Lecture 8 Machine Learning & Optimization

Today: optimality conditions for smooth unconstrained and constrained convex optim problems (Fermat's condition)

- Ex: minimizing l_2-regularized logsumexp
  projection onto convex sets

Types of Smooth Convex Optim Probs

Let f be a differentiable convex function, then

$$\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$$

Unconstrained

$$(U) \quad \min_{x \in \mathbb{R}^d} f(x) \qquad \text{requires } \operatorname{dom}(f) = \mathbb{R}^d$$

Constrained

$$(C) \quad \min_{x \in C} f(x) \qquad \text{requires } C \subseteq \operatorname{dom}(f)$$

As a caveat, (U) and (C) in general have multiple minimizers (if any), but the minimal value is unique (if it exists)

Optimality Conditions

We want to know when x^* is a minimizer of (U) or (C), because:

1) If it is a simple enough condition, we can directly use them to find x^*

2) We can use to :
   i) say a priori qualities x^* has, like sparsity
   ii) verify the quality of approximate solutions
   iii) design algorithms for solving (U) and (C)
   iv) analyze " " " " " "

Optimality Condition for (U)

$$x^* \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} f(x) \iff \nabla f(x^*) = 0$$

Prf

$\Leftarrow$ ($\nabla f(x^*) = 0$ then x^* is a minimizer of f)

Notice that $\forall y \in \mathbb{R}^d$:

$$f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle$$
$$= f(x^*)$$

so x^* is a minimizer.

$\underline{\Rightarrow}$ (x^* is a minimizer $\Rightarrow \nabla f(x^*) = 0$)

It suffices to show that if $\nabla f(x^*) \neq 0$, then x^* is not a minimizer.

Assume $\nabla f(x^*) \neq 0$. Then take $y = x^* - \varepsilon \dfrac{\nabla f(x^*)}{\|\nabla f(x^*)\|_2}$ for an $\varepsilon \in (0,1)$.

Then

$$\begin{aligned} f(y) &= f(x^*) + \langle \nabla f(x^*), y - x \rangle + o(\|y - x\|_2) \\ &= f(x^*) - \varepsilon \frac{\|\nabla f(x^*)\|_2^2}{\|\nabla f(x^*)\|_2} + o(\varepsilon) \\ &= f(x^*) - \varepsilon \|\nabla f(x^*)\|_2 + o(\varepsilon) \end{aligned}$$

Recall $o(\varepsilon)$ denotes a function $r(\varepsilon)$ that satisfies

$$\lim_{\varepsilon \to 0} \frac{r(\varepsilon)}{\varepsilon} = 0 \quad \Rightarrow \quad \lim_{\varepsilon \to 0} \frac{r(\varepsilon)}{\|\nabla f(x^*)\|_2 \varepsilon} = 0$$

$\Rightarrow$ there exist a ε_c s.t. if $\varepsilon < \varepsilon_c$

$$r(\varepsilon) < \frac{1}{2} \|\nabla f(x^*)\|_2 \varepsilon$$

This implies that if ε is chosen sufficiently small

$$\begin{aligned} f(y) &= f(x^*) - \varepsilon \|\nabla f(x^*)\|_2 + o(\varepsilon) \\ &< f(x^*) - \frac{1}{2}\varepsilon \|\nabla f(x^*)\|_2 \\ &< f(x^*) \end{aligned}$$

so x^* is not a minimizer.

__Ex__

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^d} \|x\|_2^2 + \operatorname{logsumexp}(x) =: f(x)$$

$$\nabla f(x^*) = 2x^* + \operatorname{softmax}(x^*) = 0$$

Observations:

- $\operatorname{softmax}(x^*) = -2x^*$ is a probability vector
- f is permutation-invariant: any permutation of x^* is __also__ a minimizer
- f is strongly convex so x^* is unique

