \lambda c_0 \end{bmatrix} \geq 0$$

Remark: The inhomogeneous S-Lemma is the Dual of (Q). How does this lemma imply strong duality obtains? How small can we make $x^T A_1 x + 2b_1^T x + c_1$ such that the constraint $\bar{x}^T A_0 \bar{x} + 2b^T \bar{x} + c_0 > 0$ holds? In other words, find the largest γ such that

$$x^T A_1 x + 2b_1^T x + c_1 \geq \gamma$$

with the constraint

$$\bar{x}^T A_0 \bar{x} + 2b_0^T \bar{x} + c_0 > 0$$

Notice the largest γ is the threshold at which $x^T A_1 x + 2b_1^T x + c_1$ becomes indefinite: the S-lemma says

$$\begin{aligned} & x^T A_0 x + 2b_0^T x + c_0 \geq 0 \Rightarrow x^T A_1 x + 2b_1^T x + c_1 \geq \gamma \\ \iff & \begin{bmatrix} A_1 - \lambda A_0 & b_1 - \lambda b_0 \\ (b_1 - \lambda b_0)^T & c_1 - \lambda c_0 - \gamma \end{bmatrix} \geq 0 \end{aligned}$$

If we find the maximal γ for which the above matrix inequality holds, then that is the optimal value of the problem (Q).

Proof of the Homogeneous S-Lemma. ($\Leftarrow$) Trivial: if $A_1 - \lambda A_0 \geq 0$ then $x^T A_1 x \geq \lambda x^T A_0 x$ if $x^T A_0 x \geq 0$, certainly $x^T A_1 x \geq 0$.

($\Rightarrow$) Say that $x^T A_0 x \geq 0$ implies $x^T A_1 x \geq 0$. Consider a semidefinite optimization problem of minimizing $tr(A_1 X)$ subject to

1. $tr(A_0 X) \geq 0$
2. $tr(X) = 1$
3. $X \geq 0$

This is strictly feasible because of our constraint qualifications:

$$X = \frac{\bar{x}\bar{x}^T}{\|\bar{x}\|^2}$$

satisfies $tr(A_0 X) \geq 0$. Set

$$\bar{X} = (1-\epsilon)\frac{\bar{x}\bar{x}^T}{\|\bar{x}\|^2} + \frac{\epsilon}{n} I$$

some small $\epsilon > 0$ [needs to be interior to psd cone]. Strong duality obtains [from past homework] and optimal value is w^*.

Dual problem: $L(X, \lambda, \theta) = tr(A_1 X) - \lambda tr(A_0 X) - \theta(tr(x) - 1)$.

$$\nabla_X L(X, \lambda, \theta) = A_1 - \lambda A_0 - \theta I$$

Need A_0 to be non-negative [otherwise, the infimum goes to $-\infty$]. Taking an infimum over psd matrices to get our dual value, we have:

$$\inf_{X \geq 0} L(X, \lambda, \theta) = \begin{cases} \theta \text{ if } A_1 - \lambda A_0 - \theta I \geq 0 \\ -\infty \text{ otherwise.} \end{cases}$$

gives us the dual problem $\max \theta$ such that $A_1 - \lambda A_0 \geq \theta I$ and $\lambda \geq 0$.

Optimal objective value will be θ and if $\theta \geq 0$, then we are guaranteed that $A_1 - \lambda A_0$ is psd because this is a constraint to maximizing the θ optimization problem. If $w^* \geq 0$, then $\theta \geq 0$ and $A_1 = \lambda A_0 \geq 0$.

Let X_* be optimal, let $\bar{A}_0 = X_*^{1/2} A_0 X_*^{1/2}$, $\bar{A}_1 = X_*^{1/2} A_1 X_*^{1/2}$. Thus $tr(\bar{A}_0) = tr(A_0 X_*) \geq 0$, $tr(\bar{A}_1) = w^*$.

Also $x^T \bar{A}_0 x \geq 0$ implies $x^T \bar{A}_1 x \geq 0$. Let $\bar{A}_0 = UVU^T$ [eigen decomposition] and $x = \sum_{j=1}^n \epsilon_j u_j$ where $\epsilon_j \sim \pm 1$ i.i.d. Then let $U = \begin{bmatrix} u_1 & u_2 & ... & u_n \end{bmatrix}$ be a

matrix of column vectors u_i. We have:

$$\begin{aligned} x^T \bar{A}_0 x &= \epsilon^T U^T \bar{A}_0 U \epsilon \\ &= \sum_{j=1}^{n} \lambda_j \epsilon_j^2 \\ &= \sum_{j=1}^{n} \lambda_j \\ &= tr(\bar{A}_0) \geq 0 \end{aligned}$$

So $0 \leq x^T \bar{A}_1 x \geq 0$, and

$$\begin{aligned} \mathbb{E}[x^T \bar{A}_1 x] &= tr(\mathbb{E}[\bar{A}_1 x x^T]) \\ &= tr(\bar{A}_1 \mathbb{E}[U \epsilon \epsilon^T U^T]) \end{aligned}$$

on-diagonal entries are 1 and off-diagonals are $0 = tr(\bar{A}_1 I_n) = tr(\bar{A}_1)$.

□

Proof of the Inhomogeneous S-Lemma. ($\Leftarrow$) trivial.

($\Rightarrow$) Goal: relate to the homogeneous case.

Let $\hat{A}_i = \begin{bmatrix} A_i & b_i \\ b_i^T & c_i \end{bmatrix}$ $i = 0, 1$. Define:

$$f_i(\begin{bmatrix} x \\ t \end{bmatrix}) = \begin{bmatrix} x \\ t \end{bmatrix}^T \hat{A}_i \begin{bmatrix} x \\ t \end{bmatrix} = x^T A_i x + 2t b_i^T x + c_i t^2$$

for $i = 0, 1$ Then

$$\hat{f}_0(\begin{bmatrix} x \\ 1 \end{bmatrix}) \geq 0 \Rightarrow \hat{f}_1(\begin{bmatrix} x \\ 1 \end{bmatrix}) \geq 0$$

as $t^2 \hat{f}_i(\begin{bmatrix} x/t \\ 1 \end{bmatrix}) = \hat{f}_i(\begin{bmatrix} x \\ t \end{bmatrix})$ for any $t \neq 0$. Similarly, we get that

$$\hat{f}_0(\begin{bmatrix} x \\ t \end{bmatrix}) \geq 0 \Rightarrow \hat{f}_1(\begin{bmatrix} x \\ t \end{bmatrix}) \geq 0$$

for any $t \neq 0$.

