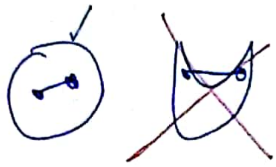
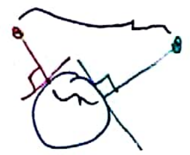
ML Theory Lecture 11

We've started on optimization segment;
today is <u>Convexity II: conjugacy & duality.</u>

First let's review key concepts from last time

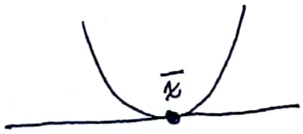
Dual view

Convex set   Supporting hyperplanes (when closed)

projection onto convex sets  (projection directions!)

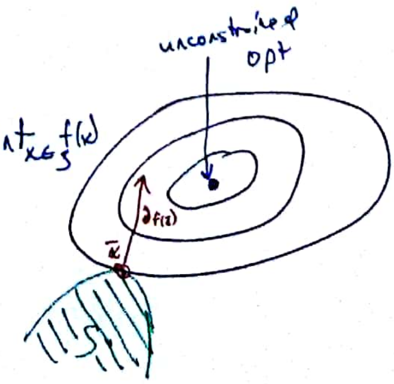
Convex function  (today's class!)

Subdifferential  (today)

First-order conditions  (today)

(unconstrained)
$0 \in \partial f(\bar{x}) \iff f(\bar{x}) = \inf_x f(x)$

(constrained)
$0 \in \partial f(\bar{x}) + N_S(\bar{x}) \iff$ $\bar{x} \in S$ & $f(\bar{x}) = \inf_{x \in S} f(x)$ (today)



normal cone
$\{ g \in \mathbb{R}^d : \forall y \in S \,.\, \langle g, y - \bar{x} \rangle \leq 0 \}$

Key types of convex functions (and their guarantees

Throughout, let $s_1 \in \partial f(x_1)$, $s_2' \in \partial f(x_2')$, $\alpha \in [0,1]$

	"direct"/"0^{th} order" view	First-order view	increasing slopes / Subdifferential view	Hessian / curvature view
Convex	$f(\alpha x + (1-\alpha) x') \leq \alpha f(x) + (1-\alpha) f(x')$	$f(x') \geq f(x) + \langle s, x'-x \rangle$	$\langle s_2 - s_1, x_2 - x_1 \rangle \geq 0$	$\nabla^2 f \succeq 0$
Strictly convex	$f(\alpha x + (1-\alpha) x') < \alpha f(x) + (1-\alpha) f(x')$ $\left(\begin{smallmatrix} x \neq x' \\ \alpha \in (0,1) \end{smallmatrix}\right)$	$f(x') > f(x) + \langle s, x'-x \rangle$ $(x' \neq x)$	$\langle s_2 - s_1, x_2 - x_1 \rangle > 0$ $(x_2 \neq x_1)$	$\nabla^2 f \succ 0$
λ-Strongly convex	$f(\alpha x + (1-\alpha) x') \leq \alpha f(x) + (1-\alpha) f(x') - \frac{\lambda \alpha (1-\alpha)}{2} \|x - x'\|^2$	$f(x') \geq f(x) + \langle s, x'-x \rangle + \frac{\lambda}{2} \|x'-x\|^2$	$\langle s_2 - s_1, x_2 - x_1 \rangle \geq \lambda \|x_2 - x_1\|^2$ (oops, should have used (x, x'), (s, s').)	$\nabla^2 f \succeq \lambda \cdot I$

Also: f has Lipschitz gradients (is strongly smooth)

when $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$.

Rem. By FTC, this implies $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|x - y\|^2$

Examples: ReLU $r \mapsto \max\{0, r\}$ is convex.

exponential $r \mapsto e^r$ is strictly convex.

If $Q \succ 0$, then $x \mapsto \frac{1}{2} x^T Q x$ is strongly convex $(\lambda = \lambda_{min}(Q))$,
and strongly smooth $(\beta = \lambda_{max}(Q))$.

$x \mapsto \|x\| + \frac{1}{2}\|x\|^2$ is strongly convex, but not differentiable, let alone smooth.

Fenchel Conjugate

First let's develop algebraic / symbolic view.

Motivation: Since ~~0 ∈~~ $\nabla f(\bar{x})$ iff $\bar{x}$ optimal,
$\nabla f^{-1}(0)$ is a useful calculation.

Informal calculation:

$$x = (\nabla f)^{-1}(s) \overset{**}{\iff} s = \nabla f(x) \overset{**}{\iff} 0 = s - \nabla f(x)$$
$$\overset{**}{\iff} x = \arg\max_{y} \langle s, y \rangle - f(y)$$

** **

This defines the Fenchel conjugate, and momentarily we'll point out it lets us invert gradients. But first:

Definition. $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed when epigraph epi (f) is closed.

