# Convex Optimization

- optimization problems
- convex sets
- convex functions
- convex problems
- duality
- additional topics

slides compiled by Neal Parikh for CS 228T, Stanford University
most content/figures from Boyd and Vandenberghe (errors mine)

# Mathematical optimization

- problems of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in S \end{array}$$

- convex optimization: minimizing a convex function over a convex set
  - tractable to solve (even with nondifferentiable objective)
  - powerful both for theory and practice

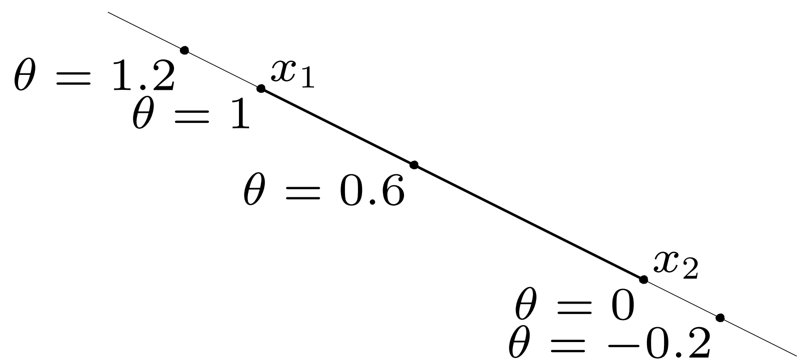
- combinatorial optimization
  - when $S$ is discrete, $e.g.$, $x \in \{0, 1\}^n$
  - when difficult, often solved via *convex relaxations*

- nonconvex optimization
  - can only find local optima
  - choice of algorithm is much more important

# Affine set

**line** through $x_1$, $x_2$: all points

$$x = \theta x_1 + (1 - \theta) x_2 \qquad (\theta \in \mathbf{R})$$




**affine set**: contains the line through any two distinct points in the set

**example**: solution set of linear equations $\{x \mid Ax = b\}$

(conversely, every affine set can be expressed as solution set of system of linear equations)

# Convex set

**line segment** between $x_1$ and $x_2$: all points

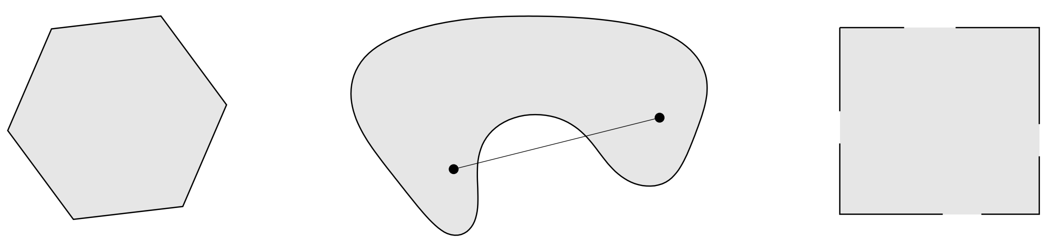
$$x = \theta x_1 + (1-\theta)x_2$$

with $0 \le \theta \le 1$

**convex set**: contains line segment between any two points in the set

$$x_1, x_2 \in C, \quad 0 \le \theta \le 1 \quad \Longrightarrow \quad \theta x_1 + (1-\theta)x_2 \in C$$

**examples** (one convex, two nonconvex sets)



# Convex combination and convex hull

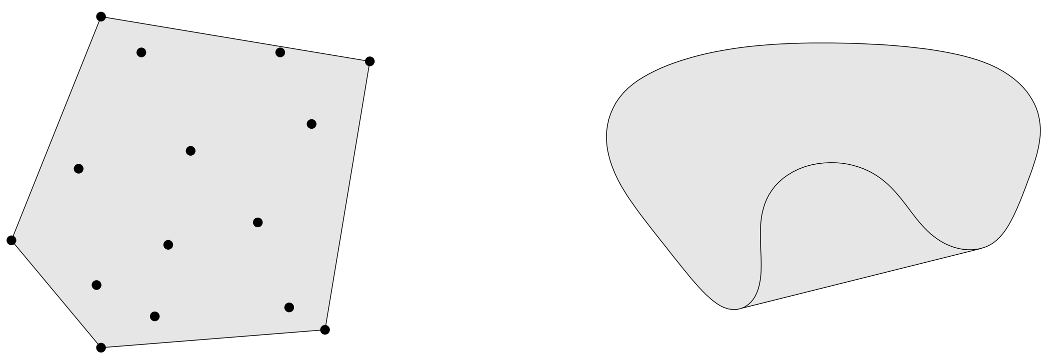
**convex combination** of $x_1$,. . . , $x_k$: any point $x$ of the form

