Optimization for Machine Learning

Lecture 4: Optimality conditions

6.881: MIT

Suvrit Sra
Massachusetts Institute of Technology

25 Feb, 2021

Optimality

(Local and global optima)

Optimality

Def. A point $x^* \in \mathcal{X}$ is ***locally optimal*** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. ***Global*** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Optimality

Def. A point $x^* \in \mathcal{X}$ is ***locally optimal*** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. ***Global*** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

Optimality

Def. A point $x^* \in \mathcal{X}$ is ***locally optimal*** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. ***Global*** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

- Let x^* be local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.

Optimality

Def. A point $x^* \in \mathcal{X}$ is *locally optimal* if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. *Global* if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

- Let x^* be local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.
- We need to show that $f(y) \geq f(x^*) = p^*$.

Optimality

Def. A point $x^* \in \mathcal{X}$ is ***locally optimal*** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. ***Global*** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

- Let x^* be local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.
- We need to show that $\boxed{f(y) \geq f(x^*) = p^*}$.
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$

Optimality

Def. A point $x^* \in \mathcal{X}$ is *locally optimal* if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. *Global* if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

- Let x^* be local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.
- We need to show that $f(y) \geq f(x^*) = p^*$.
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is cvx, and $x^*, y \in \operatorname{dom} f$, we have

$$\begin{aligned} f(x_\theta) &\leq \theta f(y) + (1-\theta)f(x^*) \\ f(x_\theta) - f(x^*) &\leq \theta(f(y) - f(x^*)). \end{aligned}$$

Optimality

Def. A point $x^* \in \mathcal{X}$ is ***locally optimal*** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. *Global* if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

- Let x^* be local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.
- We need to show that $f(y) \geq f(x^*) = p^*$.
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is cvx, and $x^*, y \in \operatorname{dom} f$, we have

$$\begin{aligned} f(x_\theta) &\leq \theta f(y) + (1-\theta) f(x^*) \\ f(x_\theta) - f(x^*) &\leq \theta(f(y) - f(x^*)). \end{aligned}$$

- Since x^* is a local minimizer, for small enough $\theta > 0$, lhs ≥ 0.

Optimality

Def. A point $x^* \in \mathcal{X}$ is *locally optimal* if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. *Global* if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex f, locally optimal point also global.

- Let x^* be local minimizer of f on $\mathcal{X}$; $y \in \mathcal{X}$ any other feasible point.
- We need to show that $\boxed{f(y) \geq f(x^*) = p^*}$.
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is cvx, and $x^*, y \in \operatorname{dom} f$, we have

$$\begin{aligned} f(x_\theta) &\leq \theta f(y) + (1-\theta)f(x^*) \\ f(x_\theta) - f(x^*) &\leq \theta(f(y) - f(x^*)). \end{aligned}$$

- Since x^* is a local minimizer, for small enough $\theta > 0$, lhs ≥ 0.
- So rhs is also nonnegative, proving $f(y) \geq f(x^*)$ as desired.

Set of Optimal Solutions

The set of optimal solutions $\mathcal{X}^*$ may be empty

Example. If $\mathcal{X} = \emptyset$, i.e., no feasible solutions, then $\mathcal{X}^* = \emptyset$

Example. When only inf not min, e.g., $\inf e^x$ as $x \to -\infty$
in general, we should worry about the question "Is $\mathcal{X}^* = \emptyset$?"

Exercise: Verify that $\mathcal{X}^*$ is always a convex set.

Optimality conditions

(Recognizing optima)

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$.
Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \quad \frac{f(x^* + td) - f(x^*)}{}$$

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \quad \frac{f(x^* + td) - f(x^*)}{t}$$

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t}$$

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt}$$

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d \rangle.$$

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d \rangle.$$

Similarly, using $-d$ it follows that $\langle \nabla f(x^*), d \rangle \leq 0$, so $\langle \nabla f(x^*), d \rangle = 0$ **must hold**. Since d is arbitrary, $\nabla f(x^*) = 0$.