Example: $s \mapsto \begin{cases} s\ln s - s & s > 0 \\ 0 & s = 0 \\ \infty & s < 0 \end{cases}$ ← convex closed | convex not closed → $s \mapsto \begin{cases} s\ln s - s & s > 0 \\ 1 & s = 0 \\ \infty & s < 0 \end{cases}$

Definition Given $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, define Fenchel conjugate

$$\boxed{f^*(s) = \sup_{x \in \mathbb{R}^d} \langle s, x \rangle - f(x)}$$

Key properties

Theorem Suppose f is closed & convex, $s \in \partial f(x)$: [ROC 23.5]

* $x \in \partial f^*(s)$ (our desired gradient inversion)
* $f(x) = f^{**}(x) = \sup_{s'} \langle s', x \rangle - f^*(s') = \langle s, x \rangle - f^*(s)$
  $\left(s = \operatorname{argmax}_{s'} \langle s', x \rangle - f^*(s') \right)$

If f is not closed (but still convex):

* $f(x) + f^*(s) = \langle s, x \rangle$ [equality case, (upcoming) Fenchel-Young]
* $f^*(s) = \sup_y \langle s, y \rangle - f(y) = \langle s, x \rangle - f(x)$
  $\left(x = \operatorname{argmax}_y \langle s, y \rangle - f(y) \right)$

Easy proof:
$s \in \partial f(x)$
$\Rightarrow \forall y.\ f(y) \geq f(x) + \langle s, y - x \rangle$
which rearranges to
$\langle s, x \rangle - f(x) \geq \sup_y \langle s, y \rangle - f(y) = f^*(s)$

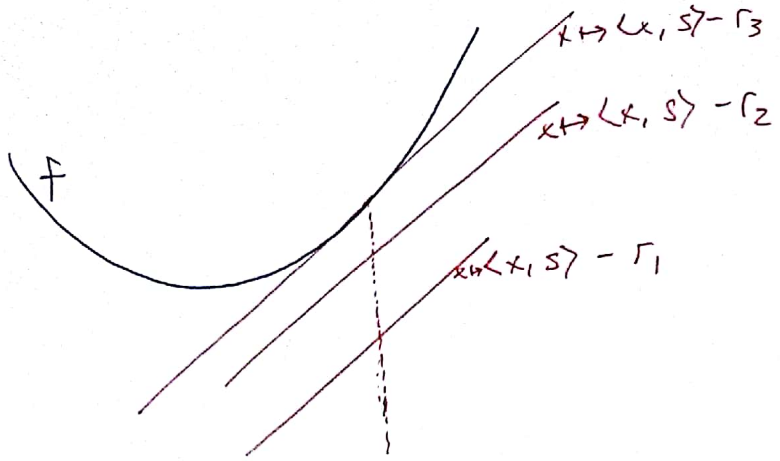
* f^* is closed convex

If f is not even convex:

* f^* is convex [because sup of affine]
* $f \geq h$ implies $f^* \leq h^*$ [By defn.]
* $f(x) + f^*(s') \geq \langle x, s' \rangle$ [Fenchel-Young inequality.]

Geometric view of Fenchel conjugate.

We would like to characterize epi(f) by its supporting hyperplanes.
In particular, given a slope s, we want to know how far up to move it to support epi(f).



So we want to consider r such that $\forall x \; f(x) \geq \langle x, s\rangle - r$,
or equivalently $\forall x \; r \geq \langle x, s\rangle - f(x)$.

But we want it to touch, so indeed we want not just $r \geq \sup_x \langle x, s\rangle - f(x)$
but in fact $r \boxed{=} \sup_x \langle x, s\rangle - f(x)$.

But this is exactly the Fenchel conjugate, and also gives meaning to the equality case
$s \in \partial f(x) \Rightarrow$ ~~[illegible]~~ $f^*(s) = \langle s, x\rangle - f(x)$.

Remark, We mentioned $f^{**} = f$ when f is closed, and f^* (and thus f^{**}) are always convex.

When f is not convex the above geometric picture makes it clear that f^{**}, a closed convex function, satisfies
$\mathrm{cl}(\mathrm{conv}(\mathrm{epi}(f))) = \mathrm{epi}(f^{**})$; f^{**} is the closed convex hull of f!

Fenchel-Rockafellar Duality Theorem

Earlier we studied optimality conditions for $x \mapsto f(x) + \iota_S(x)$.

Let's generalize this to $x \mapsto f(x) + g(Ax)$.