$$x = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

with $\theta_1 + \cdots + \theta_k = 1$, $\theta_i \geq 0$

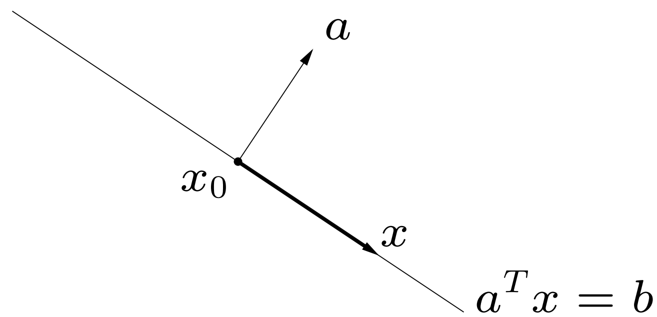
can view this probabilistically as a *mixture* or *expectation*

**convex hull** $\mathbf{conv}\, S$: set of all convex combinations of points in $S$

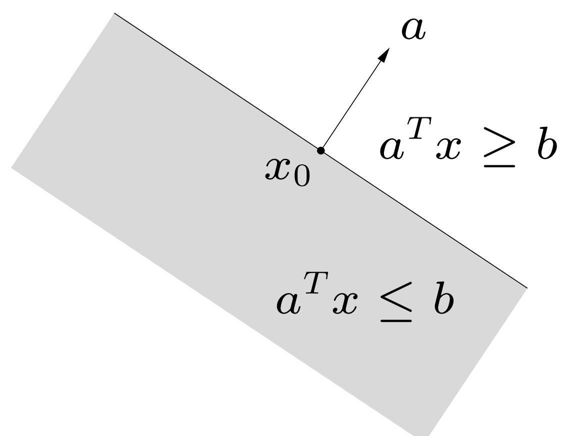


# Hyperplanes and halfspaces

**hyperplane**: set of the form $\{x \mid a^T x = b\}$ $(a \neq 0)$




**halfspace:** set of the form $\{x \mid a^T x \leq b\}$ $(a \neq 0)$




- $a$ is the normal vector
- hyperplanes are affine and convex; halfspaces are convex

# Euclidean balls and ellipsoids

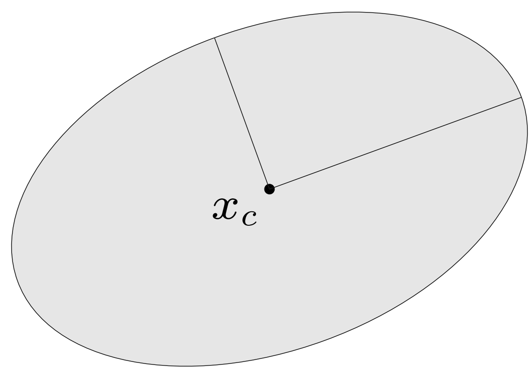
**(Euclidean) ball** with center $x_c$ and radius $r$:

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \leq r\} = \{x_c + ru \mid \|u\|_2 \leq 1\}$$

**ellipsoid:** set of the form

$$\{x \mid (x - x_c)^T P^{-1}(x - x_c) \leq 1\}$$

with $P \in \mathbf{S}^n_{++}$ (*i.e.*, $P$ symmetric positive definite)