For $t = 0$, we approximate. Use

$$\begin{bmatrix} x + t\bar{x} \\ t \end{bmatrix}^T \begin{bmatrix} A_0 & b_0 \\ b_0^T & c_0 \end{bmatrix} \begin{bmatrix} x + t\bar{x} \\ t \end{bmatrix} = x^T A_0 x + 2t x^T (A_0 \bar{x} + b_0) + t^2 (\bar{x}^T A_0 \bar{x} + 2 b_0^T \bar{x} + c_0)$$

and observe $(\bar{x}^T A_0 \bar{x} + 2b_0^T \bar{x} + c_0) > 0$ [Slater's condition]. Can take t to be positive or negative to get any sign in $2tx^T(A_0\bar{x} + b_0)$. Then if

$$\hat{f}_0 \begin{bmatrix} x \\ 0 \end{bmatrix} = x^T A_0 x \geq 0$$

then taking $t > 0$ or $t < 0$ as appropriate, we get that

$$\hat{f}_0(\begin{bmatrix} x + t\bar{x} \\ t \end{bmatrix}) > 0$$

So in particular, $\hat{f}_1(\begin{bmatrix} x + t\bar{x} \end{bmatrix}) \geq 0$. Taking $t \to 0$, we see that

$$\hat{f}_0(\begin{bmatrix} x \\ 0 \end{bmatrix}) \geq 0 \Rightarrow \hat{f}_1(\begin{bmatrix} x \\ 0 \end{bmatrix}) \geq 0$$

. i.e. $\hat{f}_0 \geq 0$ implies $\hat{f}_1 \geq 0$. So homogeneous S-lemma gives $\exists \lambda \geq 0$ such that $\hat{A}_1 \geq \lambda \hat{A}_0$ i.e.,

$$\begin{bmatrix} A_1 & b_1 \\ b_1^T & c_1 \end{bmatrix} \geq \lambda \begin{bmatrix} A_0 & b_0 \\ b_0^T & c_0 \end{bmatrix}$$

□

This is the S-lemma that guarantees us that strong duality holds for non-convex quadratic problems.

8 February 28

Motivation: It is desirable to have classes of functions such that when you apply Newton's method, the analysis looks "right". Specifically, the third derivative of this class of functions is somehow related to the second derivative. We call these functions **Self-Concordant (SC)**, and the existence of self-concordant functions allows us to solve convex optimization problems in polynomial time.

8.1 Self-Concordance

Definition 24 (Self-Concordant Functions)**.** *$f : \mathbb{R} \to \mathbb{R}$ is self-concordant (SC) if*

1. *it is C^3 [3-times continuously differentiable].*
2. *it is convex.*
3. *$|f'''(x)| \leq 2(f''(x))^{\frac{3}{2}}$*

The last condition ensures the third derivative is controlled by the second derivative. Specifically, Newton's method is accurate since the 3rd derivative looks somewhat like the 2nd derivative.

Definition 25 (Strongly Non-Degenerate Self-Concordant Functions)**.** $f : \mathbb{R}^n \to \mathbb{R}$ *is* ***(strongly non-degenerate) (SC)*** *if* $\nabla^2 f(x) > 0$ *for all* $x \in$ *domain* f *and for each* $x \in dom(f)$, $v \in \mathbb{R}^n$, $t \to f(x+tv)$ *is (SC).*

Equivalently, $\nabla^3 f(x)[u,u,u] \leq 2\langle u\ ,\ \nabla^2 f(x)u\rangle^{\frac{3}{2}}$ *for any* $u \in \mathbb{R}^n$.

Remark 3 (Third Derivatives)**.** *The third derivative is similar to the Hessian, except bigger.* $\nabla^3 f(x)[u,\cdot,\cdot]$ *is a matrix in* $\mathbb{R}^{n\times n}$ *so that*

$$\nabla^3 f(x)[u,\cdot,\cdot] = \lim_{t\to 0} \frac{\nabla^2 f(x+tu) - \nabla^2 f(x)}{t}$$

In general, if $T \in \mathbb{R}^{n\times n\times n}$, *then* $T[u,v,w] = \sum_{i,j,k} T_{i,j,k} u_i v_j w_k$.

Remark 4. *There is powerful duality theory around self-concordant functions. In particular,* f *is self-concordant* $\iff$ $f^*(y) = \sup_x\{\langle y\ ,\ x\rangle - f(x)\}$ *is self-concordant.*

Example 3. *Optimization problem:*
minimize $f(x)$ *subject to* $x \in K$ *where* K *is some convex cone.*

One simply way to solve this problem is to take the dual after introducing an auxiliary variable and setting it equal to x*:*

Let $z = x$ *and consider* $\min f(x)$ *subject to* $z = x$, $z \in K$. *Taking the Lagrangian , we have:*

$$L(x,z,\beta) = f(x) - \langle \beta\ ,\ x\rangle + \langle \beta\ ,\ z| + I_{K(z)}\rangle$$

where $I_{K(z)}$ *is the indicator function for whether* $z \in K$ *or not.*

$$\begin{aligned}\inf_\lambda\{f(x) - \langle\beta\ ,\ x\rangle\} &= -f^*(\beta)\\ &= \inf_Z \langle\beta\ ,\ z\rangle + I_{K(z)}\end{aligned}$$

if β *is in the dual cone, set* $z = 0$. *If* β *is not in the dual cone, there is some* z *in the cone so that the inner product between* β *and* z *is* $-\infty$*:*

$$\begin{cases} 0 \text{ if } \beta \in K^* \\ -\infty \text{ otherwise.}\end{cases}$$

$\max -f^*(\beta)$ *such that* $\beta \in K^*$ *is equivalent. Nice properties between dual and primal of SC functions.*

Theorem 13. *Let* f *be S.C. [strongly non-degenerate]. Then* f^* *is S.C. and for the pair* (x,s) *satisfying* $\nabla f(x) = s$ *(i.e.* $x = \nabla f^*(s)$*),* $\nabla^2 f(x)^{-1} = \nabla^2 f^*(s)$.

