Optimization for Machine Learning

(Problems; Algorithms - A)

Suvrit Sra
Massachusetts Institute of Technology

PKU Summer School on Data Science (July 2017)

Course materials

- *http://suvrit.de/teaching.html*
- Some references:
  - *Introductory lectures on convex optimization* – Nesterov
  - *Convex optimization* – Boyd & Vandenberghe
  - *Nonlinear programming* – Bertsekas
  - *Convex Analysis* – Rockafellar
  - *Fundamentals of convex analysis* – Urruty, Lemaréchal
  - *Lectures on modern convex optimization* – Nemirovski
  - *Optimization for Machine Learning* – Sra, Nowozin, Wright
  - *Theory of Convex Optimization for Machine Learning* – Bubeck
  - *NIPS 2016 Optimization Tutorial* – Bach, Sra
- Some related courses:
  - EE227A, Spring 2013, (Sra, UC Berkeley)
  - 10-801, Spring 2014 (Sra, CMU)
  - EE364a,b (Boyd, Stanford)
  - EE236b,c (Vandenberghe, UCLA)
- Venues: NIPS, ICML, UAI, AISTATS, SIOPT, Math. Prog.

Lecture Plan

– Introduction (3 lectures)
– Problems and algorithms (5 lectures)
– Non-convex optimization, perspectives (2 lectures)

Constrained problems

Optimality – constrained

♠ For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

Optimality – constrained

- ♠ For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
- ♠ Thus, x^* is optimal **if** and only if

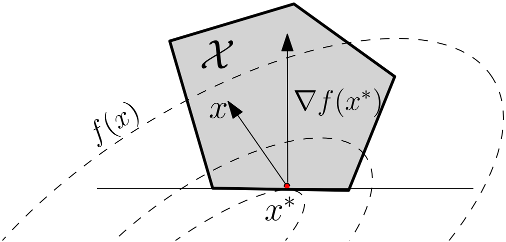
$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \qquad \text{for all } y \in \mathcal{X}.$$

Optimality – constrained

- For every $x, y \in \operatorname{dom} f$, 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^*

Optimality conditions – constrained

Proof:

- 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^*

Example: projection operator

$$P_{\mathcal{X}}(z) := \operatorname*{argmin}_{x \in \mathcal{X}} \|x - z\|^2$$

(Assume $\mathcal{X}$ is closed and convex, then projection is unique)
Let $\mathcal{X}$ be nonempty, closed and convex.

- Optimality condition: $x^* = P_{\mathcal{X}}(y)$ iff

$$\langle x^* - z, y - x^* \rangle \geq 0 \text{ for all } y \in \mathcal{X}$$

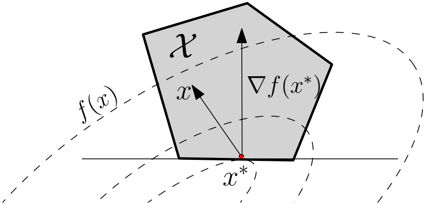
- **Exercise:** Prove that projection is **nonexpansive**, i.e.,

$$\|P_{\mathcal{X}}(x) - P_{\mathcal{X}}(y)\|^2 \leq \|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Feasible descent

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

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



Feasible descent

$$x^{k+1} = x^k + \alpha_k d^k$$

Feasible descent

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$

Feasible descent

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

Feasible descent

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

Since $\mathcal{X}$ is convex, all feasible directions are of the form

$$d^k = \gamma(z - x^k), \quad \gamma > 0,$$

where $z \in \mathcal{X}$ is any feasible vector.

Feasible descent

$$x^{k+1} = x^k + \alpha_k d^k$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

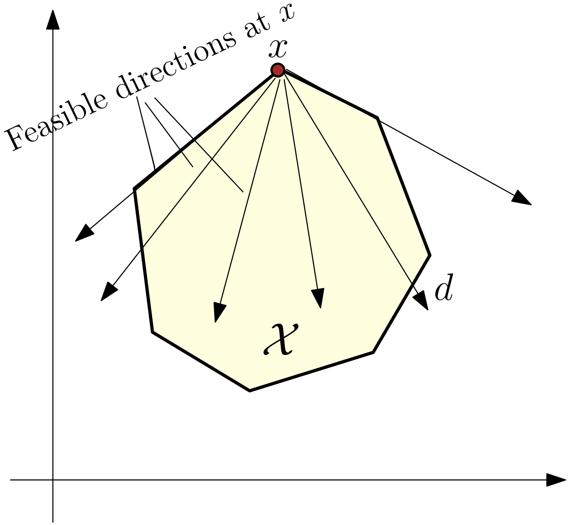
Since $\mathcal{X}$ is convex, all feasible directions are of the form

$$d^k = \gamma(z - x^k), \quad \gamma > 0,$$

where $z \in \mathcal{X}$ is any feasible vector.