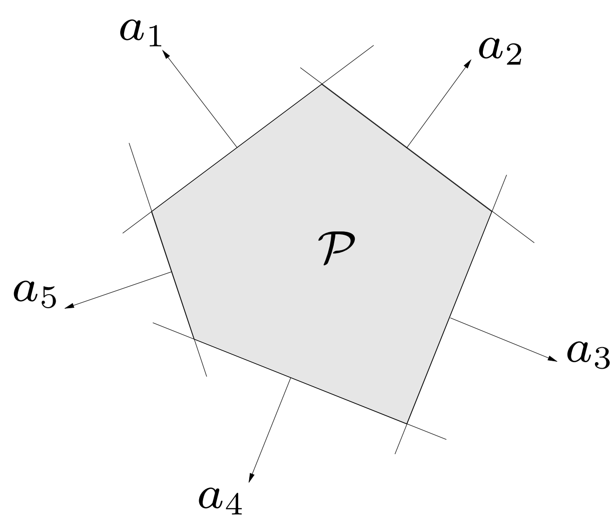

with $A$ square and nonsingular

# Polyhedra and polytopes

solution set of finitely many linear inequalities and equalities

$$Ax \preceq b, \qquad Cx = d$$

($A \in \mathbf{R}^{m \times n}$, $C \in \mathbf{R}^{p \times n}$, $\preceq$ is componentwise inequality)




polyhedron is intersection of finite number of halfspaces and hyperplanes

bounded polyhedron is called a polytope; can also be expressed as the convex hull of its vertices (Minkowski-Weyl theorem)

# Operations that preserve convexity

practical methods for establishing convexity of a set $C$

1. apply definition

$$x_1, x_2 \in C, \quad 0 \leq \theta \leq 1 \quad \Longrightarrow \quad \theta x_1 + (1-\theta) x_2 \in C$$

2. show that $C$ is obtained from simple convex sets (hyperplanes, halfspaces, norm balls, . . . ) by operations that preserve convexity

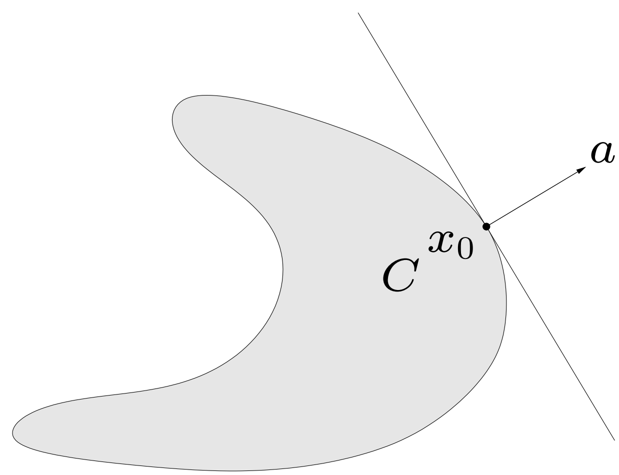
   - intersection
   - many others

# Supporting hyperplane theorem

**supporting hyperplane** to set $C$ at boundary point $x_0$:

$$\{x \mid a^T x = a^T x_0\}$$

where $a \neq 0$ and $a^T x \leq a^T x_0$ for all $x \in C$



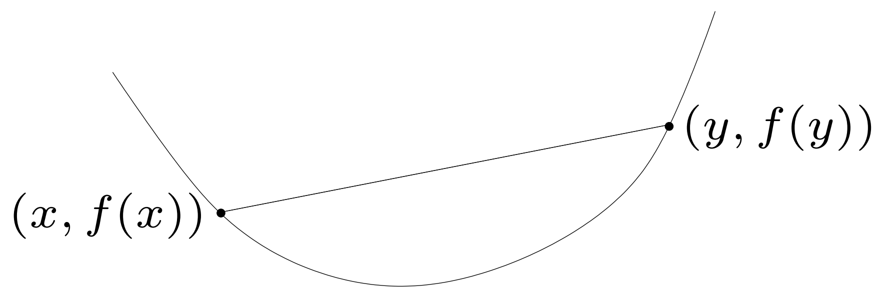

**supporting hyperplane theorem:** if $C$ is convex, then there exists a supporting hyperplane at every boundary point of $C$

# Convex functions

$f : \mathbf{R}^n \to \mathbf{R}$ is convex if $\mathbf{dom}\, f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in \mathbf{dom}\, f$, $0 \leq \theta \leq 1$




- $f$ is concave if $-f$ is convex
- $f$ is strictly convex if $\mathbf{dom}\, f$ is convex and