Proof. First recall $\nabla f(x) = s$ if and only if $x \in \partial f^*(s)$. Then $x = x(s)$ minimizes $f(x) - \langle s\ ,\ x\rangle$. [$f(x)$ is strongly convex because $\nabla^2 f(x) > 0$: follows from SC]. Then as f is SC, x is unique, and $\partial f^*(s) = \{x(s)\}$.

So $\nabla f^*(s)$ exists and is equal to x.

Hessians? As $\nabla^2 f(x) > 0$ for all $x \in dom(f)$, we can apply the inverse function theorem [requires derivative matrix be non-degenerate]. By the inverse function theorem, $\nabla f(y) = s + \Delta$ is solvable for y when Δ is small, $\nabla f(x) = s$.

Even more: $s \to x(s)$ solves $\nabla f(x(s)) = 0$ is C^2 as $\nabla^2 f(x)$ is C^2. That is $\nabla f^*(s) = x(s)$ is C^2 [2x continuously differentiable], and $\nabla f(\nabla f^*(s)) = s$, then taking ∇_s on both sides gives us:

$$\begin{aligned}
\nabla_s(\nabla f(\nabla f^*(s))) &= \nabla_s(s) = I_n \\
\nabla^2 f(\nabla f^*(s)) \underbrace{\nabla_s \nabla f^*(s)}_{\nabla^2 f^*(s)} &= I_n \\
\nabla^2 f(\nabla f^*(s)) \nabla^2 f^*(s) &= I_n
\end{aligned}$$

So we have at $x = \nabla f^*(s)$, $\nabla^2 f(x)\nabla^2 f^*(s) = I_n$.

Remark 5 (Third Derivatives of Multivariate Functions)**.** *Consider $x \in \mathbb{R}$ first. For $t \to 0$,*

$$\begin{aligned}
(f^*)''(s+t) &= \underbrace{f''(x(s+t))^{-1}}_{\text{using } x=x(s) \text{ from the previous proof about duality.}} \\
&= f''(x(s) + \dot{x}(s)t + o(t))^{-1} \\
&= f''(x(s) + \dot{x}(s)t + o(t))^{-1} \\
&= (f''(x) + f'''(x)\dot{x}(s)t + o(t))^{-1} \\
&= \frac{1}{f''(x)} - \frac{f'''(x)\dot{x}(s)t}{f''(x)^2} + o(t)
\end{aligned}$$

as $t \to 0$ because $\frac{1}{y+\Delta} = \frac{1}{y} - \frac{\Delta}{y^2} + O(\delta^2)$

Note $x(s) = (f^*)'(s)$ and ($\dot{x}$ is equal to derivative w.r.t. argument) so $\dot{x}(s) = (f^*)''(s)$. So we have

$$(f^*)''(s+t) = f''(x)^{-1}) - \underbrace{\frac{f'''(x)}{f''(x)^2}(f^*)''(s)\, t}_{\text{first order term}} + \underbrace{o(t)}_{\text{higher-order terms}}$$

So

$$\begin{aligned}(f^*)''(s) &= \frac{-f'''(x)}{f''(x)^2}(f^*)''(s) \\ &= \underbrace{\frac{-f'''(x)}{f''(x)^{3/2}}}_{\leq 2} \cdot \underbrace{\frac{(f^*)''(s)}{\sqrt{f''(x)}}}_{=(f^*)''(s)^{3/2}} \\ &\leq 2(f^*)''(s)^{3/2}\end{aligned}$$

This is the 1-dimensional proof of self-concordance.

In the full-dimensional case, it's the same argument. Let $s, v \in \mathbb{R}^n$, $t \in \mathbb{R}$ small. Recall

$$(A+E)^{-1} = A^{-1} - A^{-1}EA^{-1} + O(|E|^2)$$

Expand via first-order expansion:

$$\begin{aligned}\nabla^2 f^*(s+tv) &= \nabla^2 f(\underbrace{x(s+tv)}_{x(s)=\nabla f^*(s) \ \& \ \nabla_s x(s)=\nabla^2 f^*(s)})^{-1} \\ &= \nabla^2 f(x(s) + t\nabla^2 f^*(s)v + o(t))^{-1} \\ &= (\nabla^2 f(x) + \nabla^3 f(x)[\nabla^2 f^*(s)v, \cdot, \cdot] + o(t))^{-1} \\ &= (\nabla^2 f(x) + \nabla^3 f(x)[\nabla^2 f^*(s)v, \cdot, \cdot] + o(t))^{-1} \\ &= \nabla^2 f(x)^{-1} \underbrace{- \nabla^2 f(x)^{-1}\nabla^3 f(x)[\underbrace{\nabla^2 f^*(s)v}_{=\nabla^2 f(x)^{-1}}, \cdot, \cdot]\nabla^2 f(x)^{-1} \cdot t + o(t)}_{\text{Third order term.}}\end{aligned}$$

using earlier expansion $(A+E)^{-1}$ [note: $o(t)$ is a small error term].

So we put in arguments to the 2nd and 3rd thing into the 3-tensor: Using notation

$$u^T T[v, \cdot, \cdot]w = T[v, u, w]$$

we get that

$$\begin{aligned}\nabla^3 f^*(x)[v,v,v] &= -\nabla^3 f(x)[\nabla^2 f(x)^{-1}v, \nabla^2 f(x)^{-1}v, \nabla^2 f(x)^{-1}v] \\ &\leq 2\langle \nabla^2 f(x)^{-1}v \ , \ \nabla^2 f(x)\nabla^2 f(x)^{-1}v\rangle^{3/2} \\ &= 2\langle v \ , \ \nabla^2 f(x)^{-1}v\rangle^{3/2} \\ &= 2\langle v \ , \ \nabla^2 f^*(s)v\rangle^{3/2}\end{aligned}$$

Note that $x = x(s)$ throughout this proof. □

8.2 Logarithmically Homogeneous Functions

Logarithmically homogeneous functions form a very important class of self-concordant functions (SC).