$$x^{k+1} = x^k + \alpha_k(z^k - x^k), \quad \alpha_k \in (0, 1]$$

Cone of feasible directions



Frank-Wolfe / conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$

Frank-Wolfe / conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe / conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$

Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe (Conditional gradient) method

- Let $z^k \in \text{argmin}_{x \in \mathcal{X}} \langle \nabla f(x^k), x - x^k \rangle$
- Use different methods to select α_k
- $x^{k+1} = x^k + \alpha_k (z^k - x^k)$

Frank-Wolfe / conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe (Conditional gradient) method

- ▲ Let $z^k \in \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x^k), x - x^k \rangle$
- ▲ Use different methods to select α_k
- ▲ $x^{k+1} = x^k + \alpha_k (z^k - x^k)$

- ♠ Due to M. Frank and P. Wolfe (1956)
- ♠ Practical when solving *linear* problem over $\mathcal{X}$ easy
- ♠ Very popular in machine learning over recent years
- ♠ Refinements, several variants (including nonconvex)

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0, 1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle \nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle \nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z) \leq f(x) + \alpha\langle \nabla f(x), z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

$$f(x^{k+1}) - f(x^*) \quad = \quad f((1-\alpha_k)x^k + \alpha_k z^k) - f(x^*)$$

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle \nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

$$\begin{aligned} f(x^{k+1}) - f(x^*) &= f((1-\alpha_k)x^k + \alpha_k z^k) - f(x^*) \\ &\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, z^k - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \end{aligned}$$

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle \nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &= f((1-\alpha_k)x^k + \alpha_k z^k) - f(x^*)\\
&\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, z^k - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C\\
&\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, x^* - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C
\end{aligned}$$

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle\nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &= f((1-\alpha_k)x^k + \alpha_k z^k) - f(x^*) \\
&\leq f(x^k) - f(x^*) + \alpha_k\langle\nabla f(x^k),\, z^k - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \\
&\leq f(x^k) - f(x^*) + \alpha_k\langle\nabla f(x^k),\, x^* - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \\
&\leq f(x^k) - f(x^*) - \alpha_k(f(x^k) - f(x^*)) + \tfrac{1}{2}\alpha_k^2 C
\end{aligned}$$

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle \nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &= f((1-\alpha_k)x^k + \alpha_k z^k) - f(x^*) \\
&\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, z^k - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \\
&\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, x^* - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \\
&\leq f(x^k) - f(x^*) - \alpha_k(f(x^k) - f(x^*)) + \tfrac{1}{2}\alpha_k^2 C \\
&= (1-\alpha_k)(f(x^k) - f(x^*)) + \tfrac{1}{2}\alpha_k^2 C.
\end{aligned}$$

Frank-Wolfe: Convergence

Assum: There is a $C \geq 0$ s.t. for all $x, z \in \mathcal{X}$ and $\alpha \in (0,1)$:

$$f\big((1-\alpha)x + \alpha z\big) \leq f(x) + \alpha\langle \nabla f(x),\, z - x\rangle + \tfrac{1}{2}C\alpha^2.$$

Let $\alpha_k = \frac{2}{k+2}$. Recall $x^{k+1} = (1-\alpha_k)x^k + \alpha_k z^k$; thus,

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &= f((1-\alpha_k)x^k + \alpha_k z^k) - f(x^*) \\
&\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, z^k - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \\
&\leq f(x^k) - f(x^*) + \alpha_k\langle \nabla f(x^k),\, x^* - x^k\rangle + \tfrac{1}{2}\alpha_k^2 C \\
&\leq f(x^k) - f(x^*) - \alpha_k(f(x^k) - f(x^*)) + \tfrac{1}{2}\alpha_k^2 C \\
&= (1-\alpha_k)(f(x^k) - f(x^*)) + \tfrac{1}{2}\alpha_k^2 C.
\end{aligned}$$

A simple induction (**Verify!**) then shows that

$$f(x^k) - f(x^*) \leq \frac{2C}{k+2}, \qquad k \geq 0.$$

Example: Linear Oracle

Suppose $\mathcal{X} = \{\|x\|_p \leq 1\}$, for $p > 1$.

