Convex Optimization

—

Henry Vu

What is mathematical optimization?

- Standard form

$$\underset{x}{\text{minimize}}\ f_0(x)$$

$$\text{subject to } f_i(x) \leq b_i \qquad i = 1, 2, ..., m$$

- Vocabulary:
  - $\mathbf{f_0(x)}$: objective function
  - $\mathbf{x = (x_1, x_2, ..., x_n)}$: decision variables
  - $\mathbf{f_i(x)}$: inequality constraint functions
  - $\mathbf{b_i}$: bounds for the constraints

- A *feasible set* is the set containing all vectors x's satisfying the constraints, i.e. $f_1(x) \leq b_1$, $f_2(x) \leq b_2$, ..., $f_m(x) \leq b_m$.

- A vector x* is called an *optimal solution* if it has the smallest objective value $f_0(x^*)$. $f_0(x^*)$ is called the *optimal value*.
  i.e. for all feasible x, $f_0(x) \geq f_0(x^*)$

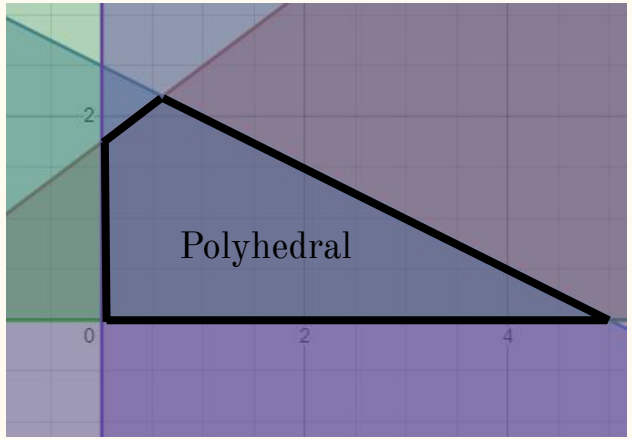
x*: optimal if $f_0(x^*) \leq f_0(x)$
$f_0(x^*)$: optimal objective value

An example: Linear Programming

- A mathematical optimization problem is called *Linear Programming* if its objective and constraint functions, $f_0, f_1, ..., f_m$ are linear.
  i.e. $f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(y) \quad \forall \alpha, \beta \in \mathcal{R}$
  $\forall x, y \in \mathcal{R}^n$

- Example:

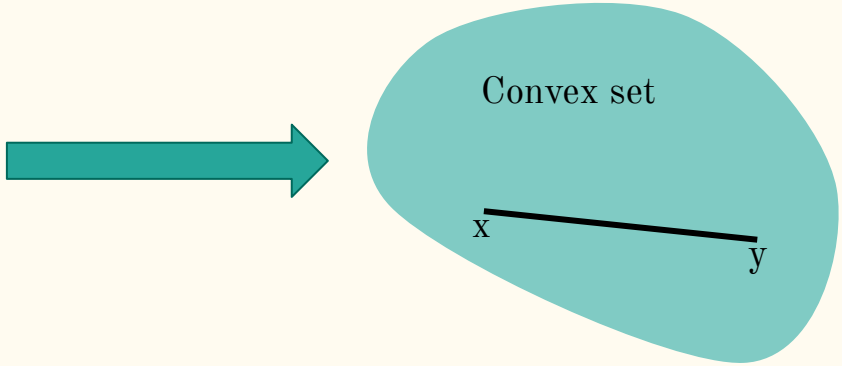
  minimize $x + 2y$
  subject to $4y - 3x \leq 7$
  $2y + x \leq 5$
  $y, x \geq 0$



Convex Optimization

- We are interested in convex optimization problems.
- A convex optimization problem is one in which its objective and constraints functions, $f_0, f_1, ..., f_m$ are convex.
  i.e. $f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y) \quad \forall \alpha, \beta \in \mathcal{R}; \alpha, \beta \geq 0; \alpha + \beta = 1$
  $\forall x, y \in \mathcal{R}^n$
- Standard form