Definition 26. *Let K be a convex cone. Then $f : intK \to \mathbb{R}$ is $\nu-$log-homogeneous if $f(tx) = f(x) - \nu \log t$*

Example 4. *$f(x) = \sum_{i=1}^{n} \log x_i$, $f : \mathbb{R}^n_{++} \to \mathbb{R}$. Then $f(tx) = f(x) - n \log t$.*

Let $X > 0$ and define $f(X) = -\log \det(X)$ is a self-concordant function and is logarithmically homogeneous.

$f(tX) = -\log \det(tX) = -\log(t^n \det(X)) = -\log \det(x) - n \log t$

In homework, you'll show that for such functions, there's an even stronger duality theory, and

$$\lambda(x) = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

This is called Newton's Decrement: how much progress one step of newton's makes] $\lambda(x) = \nu$ [is a constant].

9 March 6

9.1 General Cone Convex Optimization Problem:

$$\begin{aligned} &\min c^T x \\ &\text{such that} \\ &Ax = B \\ &x \in K \end{aligned}$$

Where K is a proper convex cone: $\int K \neq \emptyset$ and K has no lines [i.e. not an entire halfspace]. We call this a Cone Problem (CP).

To solve (CP) use interior point methods: we need a logarithmic barrier for the cone K: $\phi : K^\circ \to \mathbb{R}$ [recall K° is the interior of K and ∂K is the boundary] with the property that $\phi(tx) = \phi(x) - \nu \log t$ for any $t > 0$ and $x \in K^\circ$. ν is the homogeneity parameter of the function with $\nu > 0$.

Barrier Condition: $\|\nabla \phi(x)\| \to \infty$ or $\phi(x) \to \infty$ as $x \to \partial K$.

Example 5. *$\phi(x) = -\sum_{i=1}^{n} \log x_i$ is n-log-homogeneous with the domain of $\phi = \mathbb{R}_{++}$.*

Example 6. *$\phi(x) = -\log det(X)$ is also n-log-homogeneous with the domain of $\phi = \{X > 0\}$*

Given such a barrier function, we solve (CP) by iterative solving

$$\begin{aligned} &\min tc^T x + \phi(x) \\ &\text{subject to} \\ &Ax = b \end{aligned}$$

and take $t \to \infty$ to solve this (typical interior point method).

Recall the complexity of interior point methods scale at worst with ν [homogeneity parameter]. However, it is not clear how we find the logarithmic barrier. Specifically, given an arbitrary convex cone in $\mathbb{R}^n$, is there a logarithmic barrier for this convex cone? Yes — so we can solve any convex cone optimization problem in polynomial time after we get the logarithmic barrier.

Theorem 14. *Let $K \subset \mathbb{R}^n$ be a proper convex cone [has an interior and contains no lines]. Define*

$$f(\theta) := \log \int_{-K} e^{\langle \theta \ , \ x \rangle} dx$$

Then

$$f^*(x) := \sup_\theta \{\langle x \ , \ \theta \rangle - f(\theta)\}$$

is an n-log-homogeneous and self-concordant barrier for K^* [the dual cone]. [This is optimal]

We will take a detour through exponential families to develop this result.

Definition 27 (Exponential Family Model)**.** *Say random variable $X \in \mathbb{R}^n$ has exponential family distribution on a set $\mathcal{X} \in \mathbb{R}^n$ if for $x \in \mathcal{X}$ its density is*

$$p_\theta(x) = exp(\theta^T x - f(\theta))$$

where

$$f(\theta) = \log \int_{\mathcal{X}} exp(x^T \theta) dx$$

Clearly $\int_{\mathcal{X}} p_\theta(x) dx = 1$ and $p_\theta(x) \geq 0$ [so a valid probability measure].

This function f is the log-partition function [also called the Cumulant generating function]. Properties of f:

1. f is convex and $domf = \{\theta | f(\theta) < \infty\}$ is convex [consequence of Hölder's inequality].
2. It is C^∞ on the interior of the domain of f.

Key Insight: Derivatives of f give us moments of X. More formally, derivatives of f are in bijection with moments of X under the density p_θ.

Proposition 2. *Let*

$$\mathbb{E}_\theta[T(X)] = \int T(x)p_\theta(x)dx$$

then

$$\nabla f(\theta) = \mathbb{E}_\theta[X] = \mu(\theta)$$

[this is called the mean mapping]. For higher derivatives, we have:

$$\nabla^2 f(\theta) = \mathbb{E}_\theta[(X - \mu(\theta))(X - \mu(\theta))^T] = Cov_\theta(X)$$

and

$$\nabla^3 f(\theta) = \mathbb{E}_\theta[(X - \mu^\theta)^{\times 3}] = \mathbb{E}_\theta[(x_i - \mu_i(\theta))(x_j - \mu_j(\theta))(x_k - \mu_k(\theta))]$$

Proof. For the first derivative.

$$\nabla f(\theta) = \frac{1}{\int e^{\theta^T x}dx} \int x e^{\theta^T x}dx = \int x p_\theta(x)dx$$

For the second derivative.

$$\begin{aligned}
\nabla^2 f(\theta) &= \frac{\int x x^T e^{\theta^T x}dx}{\int e^{\theta^T x}dx} - \frac{(\int x e^{\theta^T x}dx)(\int x e^{\theta^T x}dx)^T}{(\int e^{\theta^T x}dx)^2} \\
&= \int x x^T p_\theta(x)dx - (\int x p_\theta(x)dx)(\int x p_\theta(x)dx)^T \\
&= \mathbb{E}_\theta[xx^T] - \mu(\theta)\mu(\theta)^T \\
&= Cov_\theta(X)
\end{aligned}$$

Third derivative requires more algebra... □

Question: When can we relate higher order moments of Random Variables to lower order moments? If you can say the third moments of e are bounded by the second moments, then f is self-concordant. And if f is self-concordant, then its dual is self-concordant. What families of probability distributions can we do this for? Will define a general family where this will work. Note that the family of distributions with log-concave densities is sufficient for this.