Write Linear Oracle (LO) as maximization problem:

$$\max_z \quad \langle g, z \rangle \quad \text{s.t.} \;\; \|z\|_p \leq 1.$$

What is the answer?

Example: Linear Oracle

Suppose $\mathcal{X} = \{\|x\|_p \leq 1\}$, for $p > 1$.

Write Linear Oracle (LO) as maximization problem:

$$\max_z \quad \langle g, z \rangle \quad \text{s.t.} \quad \|z\|_p \leq 1.$$

What is the answer?

Hint: Recall, $\langle g, z \rangle \leq \|z\|_p \|g\|_q$. Pick z to obtain equality.

Example: Linear Oracle

Suppose $\mathcal{X} = \{\|x\|_p \leq 1\}$, for $p > 1$.

Write Linear Oracle (LO) as maximization problem:

$$\max_z \quad \langle g, z\rangle \quad \text{s.t.} \quad \|z\|_p \leq 1.$$

What is the answer?

Hint: Recall, $\langle g, z\rangle \leq \|z\|_p \|g\|_q$. Pick z to obtain equality.

Example: Linear Oracle*

Trace norm LO

$$\max_Z \quad \langle G, Z \rangle \qquad \sum\nolimits_i \sigma_i(Z) \le 1.$$

Example: Linear Oracle*

Trace norm LO

$$\max_Z \quad \langle G, Z \rangle \qquad \sum_i \sigma_i(Z) \le 1.$$

- $\max \{\langle G, Z \rangle \mid \|Z\|_* \le 1\}$ is just the "dual-norm" to the trace norm; see Lectures 1-3 for more on dual norms.

Example: Linear Oracle*

Trace norm LO

$$\max_{Z} \quad \langle G, Z \rangle \qquad \sum\nolimits_i \sigma_i(Z) \leq 1.$$

- $\max\{\langle G, Z\rangle \mid \|Z\|_* \leq 1\}$ is just the "dual-norm" to the trace norm; see Lectures 1-3 for more on dual norms.
- can be shown that $\|G\|_2$ is the dual norm here.

Example: Linear Oracle*

Trace norm LO

$$\max_{Z} \quad \langle G, Z \rangle \qquad \sum_i \sigma_i(Z) \leq 1.$$

- $\max \{\langle G, Z\rangle \mid \|Z\|_* \leq 1\}$ is just the "dual-norm" to the trace norm; see Lectures 1-3 for more on dual norms.
- can be shown that $\|G\|_2$ is the dual norm here.
- Optimal Z satisfies $\langle G, Z\rangle = \|G\|_2 \|Z\|_* = \|G\|_2$; use Lanczos (or using power method) to compute top singular vectors.

(for more examples: Jaggi, Revisiting Frank-Wolfe: ...)

Extensions

- How about FW for nonconvex objective functions?
- What about FW methods that can converge faster than $O(1/k)$?

Extensions

- How about FW for nonconvex objective functions?
- What about FW methods that can converge faster than $O(1/k)$?

▶ Nonconvex-FW possible. It "works" (i.e., satisfies first-order optimality conditions to ϵ-accuracy in $O(1/\epsilon)$ iterations (Lacoste-Julien 2016; Reddi et al. 2016).

Extensions

- How about FW for nonconvex objective functions?
- What about FW methods that can converge faster than $O(1/k)$?

▶ Nonconvex-FW possible. It "works" (i.e., satisfies first-order optimality conditions to ϵ-accuracy in $O(1/\epsilon)$ iterations (Lacoste-Julien 2016; Reddi et al. 2016).

▶ Linear convergence under quite strong assumptions on both f and $\mathcal{X}$; alternatively, use a more complicated method: *FW with Away Steps* (Guelat-Marcotte 1986); more recently (Jaggi, Lacoste-Julien 2016)

Quadratic oracle: projection methods

- FW can be quite slow
- If $\mathcal{X}$ not compact, LO does not make sense
- A possible alternative (with other weaknesses though!)