What does the Lagrangian give us?

$$\inf_x f(x) + g(Ax)$$

$$= \inf_{\substack{x, y \\ y = Ax}} f(x) + g(y)$$

$$= \inf_{x,y} \sup_s f(x) + g(y) + \langle s, y - Ax \rangle$$

Primal

Dual

(heart of) Weak duality

$$\boxed{\geq} \sup_s \inf_{x,y} \left(f(x) - \langle s, Ax \rangle \right) + \left(g(y) - \langle -s, y \rangle \right)$$

$$= \sup_s - \sup_x \left(\langle s, Ax \rangle - f(x) \right) - \sup_y \left(\langle -s, y \rangle - g(y) \right)$$

$$= \sup_s - f^*(A^T s) - g^*(-s).$$

<u>Remark.</u> Note an easier derivation: Fenchel-Young grants

$\forall x, s.\quad f(x) + f^*(A^T s) + g(Ax) + g^*(-s) \geq \langle x, A^T s \rangle + \langle Ax, -s \rangle = 0.$

$\Rightarrow \forall x, s.\quad f(x) + g(Ax) \geq -f^*(A^T s) - g^*(-s)$

$\Rightarrow \inf_x f(x) + g(Ax) \geq \sup_s -f^*(A^T s) - g^*(-s),$

but some key intuition lost. [Note no convexity used]

Theorem (Fenchel–Rockafellar Duality).

Let convex $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, convex $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, matrix $A \in \mathbb{R}^{n \times m}$ given.

Suppose the Slater condition (a constraint qualification (CQ))

$$\mathrm{ri}(\mathrm{dom}(g)) \cap A\,\mathrm{ri}(\mathrm{dom}(f)) \neq \phi.$$

Then

$$\inf_x f(x) + g(Ax) = \max_s -f^*(A^T s) - g^*(-s).$$

Moreover, a pair $(\bar{x}, \bar{s})$ is optimal iff $A^T \bar{s} \in \partial f(\bar{x})$ and $-\bar{s} \in \partial g(A\bar{x})$.

Remarks

* Proof technique. Consider the perturbation function $h(u) = \inf_x f(x) + g(Ax + u)$, which we can motivate from the Lagrangian. (h is convex!) Slater implies $\exists -\bar{s} \in \partial h(0)$ and some algebra gives the proof.
  * A nice & powerful proof technique (from Rockafellar).
  * Has geometric interpretation with $\bar{s}$ a supporting hyperplane.

* Duality gap. Given any x, s,

$$f(x) + g(Ax) + f^*(A^T s) + g^*(-s) \geq f(x) + g(Ax) - \inf_y \left(f(y) + g(Ay)\right).$$

E.g.,

$$f(x) + g(Ax) + f^*(A^T \nabla f(x)) + g^*(-\nabla f(x)) \geq f(x) + g(Ax) - \inf_y \left(f(y) + g(Ay)\right).$$

* Asymmetry: dual optimum is attained.
  But with a dual CQ, get primal attainment.

Example (Full Rockafeller - Pshenichnii).

Let's again consider $f + \iota_C$, & compare what we get with old first-order conditions with Fenchel duality.

$\boxed{\text{OLD:}}$

Suppose $\operatorname{ri}(C) \cap \operatorname{ri}(\operatorname{dom}(f)) \neq \emptyset$ (beefed up from before) [Slater CQ].

Then for any $\bar{x} \in C$,

$$f(\bar{x}) = \inf_{x \in C} f(x) \iff \partial f(\bar{x}) \cap -N_C(\bar{x}) \neq \emptyset$$

$N_C(\bar{x}) = \partial \iota_C(\bar{x})$
$= \{ s : \forall y \in C,\ \langle s, y - x \rangle \leq 0 \}$

$\boxed{\text{NEW:}}$ (And let's include A; use $A = I$ to compare.)

Suppose $\operatorname{ri}(C) \cap A \operatorname{ri}(\operatorname{dom}(f)) \neq \emptyset$ [Slater CQ].

Then

$$\inf_x f(x) + \iota_C(Ax) = \max_s - f^*(A^T s) - \cancel{} \sup_{x \in C} \langle s, x \rangle ;$$

Moreover $(\bar{x}, \bar{s})$ optimal iff

$$A^T \bar{s} \in \partial f(\bar{x}), \qquad \bar{s} \in -N_C(A\bar{x}).$$

$\boxed{\text{Thus}}$ we have given a name and meaning to ~~the~~ a common element of $\partial f(\bar{x}) \cap -N_C(\bar{x})$.

$\boxed{\text{What does } A \text{ buy us?}}$

Suppose $C := \{ v \in \mathbb{R}^m : v \leq b \}$.

Then $\iota_C(Ax) = \begin{cases} 0 & \text{when } Ax \leq b \text{ (a } \boxed{\text{Polyhedron}}\text{),} \\ \infty & \text{o.w..} \end{cases}$

$A = \begin{bmatrix} - a_1 - \\ \vdots \\ - a_m - \end{bmatrix}$

because $\{ x : Ax \leq b \}$
$= \bigcap_{i=1}^m \{ x : a_i^T x \leq b_i \}$

A is often very convenient (we'll see more examples).