Definition 28. *A probability measure P on $\mathbb{R}^n$ is* log*-concave if for all $A, B \subset \mathbb{R}^n$*

$$P(\lambda A + (1-\lambda)B) \geq P(A)^\lambda P(B)^{1-\lambda}$$

Question: What is a sufficient condition for a probability measure to be log-concave?

Proposition 3. *If P has a continuous density, p, then P is a log-concave probability measure $\iff$ $p(x) = e^{-u(x)}$ where u is a convex function.*

Proof. 5-line argument with Prékopa-Leindler [HW]. First important, but really deep step in this result. □

Second Key to relating moments is called concentration of measure. Basic idea: if we have a log-concave distribution, then the vast majority of its probability mass is concentrated in one location, so everything behaves as it does on average. For log-concave P, vast majority of the probability mass concentrates in "average" places (all moments are roughlyh the same).

Proposition 4. *Let P be any log-concave distribution. Let $r \geq 1$ and $A \subset \mathbb{R}^n$ be a symmetric and convex set. Then*

$$1 - P(rA) \leq (\frac{1 - P(A)}{P(A)})^{r/2} \sqrt{P(A)(1 - P(A))}$$

Consequence: if $P(A) \geq \frac{e}{e+1}$. Then $(\frac{1-P(A)}{P(A)} \leq \frac{1}{e}$ and $1-P(rA) \leq e^{-r/2} \sqrt{\frac{e}{(e+1)^2}}$.

Informally, this proposition tells us that as soon as you enlarge A ever so slightly, you get exponentially close to the entire probability mass. For a set with initial probability mass $\frac{1}{2}$, a slight enlargement controls an exponentially large fraction of the entire probability space.

Proof. Homework. "Not that bad" — author's note: we will see... □

Final step: Control higher moments with lower moments.
Let $F \geq 0$ be $1-$positively homogeneous and symmetric. That is to say,

$$F(\alpha x) = |\alpha| F(x)$$

then every moment of F can be controlled by the first moment.

Corollary 7. *For any M such that*

$$P(F(X) \geq M) \leq \frac{1}{e+1} \approx \frac{1}{4}$$

then

$$\mathbb{E}[|F(X)|^q]^{1/q} \leq C \cdot Mq$$

where C is a universal constant for any $q \geq 1$ [all moments are controlled by the first moment].

In particular, if X has a log-concave density, then we get

$$\mathbb{E}[F(X - \mathbb{E}[X])^q]^{1/q} \leq Cq \cdot \mathbb{E}[F(X - \mathbb{E}[X])^2]^{1/2}$$

Consequence: If X has the exponential family density

$$p_\theta(x) = exp(\theta^T x - f(\theta))$$

then certainly its density is log-concave, and

$$\begin{aligned}\nabla^3 f(\theta)[u,u,u] &= \mathbb{E}[|(X-\mu(\theta))^T u|^3] \\ &\le C \cdot \mathbb{E}[((X-\mu(\theta))^T u)^2]^{3/2} \\ &= C \cdot (u^T \nabla^2 f(\theta) u)^{3/2}\end{aligned}$$

In words, all of probability mass concentrated in one area, so all moments are "similar", and log-partition function bounds the moments. Deep connections between convex geometry stuff and construction of barrier functions for optimization, etc.

Proof. By homogeneity of F,

$$\{x|F(x) \ge rM\} = r\underbrace{\{x|F(x) \ge M\}}_{\text{symmetric set}}$$

and $\{x|F(x) \le M\}$ is convex.

Then

$$\mathbb{P}(|F| \ge rM) \le \frac{\sqrt{e}}{e+1} e^{-r/2}$$

for all $r \ge 1$ and

$$\begin{aligned}\mathbb{E}[|F|^q] &= \underbrace{\int_0^\infty \mathbb{P}(|F|^q \ge t)dt}_{\text{as } \mathbb{E}[Z = \int_0^\infty \mathbb{P}(Z \ge t)dt \text{ if } Z \ge 0} \\ &\le 1\underbrace{\int_0^{M^q} dt}_{M^q} + \underbrace{\int_{M^q}^\infty \mathbb{P}(|F|^q \ge t)dt}_{\star} \\ \star &= qM^q \int_1^\infty \underbrace{\mathbb{P}(|F|^q \ge M^q r^q)}_{\le e^{-r/2}\sqrt{\frac{e}{(e+1)^2}}} r^{q-1} dr\end{aligned}$$

Recall that $t = M^q r^q$ i.e., $qM^q r^{q-1} dr = dt$, so

$$\begin{aligned}\star &\le qM^q \int_1^\infty e^{-r/2} r^{q-1} dr \\ &\le 2^q qM^q \underbrace{\int_0^\infty e^{-u} u^{q-1} du}_{\text{using } u=\frac{r}{2} \text{ and } 2du=dr} \\ &= (2M)^q \Gamma(q+1)\end{aligned}$$

with $\Gamma(q)^{1/q} \leq exp(\frac{q \log q}{q}) = e^{\log q} \leq q$.

Recall the gamma function is: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. □

10 March 13

Outline:

1. Model-based minimization
2. Models
3. Gradient Mappings
4. Beyond Complexity (i.e. non-convex)

10.1 Minimization Algorithms

$\min f(x)$ such that $x \in X$, $X \subset \mathbb{R}^n$ closed and convex.
At each iteration, you model f around iterate x_k, minimize model (+ regularization to stay near points where the model is accurate).