Quadratic oracle: projection methods

- FW can be quite slow
- If $\mathcal{X}$ not compact, LO does not make sense
- A possible alternative (with other weaknesses though!)

If constraint set $\mathcal{X}$ is simple, i.e., we can *easily solve projections*

$$\min \quad \tfrac{1}{2}\|x - y\|_2 \quad \text{s.t. } x \in \mathcal{X}.$$

Quadratic oracle: projection methods

- FW can be quite slow
- If $\mathcal{X}$ not compact, LO does not make sense
- A possible alternative (with other weaknesses though!)

If constraint set $\mathcal{X}$ is simple, i.e., we can *easily solve projections*

$$\min \quad \tfrac{1}{2}\|x - y\|_2 \quad \text{s.t. } x \in \mathcal{X}.$$

Projected Gradient

$$x^{k+1} = P_{\mathcal{X}}\big(x^k - \alpha_k \nabla f(x^k)\big), \quad k = 0, 1, \ldots$$

where $P_{\mathcal{X}}$ denotes above orthogonal projection.

Quadratic oracle: projection methods

- FW can be quite slow
- If $\mathcal{X}$ not compact, LO does not make sense
- A possible alternative (with other weaknesses though!)

If constraint set $\mathcal{X}$ is simple, i.e., we can *easily solve projections*

$$\min \quad \tfrac{1}{2}\|x - y\|_2 \quad \text{s.t. } x \in \mathcal{X}.$$

Projected Gradient

$$x^{k+1} = P_{\mathcal{X}}\big(x^k - \alpha_k \nabla f(x^k)\big), \quad k = 0, 1, \ldots$$

where $P_{\mathcal{X}}$ denotes above orthogonal projection.

- PG can be much faster than $O(1/k)$ of FW (e.g., $O(e^{-k})$ for strongly convex); but LO can be sometimes much faster than projections.

Projected Gradient – convergence

Depends on the following crucial properties of P:

Nonexpansivity: $\|Px - Py\|_2 \leq \|x - y\|_2$

Firm nonxpansivity: $\|Px - Py\|_2^2 \leq \langle Px - Py,\, x - y\rangle$

Projected Gradient – convergence

Depends on the following crucial properties of P:

Nonexpansivity: $\|Px - Py\|_2 \leq \|x - y\|_2$
Firm nonxpansivity: $\|Px - Py\|_2^2 \leq \langle Px - Py,\, x - y\rangle$

▶ Using projections, essentially convergence analysis with $\alpha_k = 1/L$ for the unconstrained case works.

Exercise: Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2$. Write a Matlab/Python script to minimize this function over the convex set $\mathcal{X} := \{-1 \leq x_i \leq 1\}$ using projected gradient as well as Frank-Wolfe. Compare the two.

Duality

Primal problem

Let $f_i : \mathbb{R}^n \to \mathbb{R}$ $(0 \le i \le m)$. Generic **nonlinear program**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.}\ & f_i(x) \le 0, \quad 1 \le i \le m, \\ & x \in \{\operatorname{dom} f_0 \cap \operatorname{dom} f_1 \cdots \cap \operatorname{dom} f_m\}. \end{aligned} \tag{P}$$

Def. Domain: The set $\mathcal{D} := \{\operatorname{dom} f_0 \cap \operatorname{dom} f_1 \cdots \cap \operatorname{dom} f_m\}$

- We call (P) the **primal problem**
- The variable x is the **primal variable**
- We will attach to (P) a **dual problem**
- In our initial derivation: no restriction to convexity.