First-order conditions: unconstrained

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set S containing x^*, a local minimum. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d \rangle.$$

Similarly, using $-d$ it follows that $\langle \nabla f(x^*), d \rangle \leq 0$, so $\langle \nabla f(x^*), d \rangle = 0$ **must hold**. Since d is arbitrary, $\nabla f(x^*) = 0$.

Exercise: Prove that if f is convex, then $\nabla f(x^*) = 0$ is actually **sufficient** for global optimality! For general f this is ***not*** true. (This property is what makes convex optimization special!)

First-order conditions: constrained

♠ For convex f, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

First-order conditions: constrained

- For convex f, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
- Thus, x^* is optimal if and only if

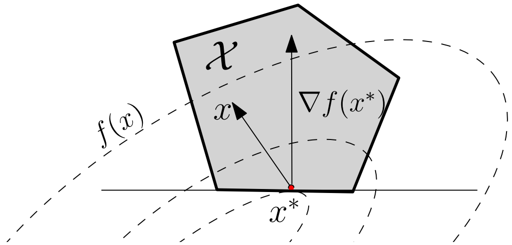
$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \qquad \text{for all } y \in \mathcal{X}.$$

First-order conditions: constrained

- For convex f, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
- Thus, x^* is optimal if and only if

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \qquad \text{for all } y \in \mathcal{X}.$$

- If $\mathcal{X} = \mathbb{R}^n$, this reduces to $\nabla f(x^*) = 0$



- If $\nabla f(x^*) \neq 0$, it defines supporting hyperplane to $\mathcal{X}$ at x^*

First-order conditions: constrained

- Let f be continuously differentiable, possibly nonconvex
- Suppose $\exists y \in \mathcal{X}$ such that $\langle \nabla f(x^*), y - x^* \rangle < 0$
- Using mean-value theorem of calculus, $\exists \xi \in [0, 1]$ s.t.

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

(we applied MVT to $g(t) := f(x^* + t(y - x^*))$)

- For sufficiently small t, since ∇f continuous, by assump on y, $\langle \nabla f(x^* + \xi t(y - x^*)), y - x^* \rangle < 0$
- This in turn implies that $f(x^* + t(y - x^*)) < f(x^*)$
- Since $\mathcal{X}$ is convex, $x^* + t(y - x^*) \in \mathcal{X}$ is also feasible
- Contradiction to local optimality of x^*

Optimality without differentiability

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{Argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}\,.$$

Optimality without differentiability

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{Argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{Argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.

Optimality without differentiability

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{Argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{Argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle \quad \forall y$,

Optimality without differentiability

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{Argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}\,.$$

Proof: $x \in \operatorname{Argmin} f$ implies that $f(x) \le f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \ge f(x) + \langle 0, y - x\rangle \quad \forall y\,, \Leftrightarrow 0 \in \partial f(x)$.

Optimality without differentiability

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{Argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{Argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle \quad \forall y, \Leftrightarrow 0 \in \partial f(x)$.

Nonsmooth problem

$$\min_x \quad f(x) \quad \text{s.t.} \quad x \in \mathcal{X}$$

$$\min_x \quad f(x) + \mathbb{1}_{\mathcal{X}}(x).$$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x\rangle$ for all y.

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle \ \forall y \in \mathcal{X}$.

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x\rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x\rangle\ \forall y \in \mathcal{X}$.
- ***Subdifferential of the indicator*** $\mathbb{1}_{\mathcal{X}}(x)$, aka **normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x\rangle \quad \forall y \in \mathcal{X}\}$$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x\rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x\rangle$ $\forall y \in \mathcal{X}$.
- *Subdifferential of the indicator* $\mathbb{1}_{\mathcal{X}}(x)$, aka **normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x\rangle \quad \forall y \in \mathcal{X}\}$$

Application

$$\boxed{\min\ f(x) + \mathbb{1}_{\mathcal{X}}(x).}$$

◇ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle\ \forall y \in \mathcal{X}$.
- ***Subdifferential of the indicator*** $\mathbb{1}_{\mathcal{X}}(x)$, aka **normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X}\}$$

Application

$$\boxed{\min\ f(x) + \mathbb{1}_{\mathcal{X}}(x).}$$

◇ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

◇ $-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*) \iff \langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in \mathcal{X}$.