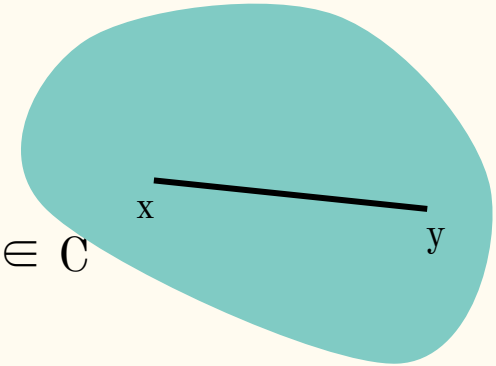
$$\begin{aligned} &\underset{x}{\text{minimize}} && f_0(x) \\ &\text{subject to} && f_i(x) \leq 0 \qquad i = 1, 2, ..., m \\ &&& Ax = b \end{aligned}$$



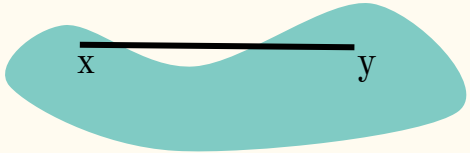
Convex Set

- A subset C of $\mathcal{R}^n$ is called *convex* if

$$\alpha x + (1\text{-}\alpha)y \in C,\ \forall \alpha \in [0, 1],\ \forall x, y \in C$$



- Informally, given any two points in the subset, the line segment joining them is also in the subset.



- Examples of convex set:
  - Hyperplanes $\{Ax = b\}$, Halfspaces $\{Ax \leq b\}$
  - Euclidean ball $B(x_0, \epsilon) = \{x|\ ||x - x_0||_2 \leq \epsilon\}$

Convex Function

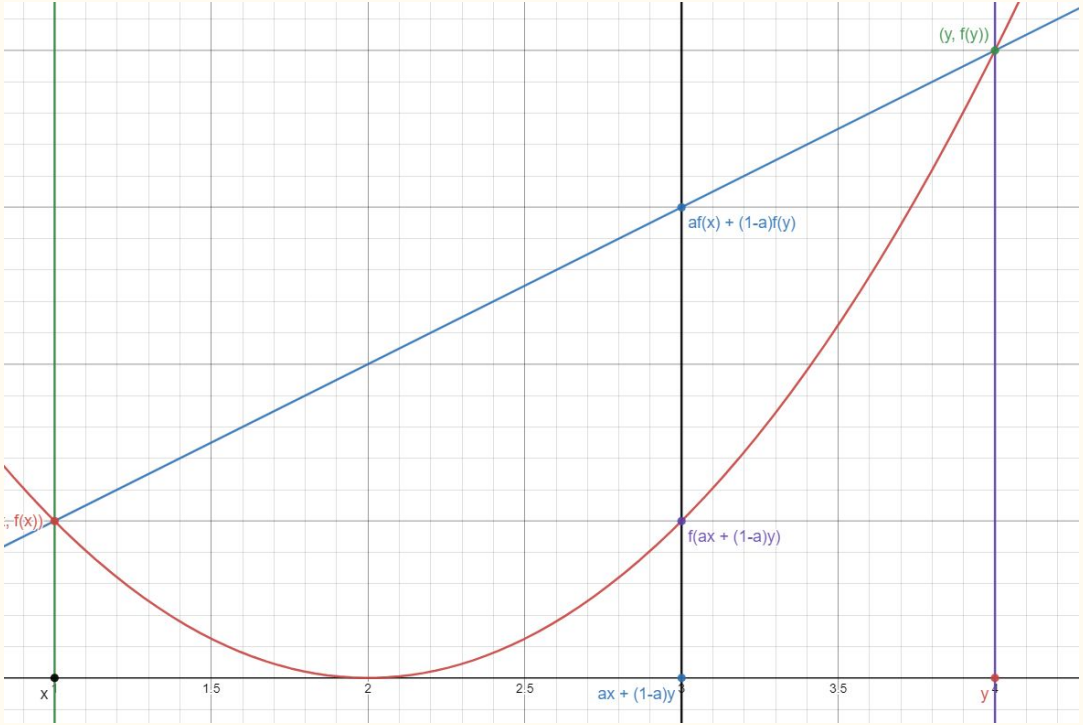
- Let C be a convex subset of $\mathcal{R}^n$ and f: C $\rightarrow \mathcal{R}$ be a function
  f is called a *convex function* if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \quad \forall \alpha \in [0, 1], \forall x, y \in C$$