Lagrangian

To the primal problem, associate **Lagrangian** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

♠ Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**

Lagrangian

To the primal problem, associate **Lagrangian** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**
- If x is feasible, $\lambda \geq 0$, then we get the lower-bound

$$f_0(x) \geq \mathcal{L}(x, \lambda) \qquad \forall x \in \mathcal{X}, \;\; \lambda \in \mathbb{R}^m_+.$$

Lagrangian

To the primal problem, associate **Lagrangian** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**
- If x is feasible, $\lambda \geq 0$, then we get the lower-bound

$$f_0(x) \geq \mathcal{L}(x, \lambda) \qquad \forall x \in \mathcal{X}, \ \ \lambda \in \mathbb{R}^m_+.$$

- Lagrangian helps write problem in **unconstrained form**

Lagrangian

Claim: Since, $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}, \quad \lambda \in \mathbb{R}^m_+$, primal optimal

$$p^* = \inf_{x \in \mathcal{X}} \sup_{\lambda \geq 0} \quad \mathcal{L}(x, \lambda).$$

Lagrangian

Claim: Since, $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}, \ \ \lambda \in \mathbb{R}^m_+$, primal optimal

$$p^* = \inf_{x \in \mathcal{X}} \sup_{\lambda \geq 0} \quad \mathcal{L}(x, \lambda).$$

Proof:

♠ If x is not feasible, then some $f_i(x) > 0$

Lagrangian

Claim: Since, $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}, \quad \lambda \in \mathbb{R}^m_+$, primal optimal

$$p^* = \inf_{x \in \mathcal{X}} \sup_{\lambda \geq 0} \quad \mathcal{L}(x, \lambda).$$

Proof:

- If x is not feasible, then some $f_i(x) > 0$
- In this case, inner sup is $+\infty$, so claim true by definition

Lagrangian

Claim: Since, $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}, \quad \lambda \in \mathbb{R}^m_+$, primal optimal

$$p^* = \inf_{x \in \mathcal{X}} \sup_{\lambda \geq 0} \quad \mathcal{L}(x, \lambda).$$

Proof:

- ♠ If x is not feasible, then some $f_i(x) > 0$
- ♠ In this case, inner sup is $+\infty$, so claim true by definition
- ♠ If x is feasible, each $f_i(x) \leq 0$, so $\sup_\lambda \sum_i \lambda_i f_i(x) = 0$

Lagrange dual function

Def. We define the **Lagrangian dual** as

$$g(\lambda) := \inf_x \quad \mathcal{L}(x, \lambda).$$

Lagrange dual function

Def. We define the **Lagrangian dual** as

$$g(\lambda) := \inf_x \quad \mathcal{L}(x, \lambda).$$

Observations:

- g is pointwise inf of affine functions of λ
- Thus, g is concave; it may take value $-\infty$

Lagrange dual function

Def. We define the **Lagrangian dual** as

$$g(\lambda) := \inf_x \quad \mathcal{L}(x, \lambda).$$

Observations:

- g is pointwise inf of affine functions of λ
- Thus, g is concave; it may take value $-\infty$
- Recall: $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}, \lambda \geq 0$; thus
- $\forall x \in \mathcal{X}, \quad f_0(x) \geq \inf_{x'} \mathcal{L}(x', \lambda) =: g(\lambda)$
- Now minimize over x on lhs, to obtain

$$\forall\, \lambda \in \mathbb{R}^m_+ \qquad p^* \geq g(\lambda).$$

Lagrange dual problem

$$\sup_{\lambda} g(\lambda) \qquad \text{s.t. } \lambda \geq 0.$$

Lagrange dual problem

$$\sup_{\lambda} g(\lambda) \qquad \text{s.t. } \lambda \geq 0.$$

- **dual feasible:** if $\lambda \geq 0$ and $g(\lambda) > -\infty$
- **dual optimal:** λ^* if sup is achieved
- Lagrange dual is **always concave**, regardless of original

Weak duality

Def. Denote **dual optimal value** by d^*, i.e.,

$$d^* := \sup_{\lambda \geq 0} \quad g(\lambda).$$

Weak duality

Def. Denote **dual optimal value** by d^*, i.e.,

$$d^* := \sup_{\lambda \geq 0} \quad g(\lambda).$$

Theorem. (Weak-duality): For problem (P), we have $p^* \geq d^*$.

Weak duality

Def. Denote **dual optimal value** by d^*, i.e.,

$$d^* := \sup_{\lambda \geq 0} \quad g(\lambda).$$

Theorem. (Weak-duality): For problem (P), we have $p^* \geq d^*$.

Proof: We showed that for all $\lambda \in \mathbb{R}^m_+$, $p^* \geq g(\lambda)$.
Thus, it follows that $p^* \geq \sup g(\lambda) = d^*$.