Example

$$\boxed{\min \quad f(x) \qquad \|x\| \leq 1.}$$

Example

$$\min \quad f(x) \qquad \|x\| \leq 1.$$

A point x is optimal if and only if

$x \in \operatorname{dom} f, \quad \|x\| \leq 1,$

Example

$$\boxed{\min \quad f(x) \qquad \|x\| \leq 1.}$$

A point x is optimal if and only if

$$x \in \operatorname{dom} f, \quad \|x\| \leq 1, \forall y \text{ s.t. } \|y\| \leq 1 \implies \nabla f(x)^T(y - x) \geq 0.$$

Example

$$\boxed{\min \quad f(x) \qquad \|x\| \leq 1.}$$

A point x is optimal if and only if

$$x \in \operatorname{dom} f, \quad \|x\| \leq 1, \forall y \text{ s.t. } \|y\| \leq 1 \implies \nabla f(x)^T(y-x) \geq 0.$$

In other words

$$\begin{aligned} \forall \|y\| \leq 1, \quad \nabla f(x)^T y &\geq \nabla f(x)^T x \\ \forall \|y\| \leq 1, \quad -\nabla f(x)^T y &\leq -\nabla f(x)^T x \end{aligned}$$

Example

$$\boxed{\min \quad f(x) \qquad \|x\| \leq 1.}$$

A point x is optimal if and only if

$$x \in \operatorname{dom} f, \quad \|x\| \leq 1, \forall y \text{ s.t. } \|y\| \leq 1 \implies \nabla f(x)^T(y-x) \geq 0.$$

In other words

$$\begin{aligned}
\forall \|y\| \leq 1, \quad \nabla f(x)^T y &\geq \nabla f(x)^T x \\
\forall \|y\| \leq 1, \quad -\nabla f(x)^T y &\leq -\nabla f(x)^T x \\
\sup\{-\nabla f(x)^T y \mid \|y\| \leq 1\} &\leq -\nabla f(x)^T x
\end{aligned}$$

Example

$$\min \quad f(x) \qquad \|x\| \leq 1.$$

A point x is optimal if and only if

$$x \in \operatorname{dom} f, \quad \|x\| \leq 1, \forall y \text{ s.t. } \|y\| \leq 1 \implies \nabla f(x)^T(y-x) \geq 0.$$

In other words

$$\begin{aligned}
\forall \|y\| \leq 1, \quad \nabla f(x)^T y &\geq \nabla f(x)^T x \\
\forall \|y\| \leq 1, \quad -\nabla f(x)^T y &\leq -\nabla f(x)^T x \\
\sup\{-\nabla f(x)^T y \mid \|y\| \leq 1\} &\leq -\nabla f(x)^T x \\
\|-\nabla f(x)\|_* &\leq -\nabla f(x)^T x \\
\|\nabla f(x)\|_* &\leq -\nabla f(x)^T x.
\end{aligned}$$

Example

$$\boxed{\min \quad f(x) \qquad \|x\| \leq 1.}$$

A point x is optimal if and only if

$$x \in \operatorname{dom} f, \quad \|x\| \leq 1, \forall y \text{ s.t. } \|y\| \leq 1 \implies \nabla f(x)^T (y - x) \geq 0.$$

In other words

$$\begin{aligned}
\forall \|y\| \leq 1, \quad \nabla f(x)^T y &\geq \nabla f(x)^T x \\
\forall \|y\| \leq 1, \quad -\nabla f(x)^T y &\leq -\nabla f(x)^T x \\
\sup\{-\nabla f(x)^T y \mid \|y\| \leq 1\} &\leq -\nabla f(x)^T x \\
\|-\nabla f(x)\|_* &\leq -\nabla f(x)^T x \\
\|\nabla f(x)\|_* &\leq -\nabla f(x)^T x.
\end{aligned}$$

Observe: If constraint satisfied strictly at optimum ($\|x\| < 1$), then $\nabla f(x) = 0$ (else we'd violate the last inequality above).