Model of f at a point $x \in \mathbb{R}^n$ is a convex function f_x [to denote the model is centered around x]

c1 $f_x(x) = f(x)$

c1' if f is convex, then $f_x(y) \leq f(y)$ $\forall y \in \mathbb{R}^n$

c2 Quadratic accuracy $\exists M < \infty$ such that $|f_x(y) - f(y)| \leq \frac{M}{2} \|x - y\|_2^2$

Example: f is convex, $g \in \partial f(x)$, then set $f_x(y) = f(x) + \langle g \ , \ y - x \rangle \leq f(y)$ [linear underestimator] (so c1, c1' are immediate).
Example: f has an L-Lipschitz gradient, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.
Claim: $f_x(y) = f(x) + \langle \nabla f(x) \ , \ y - x \rangle$ is L-quadratically accurate. i.e. $|f_x(y) - f(y)| \leq \frac{L}{2} \|x - y\|_2^2$ *Proof*: Fundamental theorem of calculus where we say $h(t) = f(x + t(y - x))$ and then apply $h'(t) = \langle \nabla f(x + t(y - x)) \ , \ y - x \rangle$.

$$
\begin{aligned}
f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y - x)) \ , \ y - x \rangle dt \\
&= f(x) + \langle \nabla f(x), y - x \ , \ + \rangle \int_0^1 \langle \underbrace{\nabla f(x + t(y - x)) - \nabla f(x)}_{\leq Lt\|x - y\|} \ , \ \underbrace{y - x}_{\leq \|y - x\|} \rangle dt \\
&= f(x) + \langle \nabla f(x) \ , \ y - x \rangle \pm L\|x - y\|_2^2 \underbrace{\int_0^1 t dt}_{\frac{1}{2}}
\end{aligned}
$$

10.2 Methods

For a fixed step size $\alpha > 0$, iterate that $x_{k+1} \leftarrow \arg\min_{x \in X}\{f_{x_k}(x) + \frac{1}{2\alpha}\|x - x_k\|_2^2$.

Note: If $X = \mathbb{R}^n$, model f_{x_k} is just linear: $f_x(y) = f(x) + \nabla f(x)^T(y - x)$ $\Rightarrow x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k)$

Key to analysis: one-step progress. Fix $x_0 \in X$, let $x_\alpha := \arg\min_{x \in X}\{f_{x_0}(x) + \frac{1}{2\alpha}\|x - x_0\|_2^2\}$. We want to understand the progress we might get out of one update.
Gradient Mapping: $G_\alpha(x_0) := \frac{1}{\alpha}(x_0 - x_\alpha)$ is a normalized version of how much progress we make normalized by the α. A generalization of the gradient – notion of how much progress each method is making.
Note: If $X = \mathbb{R}^n$ and f_{x_0} is Linear, $x_\alpha = x_0 - \alpha\nabla(x_0)$, $G_\alpha(x_0) = \nabla f(x_0)$.

Lemma 7 (Optimality in Convex Optimization)**.** *If h is convex, subdifferentiable on X, then $\hat{x}$ minimizes h on X $\iff$ for some $g \in \partial h(\hat{x})$, $\langle g\ ,\ y - \hat{x}\rangle \geq 0$ for all $y \in X$*

Proof. Exercise ft. drawing. □

Lemma 8. *One-step progress. Let C_1, C_2 hold. Then $f(x\alpha) \leq f(x_0) - \alpha(1 - \frac{M_\alpha}{2}\|G_\alpha(x_0)\|_2^2$. If $\alpha \leq \frac{1}{M}$ then $f(x_0) \leq f(x_0) - \frac{\alpha}{2}\|G_\alpha(x_0)\|_2^2$.*

Proof. Let $g_\alpha \in \partial f_{x_0}(x_\alpha)$ and then use optimality conditions for convexity:

$$\langle g_\alpha + \frac{1}{\alpha}(x_\alpha - x_0)\ ,\ y - x_\alpha\rangle \geq 0$$

for all $y \in X$. □

Then by convexity of f_{x_0} $f(x_\alpha) \leq f_{x_0}(x_\alpha) + \frac{M}{2}\|x_0 - x_\alpha\|_2^2 \leq \underbrace{f_{x_0}(x_0)}_{=f(x_0)} + \langle g_\alpha\ ,\ x_\alpha - x_0\rangle +$
$\frac{M}{2}\|x_0 - x_\alpha\|^2 \leq f(x_0) + \frac{1}{\alpha}\langle x_0 - x_\alpha\ ,\ x_\alpha - x_0\rangle + \frac{M}{2}\|x_0 - x_\alpha\|^2$ by substituting $y = x_0 in \star$ with $\star\langle g_\alpha + \frac{1}{\alpha}(x_\alpha - x_0)\ ,\ y - x_\alpha\rangle \geq 0$
$= f(x_0) - \frac{1}{\alpha}\|x_0 - x_\alpha\|^2 + \frac{M}{2}\|x_0 - x_\alpha\|^2$

Lemma 9. *Let f be convex, C1' holds and $\alpha \leq \frac{1}{M}$. Then*

$$f(x_\alpha) \leq f(y) + \frac{1}{2\alpha}[\|x_0 - y\|^2 - \|x_\alpha - y\|^2]$$

for any $y \in X$.

Proof. $f(x_\alpha) \leq f_{x_0}(x_\alpha) + \frac{M}{2}\|x_\alpha - x_0\|^2$ □

Observe that by convexity of the model $f_{x_0}(x_\alpha) \leq f_{x_0}(y) + \langle g_\alpha\ ,\ x_\alpha - y\rangle$. Can use $\star$ again

$\leq f_{x_0}(y) + \frac{1}{\alpha}\langle x_\alpha - x_0\ ,\ y - x_\alpha\rangle$ for any y by $\star$. But there's a magical 3-square identity for inner product that helps here:

$$\langle x_\alpha - x_0\ ,\ y - x_\alpha\rangle = \frac{1}{2}[\|x_0 - y\|^2 - \|x_\alpha - y\|^2 - \|x_\alpha - x_0\|^2]$$

Substitute this into the first inequality in the proof:
$f(x_\alpha) \leq \underbrace{f_{x_0}(y)}_{\leq f(y)} + \frac{1}{2\alpha}[\|x_0 - y\|^2 - \|x_\alpha - y\|^2]$ (used $\alpha \leq \frac{1}{M}$, i.e., $M \leq \frac{1}{\alpha}$).