- f is called *strictly convex over C* if for $\forall \alpha \in (0, 1)$, the strict inequality holds

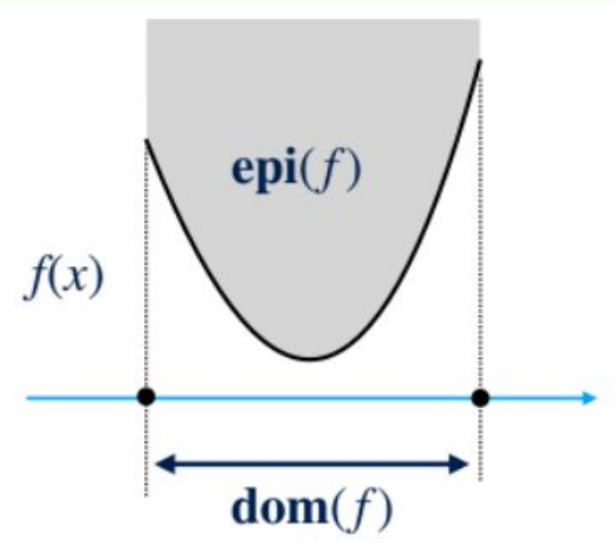
$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y)$$

Convex Function



Epigraph of Convex Functions

- **epi**(f) = {(x, t}|x ∈ **dom**(f), f(x) ≤ t}
- f is convex if and only if its epigraph **epi**(f) is a convex set



Proof:

(⇒) if f is convex

Arbitrarily pick $(x, t_x), (y, t_y) \in$ **epi**(f). Because f is convex

$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \leq \alpha t_x + (1-\alpha)t_y$ by **epi**(f) definition

i.e. $(\alpha x + (1-\alpha)y, \alpha t_x + (1-\alpha)t_y) \in$ **epi**(f)

⇒ **epi**(f) is a convex set

(⇐) if epi(f) is convex

For $(x, f(x))$ and $(y, f(y)) \in$ **epi**(f),

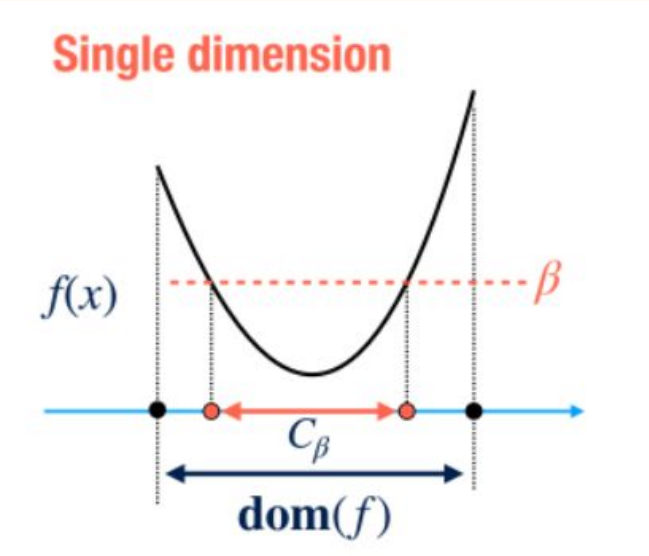
$(\alpha x + (1-\alpha)y, \alpha f(x) + (1-\alpha)f(y)) \in$ **epi**(f) by convexity of **epi**(f)

i.e. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ by **epi**(f) definition

⇒ f is a convex function

Level Sets of Convex Functions

- $\mathbf{C}_{\beta}(f) = \{x \in \mathbf{dom}(f) | \ f(x) \leq \beta\}$
- if f is convex, all level sets of f is convex



Proof:

(⇒) if f is convex

Arbitrarily pick x, y $\in \mathbf{C}_{\beta}(f)$. Then because f is convex

$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \leq \alpha\beta + (1-\alpha)\beta = \beta$

i.e. $\alpha x + (1-\alpha)y \in \mathbf{C}_{\beta}(f)$

$\Rightarrow$ $\mathbf{C}_{\beta}(f)$ is a convex set

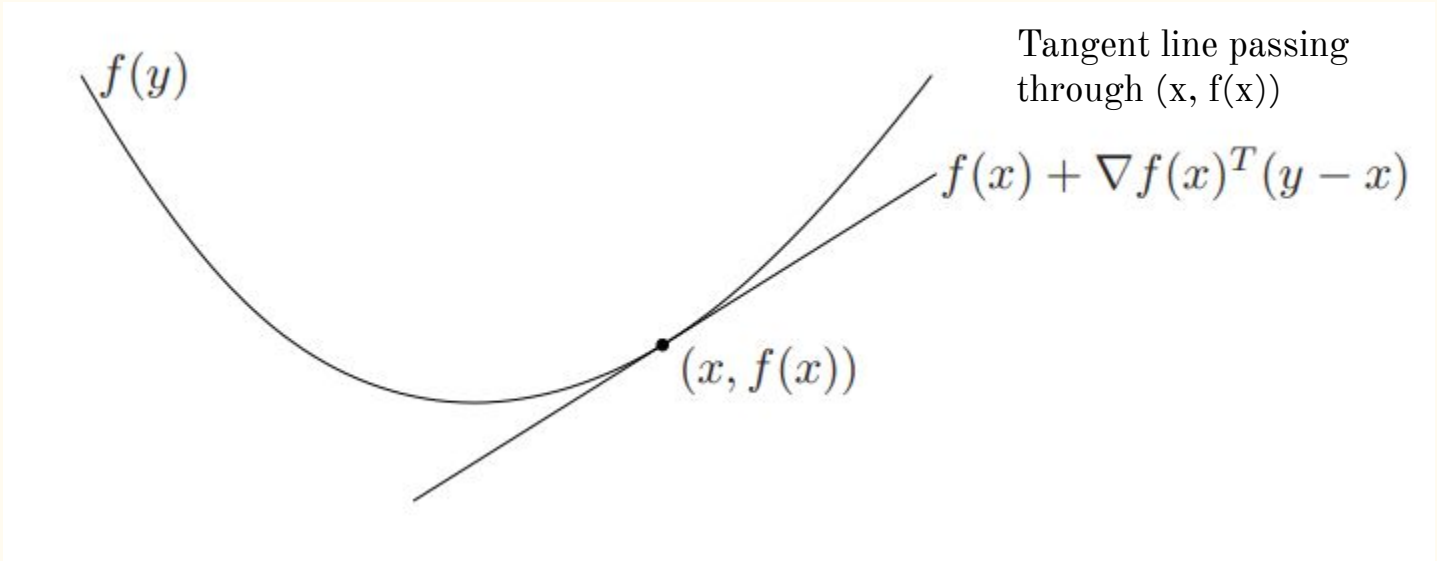
First-Order Conditions

- Suppose f is differentiable (i.e. its gradient exists at each point in **dom**(f)). Then f is convex if and only if for x, y $\in$ **dom**(f)

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

- The inequality states that the first-order Taylor expansion is a global underestimator of a convex function.
- The inequality shows that we can derive *global information* (global underestimator) from *local information* (value and derivative at a point)

First-Order Conditions



Proof:

(⇒) if f is convex. $\forall$ y, x $\in$ **dom**(f)

$$f(\alpha y + (1-\alpha)x) \leq \alpha f(y) + (1-\alpha)f(x)$$
$$\Rightarrow f(x + \alpha(y-x)) \leq f(x) + \alpha(f(y) - f(x))$$
$$\Rightarrow (y-x)\frac{f(x+\alpha(y-x))}{\alpha(y-x)} \leq f(y) - f(x)$$

Since the LHS increases as $\alpha \to 0$, we take the limit as $\alpha \to 0$ for both sides and obtain

$$(y-x)\nabla f(x)^T + f(x) \leq f(y)$$

Proof (cont.):