$\Rightarrow$ x^* is a constant vector, $x^* = c\mathbb{1}$ for some $c < 0$

Because $x^* = c \cdot \mathbb{1}$ and

$$\text{softmax}(x^*) = -2x^* = -2c \cdot \mathbb{1}$$

also, because softmax(x^*) is a probability vector that is constant,

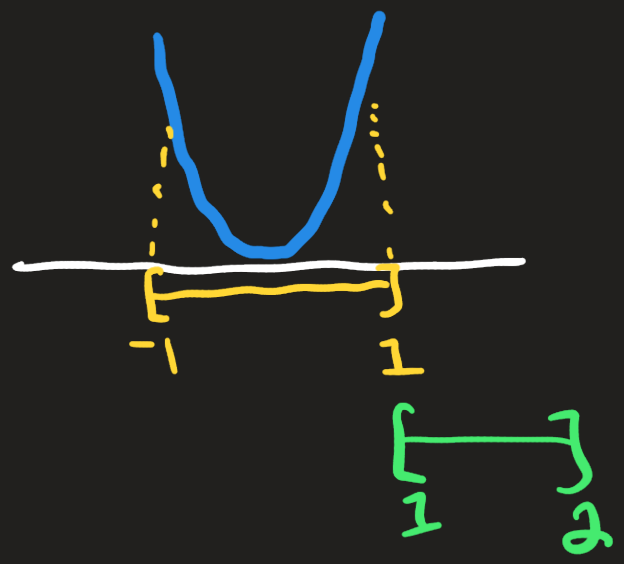
$$\text{softmax}(x^*) = \frac{1}{d} \cdot \mathbb{1}$$

so $c = -\frac{1}{2d}$ and therefore

$$x^* = \begin{bmatrix} -\frac{1}{2d} \\ \vdots \\ -\frac{1}{2d} \end{bmatrix}$$

Optimality Conditions for (C)

First, note that $\nabla f(x^*)$ is <u>not</u> a general optimality condition for (C)



e.g. $\underset{x\in[1,2]}{\operatorname{argmin}}\ x^2 = 1$, but $\left.\frac{d(x^2)}{dx}\right|_{x=1} = 2 \neq 0$

However

$\underset{x\in[-1,1]}{\operatorname{argmin}}\ x^2 = 0$ and $\left.\frac{d(x^2)}{dx}\right|_{x=0} = 0$

Takeaway: the optimality conditions depends on whether the minimizers occur on the boundary or interior of the constraint set C

Optimality Condition for (C)

$$x^* \in \underset{x \in C}{\operatorname{argmin}}\ f(x)$$

iff for all $y \in C$, $\langle \nabla f(x^*), y - x^* \rangle \geq 0$

Recall ~~that~~ b/c f is convex,

$$\forall y \in C : \ f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle$$

Prf

$\Leftarrow$ (for all $y \in C$, $\langle \nabla f(x^*), y - x^* \rangle \geq 0 \Rightarrow x^*$ is a minimizer of f)

By the first-order characterization,

$$\forall y \in C : \ f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \geq f(x^*)$$

so x^* is a minimizer of f

$\Rightarrow$ (x^* is a minimizer $\Rightarrow \forall y \in C: \langle \nabla f(x^*), y - x^* \rangle \geq 0$)

It suffices to show that if there exists $y \in C$ s.t.

$$\langle \nabla f(x^*), y - x^* \rangle < 0$$

then x is not a minimizer.

Let $z = y - x^*$ and take $\bar{y} = x^* + \varepsilon z = x^* + \varepsilon(y - x^*)$
$= (1-\varepsilon)x^* + \varepsilon y$

Notice $\bar{y} \in [x, y]$ so it is in C.

Also

$$f(\bar{y}) = f(x^*) + \langle \nabla f(x^*), \bar{y} - x^* \rangle + o(\|\bar{y} - x^*\|_2)$$

$$= f(x^*) + \underbrace{\langle \nabla f(x^*), \varepsilon z \rangle}_{\text{this is negative}} + o(\varepsilon \|z\|_2)$$

$$= f(x^*) - \varepsilon |\langle \nabla f(x^*), y - x^* \rangle| + o(\varepsilon)$$

As before, there exist ε_c s.t. when $\varepsilon < \varepsilon_c$, the $o(\varepsilon)$ remainder term is no larger than $\frac{1}{2}\varepsilon |\langle \nabla f(x^*), y - x^* \rangle|$, so

$$f(\bar{y}) \leq f(x^*) - \frac{1}{2}\varepsilon |\langle \nabla f(x^*), y - x^* \rangle|$$
$$< f(x^*)$$

Ex Projection onto Convex Sets

Let C be a convex set and x be a point

$$P_C(x) = \underset{z \in C}{\operatorname{argmin}} \|x - z\|_2^2$$

Facts

1) $P_C(x)$ is unique because $z \mapsto \|x-z\|_2^2$ is strongly convex

2) $P_C(x)$ is the unique point in C satisfying

$$\forall y \in C: \langle P_C(x) - x, y - P_C(x) \rangle \geq 0$$

We will show next class that if

$$C = B_{\|\cdot\|_2}(0,1) = \{x \mid \|x\|_2 \leq 1\}$$

then

$$P_C(x) = \begin{cases} x & \text{if } \|x\|_2 \leq 1 \\ \dfrac{x}{\|x\|_2} & \text{if } \|x\|_2 > 1 \end{cases}$$