Theorem 15. *If f is convex, c1', c2 and $\alpha \leq \frac{1}{M}$, then*

$$f(x_{k+1}) - f(x^*) \leq \frac{1}{2\alpha k}\|x_1 - x^*\|_2^2$$

Proof. $f_k = f(x_k)$ is decreasing. Take $y = x^*$, then

$$k(f_{k+1} - f^*) \leq \sum_{i=1}^{k} f_{i+1} - f(x^*) \leq \frac{1}{2\alpha}\sum_{i=1}^{k}[\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2]$$

□

10.3 Beyond Convexity

Consider composite optimization problem of minimizing $h(c(x))$ where h is convex, c is smooth (i.e. differentiable).

Example: Phase retrieval [molecule, shoot a laser at it, record diffraction patterns – can only recover magnitude of the wave, not the phase]. $b_i = |\langle a_i\ ,\ x_*\rangle|^2$ for $i = 1, ..., m$, $x_* \in \mathbb{C}^n$. A very natural objective is to define

Let $h(z) = \|z\|_1$ and let $c(x) = |Ax|^2 - b$ where $A = \begin{bmatrix} - - - a_1^* - - \\ \vdots \\ - - - a_m^* - - \end{bmatrix}$. Solve $\min_{x\in\mathbb{C}^n} h(c(x))$.

General recipe: Assume h is L_h-Lipschitz and ∇C is L_C-Lipschitz. h is convex, but c very much isn't. but if we take a lienar approximation to c, then ... Natural model becomes $f_x(y) = h(\underbrace{c(x) + \nabla c(x)^T(y - x))_{convexiny}}$ therefore $\|f_x(y) - f(y)| \leq \|\| \leq L_h\|c(x) + \nabla c(x)^T(y - x) - c(y)\| \leq \frac{L_h \cdot L_c}{2}\|x - y\|_2^2$. Make sequences of convex approximations.
General Result: $x_{k+1} \leftarrow \arg\min_{x\in X}\{f_{x_k}(x) + \frac{1}{2\alpha}\|x - x_k\|_2^2\}$ and take $\alpha = \frac{1}{L_h L_c} = \frac{1}{M}$ we get that the minimium $\min_{i\leq k}\|G_\alpha(x_*)\|_2^2 \leq \frac{L_h L_c}{K} \cdot \Delta_f$ where $\Delta_f = f(x_1) - f^*$

Proof.

$$f(x_{k+1}) - f(x_k) \leq -\frac{\alpha}{2}\|G_\alpha(x_k)\|_2^2$$

$\Rightarrow \sum_{i=1}^{k} \frac{\alpha}{2} \|G_\alpha(x_i)\|_2^2 \leq \sum_{i=1}^{k} f(x_i) - f(x_{i+1}) \leq f(x_1) - f(x_{k+1}) \leq f(x_1) - f^*$ □

Even if you have super weird non-convex funcitons, can effectively find "stationary points".

11 Acknowledgements

11.1 Identifying Errors and Fixing Them

1. Kasper Johansson
2. Danny Tse

12 Appendix

12.1 Matrix Calculus

One of the things I've struggled with is differentiating through terms in matrix calculus. When does a term become transposed when differentiating and when does it stay the same? These questions are easily answered, but all too often, poorly explained. When in doubt, return to the Jacobian.

Definition 29 (Jacobian)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a multivariate function. The Jacobian, or equivalently $\frac{\partial f}{\partial x}$, is the matrix of partial derivatives:*

$$\begin{aligned}\frac{\partial f}{\partial x} &= \nabla_x f \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & & \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}\end{aligned}$$

*This is **just** a definition. It could have just as easily have been defined the other way:*

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & & & \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

but early on, mathematicians agreed on the first form.

Let's look at a concrete example of how this might help us.

Example 7. *Let $x \in \mathbb{R}^{n\times 1}$ and $\theta \in \mathbb{R}^{n\times 1}$. Then define the function*

$$\begin{aligned} f(x) &= x^T\theta \\ &= \begin{bmatrix} x_1 & x_2 & \dots x_n \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \\ &= x_1\theta_1 + x_2\theta_2 + ... + x_n\theta_n \end{aligned}$$

We can compute the Jacobian as defined above:

$$\begin{aligned} \frac{\partial f}{\partial x}\left[x^T\theta\right] &= \nabla_x \left[x^T\theta\right] \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1}(x_1\theta_1 + x_2\theta_2 + ... + x_n\theta_n) & \dots & \frac{\partial}{\partial x_n}(x_1\theta_1 + x_2\theta_2 + ... + x_n\theta_n) \end{bmatrix} \\ &= \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_n \end{bmatrix} \\ &= \theta^T \end{aligned}$$

Example 8. *We can extend this to $m > 1$. Let's consider $\theta \in \mathbb{R}^{n\times m}$.*
We define the function

$$\begin{aligned} f(x) &= x^T\theta \\ &= \begin{bmatrix} x_1 & x_2 & \dots x_n \end{bmatrix} \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \dots & \theta_{1,m} \\ \theta_{2,1} & \theta_{2,2} & \dots & \theta_{2,m} \\ \vdots & & & \\ \theta_{n,1} & \theta_{n,2} & \dots & \theta_{n,m} \end{bmatrix} \\ &= \begin{bmatrix} (x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) & \dots & (x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) \end{bmatrix} \end{aligned}$$

We can compute the Jacobian as defined above:

$$\begin{aligned} \frac{\partial f}{\partial x}\left[x^T\theta\right] &= \nabla_x \left[x^T\theta\right] \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & & \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1}(x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) & \dots & \frac{\partial}{\partial x_n}(x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) \\ & \vdots & \\ \frac{\partial}{\partial x_1}(x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) & \dots & \frac{\partial}{\partial x_n}(x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) \end{bmatrix} \\ &= \begin{bmatrix} \theta_{1,1} & \theta_{2,1} & \dots & \theta_{n,1} \\ \vdots & & & \\ \theta_{1,m} & \theta_{2,m} & \dots & \theta_{n,m} \end{bmatrix} \\ &= \theta^T \end{aligned}$$