(⇐) if the first-order conditions hold

Let z = αx + (1-α)y

Applying the first order condition for the (x, z) pair and (y, z) pair:

$$f(x) \geq f(z) + \nabla f(z)^T (x - z)$$

$$f(y) \geq f(z) + \nabla f(z)^T (y - z)$$

Multiplying the inequalities by α and (1-α) respectively then add them together, we get:

$$\alpha f(x) + (1 - \alpha) f(y) \geq f(z)$$

i.e $\alpha f(x) + (1 - \alpha) f(y) \geq f(\alpha x + (1 - \alpha) y)$

Second-Order Conditions

- Suppose f is twice differentiable. Then f is convex if and only if for x $\in$ **dom**(f)

$$\nabla^2 f(x) \succeq 0$$

- The inequality implies that the gradient of f is a nondecreasing function.
- Geometrically speaking, the function has an upward curvature at x.

Proof

(⇒) if f is convex

We prove in dimension 1 then generalize

Applying the first order condition for the (x, y) pair and (y, x) pair:

$$f(x) \geq f(y) + f'(y)(x - y)$$

$$f(y) \geq f(x) + f'(x)(y - x)$$

Adding them together, we get:

$$f(x) + f(y) \geq f(y) + f(x) + (x - y)(f'(y) - f'(x))$$

$$\Rightarrow (y - x)(f'(y) - f'(x)) \geq 0$$

Dividing both sides by a positive quantity $(y - x)^2$ for $y \neq x$, we get:

$$\frac{f'(y) - f'(x)}{y - x} \geq 0$$

As $y \to x$:

$$f''(x) \geq 0$$

Proof (cont.):

(⇐) if the second-order conditions hold

Using the 2nd order Taylor expansion and the assumption that $\nabla^2 f(x) \succeq 0$ we get the first-order conditions:

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

The first-order conditions imply convexity of f

Important Property:

- **Theorem**: Let f(x) be convex. If x* is a local minimum of f over a convex set C, then x* is also a global minimum of f over a convex set C.
- Can be proved by setting the gradient at x* to 0 and use the first-order conditions.

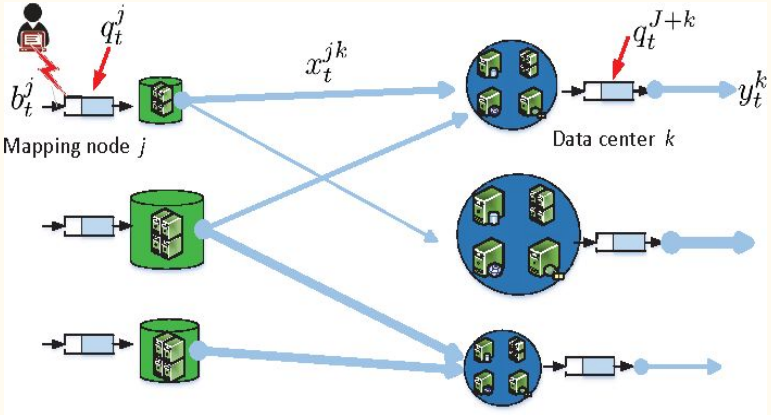
Why Convex Optimization

Nice properties:

- Local minimum is also global minimum
- Differentiable, continuous, has duality
- A convex set has nice shape

Application of Convex Optimization

- Automatic control systems
- Resource allocation
- Data analysis and modeling
- Communications and networks
- ...



A Closer Look at Convex Optimization

- First-order optimality conditions
- Lagrange multiplier
- Duality theory
- KKT conditions

Epigraph form (standard form)

$$\begin{aligned} \underset{x}{\text{minimize}}\ & f_0(x) \\ \text{subject to}\ & f_i(x) \leq 0 \qquad \text{i} = 1, 2, ..., \text{m} \\ & Ax = b \end{aligned}$$

First-Order Optimality Conditions

- If x* is the solution to an unconstrained program

$$x^* = \underset{x}{\operatorname{argmin}} f_0(x) \Leftrightarrow \nabla f_0(x^*) = 0$$

(Fermat's theorem)

- If x* is the solution to a convex program with feasible set C

$$x^* = \underset{x}{argmin} f_0(x) \Leftrightarrow \nabla f_0(x)^T (x - x^*) \geq 0 \, \forall x \in C$$