Optimality conditions

(KKT and friends)

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

Assume strong duality and that p^*, d^* attained!

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x) \right)$

Assume strong duality and that p^*, d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$p^* = f(x^*)$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

Assume strong duality and that p^*, d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*)$$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

$$\boxed{\text{Assume strong duality and that } p^*, d^* \text{ attained!}}$$

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*)$$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

Assume strong duality and that p^*, d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*)$$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x) \right)$

Assume strong duality and that p^*, d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \le \mathcal{L}(x^*, \lambda^*) \le f(x^*) = p^*$$

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

$$\boxed{\text{Assume strong duality and that } p^*, d^* \text{ attained!}}$$

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f(x^*) = p^*$$

- Thus, equalities hold in above chain, and

Optimality conditions via Lagrangian

$$\min \quad f(x), \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \left(\mathcal{L}(x, \lambda) := f(x) + \sum_i \lambda_i f_i(x)\right)$

$$\boxed{\text{Assume strong duality and that } p^*, d^* \text{ attained!}}$$

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f(x^*) = p^*$$

- Thus, equalities hold in above chain, and

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

Optimality conditions via Lagrangian

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

Optimality conditions via Lagrangian

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Optimality conditions via Lagrangian

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f(x^*)$, we also have

Optimality conditions via Lagrangian

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f(x^*)$, we also have

$$\sum\nolimits_i \lambda_i^* f_i(x^*) = 0.$$

Optimality conditions via Lagrangian

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f(x^*)$, we also have

$$\sum\nolimits_i \lambda_i^* f_i(x^*) = 0.$$

But $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$,

Optimality conditions via Lagrangian

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f(x^*)$, we also have

$$\sum\nolimits_i \lambda_i^* f_i(x^*) = 0.$$

But $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$, so *complementary slackness*

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m.$$

KKT conditions

Karush-Kuhn-Tucker Conditions (KKT)

$$
\begin{array}{rcll}
f_i(x^*) & \leq & 0, \quad i = 1, \ldots, m & \text{(primal feasibility)} \\
\lambda_i^* & \geq & 0, \quad i = 1, \ldots, m & \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) & = & 0, \quad i = 1, \ldots, m & \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} & = & 0 & \text{(Lagrangian stationarity)}
\end{array}
$$

KKT conditions

Karush-Kuhn-Tucker Conditions (KKT)

$$
\begin{array}{rcll}
f_i(x^*) & \leq & 0, \quad i = 1, \ldots, m & \text{(primal feasibility)} \\
\lambda_i^* & \geq & 0, \quad i = 1, \ldots, m & \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) & = & 0, \quad i = 1, \ldots, m & \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} & = & 0 & \text{(Lagrangian stationarity)}
\end{array}
$$

- Thus, if strong duality holds, and (x^*, λ^*) exists, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal

KKT conditions

Karush-Kuhn-Tucker Conditions (KKT)

$$
\begin{aligned}
f_i(x^*) &\le 0, \quad i = 1, \dots, m && \text{(primal feasibility)} \\
\lambda_i^* &\ge 0, \quad i = 1, \dots, m && \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \dots, m && \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}
$$

- Thus, if strong duality holds, and (x^*, λ^*) exists, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal
- If problem is convex, then KKT also **sufficient**

KKT conditions

Karush-Kuhn-Tucker Conditions (KKT)

$$\begin{array}{rcll} f_i(x^*) & \leq & 0, \quad i = 1, \ldots, m & \text{(primal feasibility)} \\ \lambda_i^* & \geq & 0, \quad i = 1, \ldots, m & \text{(dual feasibility)} \\ \lambda_i^* f_i(x^*) & = & 0, \quad i = 1, \ldots, m & \text{(compl. slackness)} \\ \nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} & = & 0 & \text{(Lagrangian stationarity)} \end{array}$$

- Thus, if strong duality holds, and (x^*, λ^*) exists, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal
- If problem is convex, then KKT also **sufficient**

Exercise: Prove the above sufficiency of KKT.