Duality gap

$$p^* - d^* \geq 0$$

Duality gap

$$p^* - d^* \geq 0$$

Strong duality if duality gap is zero: $p^* = d^*$

Notice: both p^* and d^* may be $+\infty$

Duality gap

$$p^* - d^* \geq 0$$

Strong duality if duality gap is zero: $p^* = d^*$

Notice: both p^* and d^* may be $+\infty$

Several **sufficient** conditions known, especially for convex optimization.

"Easy" necessary and sufficient conditions: **unknown**

Example: Slater's sufficient conditions

$$
\begin{aligned}
\min \quad & f_0(x) \\
\text{s.t.} \quad & f_i(x) \le 0, \quad 1 \le i \le m, \\
& Ax = b.
\end{aligned}
$$

Example: Slater's sufficient conditions

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad 1 \le i \le m, \\ & Ax = b. \end{aligned}$$

Constraint qualification: There exists $x \in \operatorname{ri} \mathcal{D}$ s.t.

$$f_i(x) < 0, \qquad Ax = b.$$

That is, there is a **strictly feasible** point.

> **Theorem.** Let the primal problem be convex. If there is a feasible point such that is strictly feasible for the non-affine constraints (and merely feasible for affine, linear ones), then strong duality holds. Moreover, the dual optimal is attained (i.e., $d^* > -\infty$).

Reading: Read BV §5.3.2 for a proof.

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.
Clearly, only feasible $x = 0$. So $p^* = 1$

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x,y) \mid y > 0\}$.
Clearly, only feasible $x = 0$. So $p^* = 1$

$$\mathcal{L}(x,y,\lambda) = e^{-x} + \lambda x^2/y,$$

so dual function is

$$g(\lambda) = \inf_{x,y>0} e^{-x} + \lambda x^2 y = \begin{cases} 0 & \lambda \geq 0 \\ -\infty & \lambda < 0. \end{cases}$$

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.
Clearly, only feasible $x = 0$. So $p^* = 1$

$$\mathcal{L}(x, y, \lambda) = e^{-x} + \lambda x^2/y,$$

so dual function is

$$g(\lambda) = \inf_{x,y>0} e^{-x} + \lambda x^2 y = \begin{cases} 0 & \lambda \geq 0 \\ -\infty & \lambda < 0. \end{cases}$$

Dual problem

$$d^* = \max_{\lambda} 0 \qquad \text{s.t. } \lambda \geq 0.$$

Thus, $d^* = 0$, and gap is $p^* - d^* = 1$.
Here, we had no strictly feasible solution.

Zero duality gap: nonconvex example

Trust region subproblem (TRS)

$$\min \quad x^T A x + 2b^T x \qquad x^T x \leq 1.$$

A is symmetric but not necessarily semidefinite!

Theorem. TRS always has zero duality gap.

Remark: Above theorem extremely important; part of family of related results for certain quadratic nonconvex problems.

Example: Maxent

$$\begin{aligned} \min \quad & \sum\nolimits_i x_i \log x_i \\ & Ax \leq b, \quad 1^T x = 1, \quad x > 0. \end{aligned}$$

Example: Maxent

$$\begin{aligned} \min \quad & \sum_i x_i \log x_i \\ & Ax \leq b, \quad 1^T x = 1, \quad x > 0. \end{aligned}$$

Convex conjugate of $f(x) = x \log x$ is $f^*(y) = e^{y-1}$.

Example: Maxent

$$\min \quad \sum_i x_i \log x_i$$
$$Ax \le b, \quad 1^T x = 1, \quad x > 0.$$

Convex conjugate of $f(x) = x \log x$ is $f^*(y) = e^{y-1}$.

$$\max_{\lambda,\nu} \quad g(\lambda, \nu) = -b^T\lambda - v - \sum_{i=1}^n e^{-(A^T\lambda)_i - \nu - 1}$$
$$\text{s.t.} \quad \lambda \ge 0.$$

Example: Maxent

$$\min \quad \sum_i x_i \log x_i$$
$$Ax \leq b, \quad 1^T x = 1, \quad x > 0.$$

Convex conjugate of $f(x) = x \log x$ is $f^*(y) = e^{y-1}$.

$$\max_{\lambda,\nu} \quad g(\lambda, \nu) = -b^T\lambda - v - \sum_{i=1}^{n} e^{-(A^T\lambda)_i - \nu - 1}$$
$$\text{s.t.} \quad \lambda \geq 0.$$

If there is $x > 0$ with $Ax \leq b$ and $1^T x = 1$, strong duality holds.