Proof

(⇒) if $x^* = \underset{x}{\operatorname{argmin}} f_0(x)$

$$\text{Assume } \exists \overline{x} \in C \text{ s.t. } \nabla f(x^*)^T(\overline{x} - x^*) < 0$$

$$\Rightarrow \lim_{t \longrightarrow 0} \frac{f(x^* + t(\overline{x} - x^*)) - f(x^*)}{t} < 0$$

$$\text{As } t \longrightarrow 0, \text{ we get}$$

$$f(x^* + t(\overline{x} - x^*)) < f(x^*) \text{ contradiction}$$

Proof (cont.):

(⇐) if $\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in C$

By the first-order convexity conditions:

$$f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*) \geq f(x^*)$$

$$\text{i.e. } x^* = \underset{x}{argmin} f(x)$$

Equivalent Form of Convex Problem

- A convex problem with feasible set C is equivalent to the following unconstrained problem

$$\underset{x}{minimize}\ f_0(x) + I_C(x)$$

- $I_C(x)$ is called an *indicator function*

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & \text{else} \end{cases}$$

Lagrange Multipliers

$$\begin{aligned} \underset{x}{\text{minimize}}\ & f_0(x) \\ \text{subject to}\ & f_i(x) \leq 0 \qquad \text{i} = 1, 2, ..., \text{m} \\ & Ax = b \end{aligned}$$

- Lagrangian function for the above optimization problem is

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T (Ax - b)$$

- λ_i's are called *Lagrange multipliers* (or *dual variables*)

- Notice that supremum over the Lagrangian function gives back $f_0(x)+I_C(x)$

$$\sup_{\lambda \geq 0} L(x, \lambda, \nu) = \sup_{\lambda \geq 0} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T (Ax - b)\right)$$

$$= \begin{cases} f_0(x) & \text{when x in feasible} \\ \infty & \text{otherwise} \end{cases}$$

- The original problem (now called *Primal Form*) is:

$$p^* = \inf_x \sup_{\lambda \geq 0} L(x, \lambda, \nu)$$

- We get the *Lagrangian dual problem* by "swapping the inf and the sup"

$$d^* = \sup_{\lambda \geq 0} \inf_x L(x, \lambda, \nu)$$

- We can show that d* gives a lower bound on p* (*weak duality*)

$$\mathbf{d^* \leqslant p^*}$$

Dual Problem

$$Let\ g(\lambda,\nu) = \inf_x L(x,\lambda,\nu) \text{ then for any } \lambda \geq 0 \text{ and } \nu$$

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu)$$

$$= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T(Ax-b)\right)$$

$$\leq \inf_{x\in C} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \nu^T(Ax-b)\right)$$

$$\leq \inf_{x\in C} f_0(x) = p^*$$

$$\text{Since this is true for any } \lambda \geq 0 \text{ and } \nu$$

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

Strong Duality and Slater's Conditions

- The difference p* - d* is called the *duality gap*
- For convex problems, we often have *strong duality*: $p^* = d^*$

Slater's Conditions: Strong duality holds for a convex problem if it is strictly feasible, i.e. $\exists\, x \in \mathbf{relint}(C)$: $f_i(x) < 0 \;\forall\, i$, $Ax - b = 0$.

- **relint**(C) is the relative interior of C and is defined as
  $\mathbf{relint}(C) = \{x \in C \mid \exists\, \epsilon \geq 0: N_\epsilon(x) \cap \mathbf{aff}(C) \subseteq C\}$
- Informally, the Slater's Conditions state that the feasible region must have an interior point

Complementary Slackness

- Assuming *strong duality* holds: p* = d*

$$f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + (\nu^*)^T (Ax - b)\right)$$

$$\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + (\nu^*)^T (Ax^* - b) \leq f_0(x^*)$$

$$\Rightarrow 0 \leq \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \leq 0 \Rightarrow \sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

$$\text{But since } \lambda^* \geq 0 \text{ and } f_i(x*) \leq 0 \; \forall i$$

$$\lambda_i^* f_i(x^*) = 0 \; \forall i$$

- $\lambda_i^* f_i(x^*) = 0 \; \forall i$ is called *Complementary Slackness*

Stationarity

- By the first-order conditions for optimality

$$\nabla_x L(x^*, \lambda^*, \nu^*) = \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + A^T v^* = 0$$

- Informally speaking, the stationarity condition states that the gradient of the objective function and its constraints must be parallel, i.e. moving along the constraint surface does not improve the objective function.