Hint: Use that $\mathcal{L}(x, \lambda^*)$ is convex, and conclude from KKT conditions that $g(\lambda^*) = f_0(x^*)$, so that (x^*, λ^*) optimal primal-dual pair.

Read Ch. 5 of BV

Examples

Projection onto a hyperplane

$$\min_x \ \tfrac{1}{2}\|x - y\|^2, \quad \text{s.t.} \ \ a^T x = b.$$

Projection onto a hyperplane

$$\min_x \; \tfrac{1}{2}\|x-y\|^2, \quad \text{s.t.} \;\; a^T x = b.$$

KKT Conditions

$$\begin{aligned} L(x,\nu) &= \tfrac{1}{2}\|x-y\|^2 + \nu(a^T x - b) \\ \frac{\partial L}{\partial x} &= x - y + \nu a = 0 \end{aligned}$$

Projection onto a hyperplane

$$\min_x \ \tfrac{1}{2}\|x - y\|^2, \quad \text{s.t.} \ \ a^T x = b.$$

KKT Conditions

$$\begin{aligned} L(x, \nu) &= \tfrac{1}{2}\|x - y\|^2 + \nu(a^T x - b) \\ \frac{\partial L}{\partial x} &= x - y + \nu a = 0 \\ & x = y - \nu a \end{aligned}$$

Projection onto a hyperplane

$$\min_x \ \tfrac{1}{2}\|x - y\|^2, \quad \text{s.t.} \ \ a^T x = b.$$

KKT Conditions

$$\begin{aligned}
L(x,\nu) \quad &= \tfrac{1}{2}\|x-y\|^2 + \nu(a^T x - b) \\
\frac{\partial L}{\partial x} \quad &= x - y + \nu a = 0 \\
&x = y - \nu a \\
&a^T x = a^T y - \nu a^T a \\
&\|a\|^2 \nu = a^T y - b
\end{aligned}$$

Projection onto a hyperplane

$$\min_x \ \tfrac{1}{2}\|x-y\|^2, \quad \text{s.t.} \ \ a^T x = b.$$

KKT Conditions

$$\begin{aligned} L(x,\nu) &= \tfrac{1}{2}\|x-y\|^2 + \nu(a^T x - b) \\ \frac{\partial L}{\partial x} &= x - y + \nu a = 0 \\ & x = y - \nu a \\ & a^T x = a^T y - \nu a^T a \\ & \|a\|^2 \nu = a^T y - b \end{aligned}$$

$$\boxed{x = y - \tfrac{1}{\|a\|^2}(a^T y - b)a}$$

Projection onto simplex

$$\min_x \ \tfrac{1}{2}\|x - y\|^2, \quad \text{s.t.} \quad x^T 1 = 1, x \geq 0.$$

KKT Conditions

$$L(x, \lambda, \nu) \qquad = \tfrac{1}{2}\|x - y\|^2 - \sum\nolimits_i \lambda_i x_i + \nu(x^T 1 - 1)$$

Projection onto simplex

$$\min_x \ \tfrac{1}{2}\|x - y\|^2, \quad \text{s.t.} \quad x^T 1 = 1, x \geq 0.$$

KKT Conditions

$$\begin{aligned}
L(x, \lambda, \nu) &= \tfrac{1}{2}\|x - y\|^2 - \textstyle\sum_i \lambda_i x_i + \nu(x^T 1 - 1) \\
\frac{\partial L}{\partial x_i} &= x_i - y_i - \lambda_i + \nu = 0 \\
\lambda_i x_i &= 0 \\
\lambda_i &\geq 0 \\
x^T 1 &= 1, x \geq 0
\end{aligned}$$