Example: Maxent

$$\min \quad \sum_i x_i \log x_i$$
$$Ax \leq b, \quad 1^T x = 1, \quad x > 0.$$

Convex conjugate of $f(x) = x \log x$ is $f^*(y) = e^{y-1}$.

$$\max_{\lambda,\nu} \quad g(\lambda, \nu) = -b^T\lambda - v - \sum_{i=1}^n e^{-(A^T\lambda)_i - \nu - 1}$$
$$\text{s.t.} \quad \lambda \geq 0.$$

If there is $x > 0$ with $Ax \leq b$ and $1^T x = 1$, strong duality holds.
Exercise: Simplify above dual by optimizing out ν

Example: dual for Support Vector Machine

$$\begin{aligned} \min_{x,\xi} \quad & \tfrac{1}{2}\|x\|_2^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & Ax \geq 1 - \xi, \quad \xi \geq 0. \end{aligned}$$

Example: dual for Support Vector Machine

$$\min_{x,\xi} \quad \tfrac{1}{2}\|x\|_2^2 + C\sum_i \xi_i$$
$$\text{s.t.} \quad Ax \geq 1 - \xi, \quad \xi \geq 0.$$

$$L(x, \xi, \lambda, \nu) = \tfrac{1}{2}\|x\|_2^2 + C1^T\xi - \lambda^T(Ax - 1 + \xi) - \nu^T\xi$$

Example: dual for Support Vector Machine

$$\min_{x,\xi} \quad \tfrac{1}{2}\|x\|_2^2 + C\sum\nolimits_i \xi_i$$
$$\text{s.t.} \quad Ax \geq 1 - \xi, \quad \xi \geq 0.$$

$$L(x,\xi,\lambda,\nu) = \tfrac{1}{2}\|x\|_2^2 + C1^T\xi - \lambda^T(Ax - 1 + \xi) - \nu^T\xi$$

$$\begin{aligned} g(\lambda,\nu) &:= \inf L(x,\xi,\lambda,\nu) \\ &= \begin{cases} \lambda^T 1 - \tfrac{1}{2}\|A^T\lambda\|_2^2 & \lambda + \nu = C\mathbf{1} \\ +\infty & \text{otherwise} \end{cases} \\ d^* &= \max_{\lambda \geq 0, \nu \geq 0} \quad g(\lambda,\nu) \end{aligned}$$

Exercise: Using $\nu \geq 0$, eliminate ν from above problem.

Dual via Fenchel conjugates

$$\min f(x) \quad \text{s.t.} \quad f_i(x) \leq 0, \;\; Ax = b.$$

$$\mathcal{L}(x, \lambda, \nu) \quad := \quad f_0(x) + \sum\nolimits_i \lambda_i f_i(x) + \nu^T(Ax - b)$$

Dual via Fenchel conjugates

$$\min f(x) \quad \text{s.t.} \quad f_i(x) \le 0, \;\; Ax = b.$$

$$\begin{aligned} \mathcal{L}(x, \lambda, \nu) &:= f_0(x) + \sum\nolimits_i \lambda_i f_i(x) + \nu^T(Ax - b) \\ g(\lambda, \nu) &= \inf_x \mathcal{L}(x, \lambda, \nu) \end{aligned}$$

Dual via Fenchel conjugates

$$\min f(x) \quad \text{s.t.} \quad f_i(x) \le 0, \; Ax = b.$$

$$\begin{aligned}
\mathcal{L}(x,\lambda,\nu) &:= f_0(x) + \sum\nolimits_i \lambda_i f_i(x) + \nu^T(Ax - b)\\
g(\lambda,\nu) &= \inf_x \mathcal{L}(x,\lambda,\nu)\\
g(\lambda,\nu) &= -\nu^T b + \inf_x x^T A^T \nu + F(x)\\
F(x) &:= f_0(x) + \sum\nolimits_i \lambda_i f_i(x)
\end{aligned}$$