KKT Conditions

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \nu^*) = \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + A^\intercal \nu^* = 0$$ --- Stationarity

$$\lambda_i^* f_i(x^*) = 0 \quad i = 1, \cdots, m$$ ---------- Complementary slackness

$$Ax^* = b, \quad f_i(x^*) \leq 0 \quad i = 1, \cdots, m$$ ---------- Primal feasibility

$$\lambda_i^* \geq 0, \; i = 1, \cdots, m$$ ---------- Dual feasibility

I. For any *differential* problem (potentially non-convex):

   Strong duality + Optimality ⇒ KKT

II. For convex problems:

   KKT ⇒ Strong duality + Optimality

III. So by I and II, the KKT conditions are both *necessary* and *sufficient* for primal/dual optimality

Some Algorithms

- Gradient Descent
- Newton's Method

Gradient Descent

- Schema for a general descent method:
  Update $x_{k+1} = x_k + t_k \nabla_k$ such that $f(x_{k+1}) \leq f(x_k)$ where t_k is the step size and ∇_k is the search direction

- Using directional derivative, we can deduce the direction of steepest ascent for function f differentiable at x_0 to be the gradient $\nabla_k = \nabla f(x_0)/||\nabla f(x_0)||$, and the direction of steepest descent is $\nabla_k = -\nabla f(x_0)/||\nabla f(x_0)||$

- So the updating step for Gradient Descent can be set as:
  $$x_{k+1} = x_k - t\nabla f(x_k)$$

Convergence of Gradient Descent

- **Assumption**: Suppose f is *convex* and *differentiable*, has optimal value f(x*) and the gradient ∇f is *Lipschitz continuous* with constant L > 0.
  That is: $||\nabla f(x) - \nabla f(y)|| \leq L||x-y|| \quad \forall x, y$

  Then if we run Gradient Descent for k iterations with fixed step size $t \leq 1/L$, it will yield solution $f(x_k)$ satisfying:

$$f(x_k) - f(x^*) \leq \frac{||x_0 - x^*||^2}{2tk}$$

  That is, Gradient Descent is guaranteed to converge with O(1/k) convergence rate.

Strongly Convex Case

- If additional f is strongly convex with parameter m, that is

$$f(y) \geq f(x) + \nabla^T f(x)(y-x) + \frac{m}{2}||y-x||^2$$

  then Gradient Descent with step size $t \leq 2/(m+L)$ will satisfy:

$$f(x_k) - f(x^*) \leq \gamma^k \frac{L}{2}||x_0 - x^*||^2 \quad where\ \gamma \in (0,1)$$

- The algorithm converges exponentially fast ($O(\gamma^k)$) if the function is strongly convex

Newton's Method

- Using 2nd-order Taylor's theorem for a twice differentiable function f:

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) + o(||h^2||)$$

Then

$$\underset{h}{argmin} f(x+h) = \underset{h}{argmin}(f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x))$$

$$= -(\nabla^2 f(x))^{-1} \nabla f(x) \text{ obtained by setting } \nabla f(x+h) \text{ to } 0$$

- So the updating step for Newton's Method can be set as:

$$x_{k+1} = x_k - t(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

Convergence of Newton's Method

- **Theorem:** suppose f satisfies

$$\begin{cases} (1)\ ||\nabla^2 f(x) - \nabla^2 f(y)|| \leq R||x-y|| \\ (2)\ \nabla^2 f(x^*) \succeq \mu I \\ (3)\ ||x_0 - x^*|| \leq \frac{\mu}{2R} \end{cases}$$

Then:

$$||x_{k+1} - x^*|| \leq \frac{R}{\mu}||x_k - x^*||^2$$

In other words, Newton's Method converges quadratically