Projection onto simplex

$$\min_x \ \tfrac{1}{2}\|x-y\|^2, \quad \text{s.t.} \quad x^T 1 = 1, x \geq 0.$$

KKT Conditions

$$\begin{aligned}
L(x,\lambda,\nu) &= \tfrac{1}{2}\|x-y\|^2 - \textstyle\sum_i \lambda_i x_i + \nu(x^T 1 - 1)\\
\frac{\partial L}{\partial x_i} &= x_i - y_i - \lambda_i + \nu = 0\\
\lambda_i x_i &= 0\\
\lambda_i &\geq 0\\
x^T 1 &= 1, x \geq 0
\end{aligned}$$

Challenge A. Solve the above conditions in $O(n \log n)$ time.

Challenge A+. Solve the above conditions in $O(n)$ time.

Total variation minimization

$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \sum_i |x_{i+1} - x_i|,$$

$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda\|Dx\|_1,$$

(the matrix D is also known as a differencing matrix).

Total variation minimization

$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \textstyle\sum_i |x_{i+1}-x_i|,$$
$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \|Dx\|_1,$$

(the matrix D is also known as a differencing matrix).
Step 1. Take the dual (recall from L3-25) to obtain:

$$\min_u \tfrac{1}{2}\|D^T u\|^2 - u^T D y, \quad \text{s.t.} \quad \|u\|_\infty \le \lambda.$$

Step 2. Replace obj by $\|D^T u - y\|^2$ (argmin is unchanged)

Total variation minimization

$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \sum\nolimits_i |x_{i+1} - x_i|,$$
$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \|Dx\|_1,$$

(the matrix D is also known as a differencing matrix).

Step 1. Take the dual (recall from L3-25) to obtain:

$$\min_u \tfrac{1}{2}\|D^T u\|^2 - u^T D y, \quad \text{s.t.} \quad \|u\|_\infty \le \lambda.$$

Step 2. Replace obj by $\|D^T u - y\|^2$ (argmin is unchanged)

Step 3. Add dummies $u_0 = u_n = 0$; write $s = r - u$ for $r = \sum_{k=1}^{i} y_k$

$$\min_s \sum\nolimits_{i=1}^{n} (s_{i-1} - s_i)^2, \quad \text{s.t.} \quad \|s - r\|_\infty \le \lambda, s_0 = 0, s_n = r_n.$$

Total variation minimization

$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \sum\nolimits_i |x_{i+1} - x_i|,$$
$$\min \quad \tfrac{1}{2}\|x-y\|^2 + \lambda \|Dx\|_1,$$

(the matrix D is also known as a differencing matrix).

Step 1. Take the dual (recall from L3-25) to obtain:

$$\min_u \tfrac{1}{2}\|D^T u\|^2 - u^T D y, \quad \text{s.t.} \quad \|u\|_\infty \le \lambda.$$

Step 2. Replace obj by $\|D^T u - y\|^2$ (argmin is unchanged)

Step 3. Add dummies $u_0 = u_n = 0$; write $s = r - u$ for $r = \sum_{k=1}^{i} y_k$

$$\min_s \sum\nolimits_{i=1}^{n} (s_{i-1} - s_i)^2, \quad \text{s.t.} \quad \|s - r\|_\infty \le \lambda, s_0 = 0, s_n = r_n.$$

Step 4 (Challenge). Look at KKT conditions, and keep working …finally, obtain $O(n)$ method!
For full-story look at: *A. Barbero, S. Sra. "Modular proximal optimization for multidimensional total-variation regularization" (JMLR 2019, pp. 1–82)*

Nonsmooth KKT

(via subdifferentials)

KKT via subdifferentials*

Assume all $f_i(x)$ are finite valued, and $\operatorname{dom} f = \mathbb{R}^n$

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \;\; f_i(x) \leq 0, \;\; i \in [m].$$