Dual via Fenchel conjugates

$$\min f(x) \quad \text{s.t.} \quad f_i(x) \leq 0, \;\; Ax = b.$$

$$\begin{aligned}
\mathcal{L}(x, \lambda, \nu) &:= f_0(x) + \sum\nolimits_i \lambda_i f_i(x) + \nu^T(Ax - b) \\
g(\lambda, \nu) &= \inf_x \mathcal{L}(x, \lambda, \nu) \\
g(\lambda, \nu) &= -\nu^T b + \inf_x x^T A^T \nu + F(x) \\
F(x) &:= f_0(x) + \sum\nolimits_i \lambda_i f_i(x) \\
g(\lambda, \nu) &= -\nu^T b - \sup_x \langle x, -A^T \nu \rangle - F(x)
\end{aligned}$$

Dual via Fenchel conjugates

$$\min\ f(x) \quad \text{s.t.} \quad f_i(x) \le 0, \;\; Ax = b.$$

$$\begin{aligned}
\mathcal{L}(x,\lambda,\nu) &:= f_0(x) + \textstyle\sum_i \lambda_i f_i(x) + \nu^T(Ax - b) \\
g(\lambda,\nu) &= \inf_x \mathcal{L}(x,\lambda,\nu) \\
g(\lambda,\nu) &= -\nu^T b + \inf_x x^T A^T \nu + F(x) \\
{\color{red} F(x)} &:= {\color{red} f_0(x) + \textstyle\sum_i \lambda_i f_i(x)} \\
g(\lambda,\nu) &= -\nu^T b - \sup_x \langle x, -A^T\nu \rangle - F(x) \\
g(\lambda,\nu) &= -\nu^T b - F^*(-A^T\nu).
\end{aligned}$$

Not so useful! F^* hard to compute.

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Dual problem

$$\min_y \quad f^*(-A^T y) \quad \text{s.t. } \|y\|_* \leq 1.$$

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Dual problem

$$\min_y \quad f^*(-A^T y) \quad \text{s.t. } \|y\|_* \leq 1.$$

Say $\|\bar{y}\|_* < 1$, such that $A^T\bar{y} \in \mathrm{ri}(\mathrm{dom} f^*)$, then we have strong duality (e.g., for instance $0 \in \mathrm{ri}(\mathrm{dom} f^*)$)

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \le 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \le 1\right\}.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \leq 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \leq 1\right\}.$$

Saddle-point formulation

$$p^* \quad = \quad \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda\right\}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \le 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \le 1\right\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda\right\} \\ &= \max_{u,v} \min_x \left\{u^T(b - Ax) + x^T v \mid \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda\right\} \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$
$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

Saddle-point formulation

$$\begin{aligned}
p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \ \|v\|_\infty \leq \lambda \right\} \\
&= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1, \ \|v\|_\infty \leq \lambda \right\} \\
&= \max_{u,v} u^T b \qquad A^T u = v, \ \|u\|_2 \leq 1, \ \|v\|_\infty \leq \lambda
\end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$
$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

Saddle-point formulation

$$\begin{aligned}
p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1,\ \ \|v\|_\infty \leq \lambda \right\} \\
&= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1,\ \ \|v\|_\infty \leq \lambda \right\} \\
&= \max_{u,v} u^T b \qquad A^T u = v,\ \|u\|_2 \leq 1,\ \ \|v\|_\infty \leq \lambda \\
&= \max_{u} u^T b \qquad \|u\|_2 \leq 1, \quad \|A^T v\|_\infty \leq \lambda.
\end{aligned}$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$p^* = f_0(x^*)$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$$p^* = f_0(x^*) = d^* = g(\lambda^*)$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*)$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \le \mathcal{L}(x^*, \lambda^*)$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^*$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \dots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^*$$

- Thus, equalities hold in above chain.

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

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

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^*$$

- Thus, equalities hold in above chain.

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

Example: KKT conditions

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

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

Example: KKT conditions

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

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Example: KKT conditions

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

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

Example: KKT conditions

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

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

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

Example: KKT conditions

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

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

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

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

Example: KKT conditions

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

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, 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

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

KKT conditions

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

- We showed: if strong duality holds, and (x^*, λ^*) exist, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal

KKT conditions

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

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

KKT conditions

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

- We showed: if strong duality holds, and (x^*, λ^*) exist, 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.