Example 9. *Nice! So we can now differentiate what happens when $x^T\theta$. But what about the other way around? What if we do $\theta^T x$?*

$$\begin{aligned} f(x) &= \theta^T x \\ &= \begin{bmatrix} \theta_{1,1} & \theta_{2,1} & \dots & \theta_{n,1} \\ \theta_{1,2} & \theta_{2,2} & \dots & \theta_{n,2} \\ \vdots & & & \\ \theta_{1,m} & \theta_{2,m} & \dots & \theta_{n,m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} (x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) \\ \vdots \\ (x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) \end{bmatrix} \end{aligned}$$

We can compute the Jacobian:

$$\begin{aligned} \nabla_x \left[\theta^T x\right] &= \nabla_x \begin{bmatrix} (x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) \\ \vdots \\ (x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1}(x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) & \dots & \frac{\partial}{\partial x_n}(x_1\theta_{1,1} + x_2\theta_{2,1} + ... + x_n\theta_{n,1}) \\ \vdots & & \\ \frac{\partial}{\partial x_1}(x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) & \dots & \frac{\partial}{\partial x_n}(x_1\theta_{1,m} + x_2\theta_{2,m} + ... + x_n\theta_{n,m}) \end{bmatrix} \\ &= \begin{bmatrix} \theta_{1,1} & \dots & \theta_{n,1} \\ \vdots & & \\ \theta_{1,m} & \dots & \theta_{n,m} \end{bmatrix} \\ &= \theta^T \end{aligned}$$

Notice that this time the gradient is θ^T i.e. the input to the function. It does not transpose when taking derivatives! If we set $W = \theta^T$ and then compute $\nabla_x Wx$, then the derivative w.r.t. x would be W. On the other hand, $x^T W$ has a derivative of W^T!

Example 10. *Let's now consider a more difficult setting (often seen in deep learning). Let $X \in \mathbb{R}^{b\times d}$ and $W \in \mathbb{R}^{d\times m}$. In this context, b is the number of examples you have in a "batch" of data and d is the dimension of each example. The forward propagation rule is:*

$$Y = XW \in \mathbb{R}^{b\times m}$$

What is the derivative of the output with respect to the input?

We can't naively use the Jacobian, it's defined for $f : \mathbb{R}^n \to \mathbb{R}^m$, and in this example, $f(X) = XW$ goes from $\mathbb{R}^{b\times d} \to \mathbb{R}^{b\times m}$. However, we can use it when looking a ***how a single example within the batch changes****.*

For example, if we select the first element of X (i.e. $X[0]$ is the code that will select the first row from matrix X), we get $X \in \mathbb{R}^{1\times d}$. From Example 8, $\frac{\partial}{\partial X}(XW) = W^T$.

This holds for any row we select from X. For example, if we select the third example, i.e., say $X = X[2]$, then $\frac{\partial}{\partial X}(XW) = W^T$.

In reality, the Jacobian can be generalized for $\frac{\partial XW}{\partial X}$ and it will be a multi-dimensional matrix (or tensor) of size $\mathbb{R}^{b\times m\times b\times d}$ composed of the terms $\frac{\partial Y_{i,j}}{\partial X_{p,q}}$ where we define $Y_{i,j}$ as below:

$$Y_{i,j} = \sum_{k=1}^{d} X_{i,k}W_{k,j}$$

Notice there'll be a lot of 0s when we differentiate with respect to $X_{p,q}$ because when $p \neq i$, $Y_{i,j}$ is not a function of $X_{p,q}$. However, when $\mathbf{p = i}$,

$$Y_{i,j} = X_{i,1}W_{1,j} + ... + X_{i,q}W_{q,j} + X_{i,d}W_{d,j}$$

so

$$\begin{aligned}\frac{\partial Y_{i,j}}{\partial X_{p,q}} &= \frac{\partial Y_{i,j}}{\partial X_{i,q}} \\ &= W_{q,j}\end{aligned}$$

Nice! So we've now solved all terms in the full $b \times m \times b \times d$ Jacobian tensor! For $\frac{\partial Y_{i,j}}{\partial X_{p,q}}$, when $p \neq i$, this term is 0. Otherwise, it is $W_{q,j}$! This is exactly what our row-by-row analysis will give us. Differentiating the appropriate row of Y with respect to the appropriate row of X will result in a derivative of W^T!

In short, the Jacobian looks at how each entry of the output Y changes with respect to each entry of the input X. We often simplify this tensor in $\mathbb{R}^{b\times m\times b\times d}$ to a 2D matrix of shape $\mathbb{R}^{(bm)\times(bd)}$ that is easier to visualize. We showed earlier that the derivative with respect to any one row of X is simply W^T, so the full

Jacobian will simply be:

$$
\begin{aligned}
\nabla_X Y &= I_b \bigotimes W^T \\
&= \begin{bmatrix} 1W^T & 0W^T & \dots & 0W^T \\ \vdots & & & \\ 0W^T & 0W^T & \dots & 1W^T \end{bmatrix} \\
&= \begin{bmatrix} & \begin{bmatrix} 0 & \dots 0 \\ \vdots & \\ 0 & \dots 0 \end{bmatrix} & \dots & \begin{bmatrix} 0 & \dots 0 \\ \vdots & \\ 0 & \dots 0 \end{bmatrix} \\ 1W^T & & & \\ & \vdots & & \\ \begin{bmatrix} 0 & \dots 0 \\ \vdots & \\ 0 & \dots 0 \end{bmatrix} & \begin{bmatrix} 0 & \dots 0 \\ \vdots & \\ 0 & \dots 0 \end{bmatrix} & \dots & 1W^T \end{bmatrix} \\
&\in \mathbb{R}^{(bm)\times(bd)}
\end{aligned}
$$

where $\bigotimes$ is the Kronecker product. Even this is quite verbose to write out, so we often abbreviate $\frac{\partial XW}{\partial X}$ simply as W^T since this contains all the information we need to piece-together the full [generalized] Jacobian matrix. Saying $\frac{\partial (XW)}{\partial X} = W^T$ says that [clearly] many entries in the full Jacobian will be 0, but for each individual row of X (i.e. each example in your batch), the corresponding output row in Y has derivative W^T.