KKT via subdifferentials*

Assume all $f_i(x)$ are finite valued, and $\mathrm{dom}\ f = \mathbb{R}^n$

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \;\; f_i(x) \le 0,\; i \in [m].$$

Assume Slater's condition: $\exists x$ such that $f_i(x) < 0$ for $i \in [m]$

KKT via subdifferentials*

Assume all $f_i(x)$ are finite valued, and $\operatorname{dom} f = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad f_i(x) \leq 0,\ i \in [m].$$

Assume Slater's condition: $\exists x$ such that $f_i(x) < 0$ for $i \in [m]$
Write $C_i := \{x \mid f_i(x) \leq 0\}$. Then, above problem becomes

$$\min_{x} \quad \phi(x) := f(x) + \mathbb{1}_{C_1}(x) + \cdots + \mathbb{1}_{C_m}(x).$$

KKT via subdifferentials*

Assume all $f_i(x)$ are finite valued, and $\mathrm{dom}\, f = \mathbb{R}^n$

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \;\; f_i(x) \le 0,\; i \in [m].$$

Assume Slater's condition: $\exists x$ such that $f_i(x) < 0$ for $i \in [m]$
Write $C_i := \{x \mid f_i(x) \le 0\}$. Then, above problem becomes

$$\min_x \quad \phi(x) := f(x) + \mathbb{1}_{C_1}(x) + \cdots + \mathbb{1}_{C_m}(x).$$

An optimal solution to this problem is a vector $\bar{x}$ such that

$$0 \in \partial\phi(\bar{x}).$$

KKT via subdifferentials*

Assume all $f_i(x)$ are finite valued, and $\operatorname{dom} f = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad f_i(x) \leq 0, \; i \in [m].$$

Assume Slater's condition: $\exists x$ such that $f_i(x) < 0$ for $i \in [m]$
Write $C_i := \{x \mid f_i(x) \leq 0\}$. Then, above problem becomes

$$\min_{x} \quad \phi(x) := f(x) + \mathbb{1}_{C_1}(x) + \cdots + \mathbb{1}_{C_m}(x).$$

An optimal solution to this problem is a vector $\bar{x}$ such that

$$0 \in \partial \phi(\bar{x}).$$

Slater's condition tells us that

$$\operatorname{int} C_1 \cap \cdots \cap \operatorname{int} C_m \neq \emptyset.$$

Exercise: Rigorously justify the above (*Hint:* use continuity of f_i)

KKT via subdifferentials*

Since $\operatorname{int} C_1 \cap \cdots \cap \operatorname{int} C_m \neq \emptyset$, Rockafellar's theorem tells us

$$\partial\phi(x) = \partial f(x) + \partial \mathbb{1}_{C_1}(x) + \cdots + \partial \mathbb{1}_{C_m}(x).$$

KKT via subdifferentials*

Since $\operatorname{int} C_1 \cap \cdots \cap \operatorname{int} C_m \neq \emptyset$, Rockafellar's theorem tells us

$$\partial\phi(x) = \partial f(x) + \partial \mathbb{1}_{C_1}(x) + \cdots + \partial \mathbb{1}_{C_m}(x).$$

Recall: $\partial \mathbb{1}_{C_i} = \mathcal{N}_{C_i}$ (normal cone). **Verify (Challenge)** that

$$\mathcal{N}_{C_i}(x) = \begin{cases} \bigcup \{\lambda_i \partial f_i(x) \mid \lambda_i \geq 0\}, & \text{if } f_i(x) = 0, \\ \{0\}, & \text{if } f_i(x) < 0, \\ \emptyset, & \text{if } f_i(x) > 0. \end{cases}$$

KKT via subdifferentials*

Since $\operatorname{int} C_1 \cap \cdots \cap \operatorname{int} C_m \neq \emptyset$, Rockafellar's theorem tells us

$$\partial\phi(x) = \partial f(x) + \partial \mathbb{1}_{C_1}(x) + \cdots + \partial \mathbb{1}_{C_m}(x).$$

Recall: $\partial \mathbb{1}_{C_i} = \mathcal{N}_{C_i}$ (normal cone). **Verify (Challenge)** that

$$\mathcal{N}_{C_i}(x) = \begin{cases} \bigcup \{\lambda_i \partial f_i(x) \mid \lambda_i \geq 0\}, & \text{if } f_i(x) = 0, \\ \{0\}, & \text{if } f_i(x) < 0, \\ \emptyset, & \text{if } f_i(x) > 0. \end{cases}$$

Thus, $\partial\phi(x) \neq \emptyset$ iff x satisfies $f_i(x) \leq 0$
(**Verify:** that the Minkowski sum $A + \emptyset = \emptyset$)

KKT via subdifferentials*

Thus, $\partial\phi(x) = \bigcup\{\partial f(x) + \lambda_1 \partial f_1(x) + \cdots + \lambda_m \partial f_m(x)\}$, *over all choices* of $\lambda_i \geq 0$ such that

$$\lambda_i f_i(x) = 0.$$

If $f_i(x) < 0$, $\partial\mathbb{1}_{C_i} = \{0\}$, while for $f_i(x) = 0$, $\partial\mathbb{1}_{C_i}(x) = \{\lambda_i \partial f_i(x) \mid \lambda_i \geq 0\}$, and we cannot jointly have $\lambda_i \geq 0$ and $f_i(x) > 0$.

KKT via subdifferentials*

Thus, $\partial\phi(x) = \bigcup\{\partial f(x) + \lambda_1 \partial f_1(x) + \cdots + \lambda_m \partial f_m(x)\}$, *over all choices* of $\lambda_i \geq 0$ such that

$$\lambda_i f_i(x) = 0.$$

If $f_i(x) < 0$, $\partial \mathbb{1}_{C_i} = \{0\}$, while for $f_i(x) = 0$, $\partial \mathbb{1}_{C_i}(x) = \{\lambda_i \partial f_i(x) \mid \lambda_i \geq 0\}$, and we cannot jointly have $\lambda_i \geq 0$ and $f_i(x) > 0$.

In other words, $0 \in \partial\phi(x)$ **iff** there exist $\lambda_1, \ldots, \lambda_m$ that satisfy the KKT conditions.

Exercise: Double check the above for differentiable f, f_i

Example: Constrained regression

$$\min_x \ \tfrac{1}{2}\|Ax - b\|^2, \quad \text{s.t. } \|x\| \leq \theta.$$

KKT Conditions

$$\begin{aligned}
L(x, \lambda) &= \tfrac{1}{2}\|Ax - b\|^2 + \lambda(\|x\| - \theta) \\
0 &\in A^T(Ax - b) + \lambda \partial\|x\| \\
\partial\|x\| &= \begin{cases} \|x\|^{-1}x & x \neq 0, \\ \{z \mid \|z\| \leq 1\} & x = 0. \end{cases}
\end{aligned}$$

Hmmm...?

Example: Constrained regression

$$\min_x \ \tfrac{1}{2}\|Ax - b\|^2, \quad \text{s.t.} \ \|x\| \le \theta.$$

KKT Conditions

$$\begin{aligned} L(x,\lambda) &= \tfrac{1}{2}\|Ax-b\|^2 + \lambda(\|x\| - \theta) \\ 0 &\in A^T(Ax-b) + \lambda\partial\|x\| \\ \partial\|x\| &= \begin{cases} \|x\|^{-1}x & x \neq 0, \\ \{z \mid \|z\| \le 1\} & x = 0. \end{cases} \end{aligned}$$

Hmmm...?

- *Case (i).* $x \leftarrow \texttt{pinv}(A)b$ and $\|x\| < \theta$, then $x^* = x$
- *Case (ii).* If $\|x\| \ge \theta$, then $\|x^*\| = \theta$. Thus, consider instead $\frac{1}{2}\|Ax-b\|^2$ s.t. $\|x\|^2 = \theta^2$. (**Exercise:** complete the idea.)