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

  for $x, y \in \mathbf{dom}\, f$, $x \neq y$, $0 < \theta < 1$

# Examples on $\mathbf{R}$

convex:

- affine: $ax + b$ on $\mathbf{R}$, for any $a, b \in \mathbf{R}$
- exponential: $e^{ax}$, for any $a \in \mathbf{R}$
- powers: $x^\alpha$ on $\mathbf{R}_{++}$, for $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute value: $|x|^p$ on $\mathbf{R}$, for $p \geq 1$
- negative entropy: $x \log x$ on $\mathbf{R}_{++}$

concave:

- affine: $ax + b$ on $\mathbf{R}$, for any $a, b \in \mathbf{R}$
- powers: $x^\alpha$ on $\mathbf{R}_{++}$, for $0 \leq \alpha \leq 1$
- logarithm: $\log x$ on $\mathbf{R}_{++}$

# Extended-value extension

extended-value extension $\tilde{f}$ of $f$ is

$$\tilde{f}(x) = f(x), \quad x \in \mathbf{dom}\, f, \qquad \tilde{f}(x) = \infty, \quad x \notin \mathbf{dom}\, f$$

often simplifies notation; for example, the condition

$$0 \leq \theta \leq 1 \quad \Longrightarrow \quad \tilde{f}(\theta x + (1-\theta)y) \leq \theta \tilde{f}(x) + (1-\theta)\tilde{f}(y)$$

(as an inequality in $\mathbf{R} \cup \{\infty\}$), means the same as the two conditions

- $\mathbf{dom}\, f$ is convex
- for $x, y \in \mathbf{dom}\, f$,

$$0 \leq \theta \leq 1 \quad \Longrightarrow \quad f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

# First-order condition

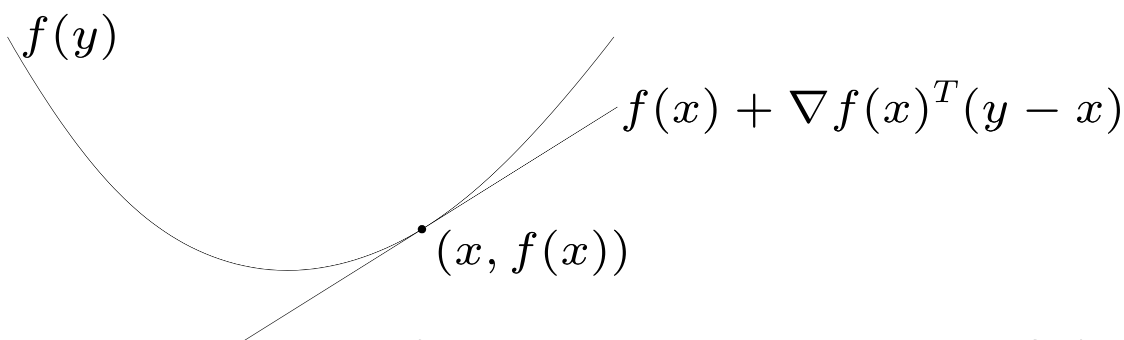
$f$ is **differentiable** if $\mathbf{dom}\, f$ is open and the gradient

$$\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)$$

exists at each $x \in \mathbf{dom}\, f$

**1st-order condition:** differentiable $f$ with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathbf{dom}\, f$$




first-order approximation of $f$ is global underestimator

# Second-order conditions

$f$ is **twice differentiable** if $\mathbf{dom}\, f$ is open and the Hessian $\nabla^2 f(x) \in \mathbf{S}^n$,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n,$$

exists at each $x \in \mathbf{dom}\, f$

**2nd-order conditions:** for twice differentiable $f$ with convex domain

- $f$ is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbf{dom}\, f$$

- if $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{dom}\, f$, then $f$ is strictly convex

# Examples

**quadratic function:** $f(x) = (1/2)x^T Px + q^T x + r$ (with $P \in \mathbf{S}^n$)

$$\nabla f(x) = Px + q, \qquad \nabla^2 f(x) = P$$

convex if $P \succeq 0$

**least-squares objective**: $f(x) = \|Ax - b\|_2^2$

$$\nabla f(x) = 2A^T(Ax - b), \qquad \nabla^2 f(x) = 2A^T A$$

convex (for any $A$)

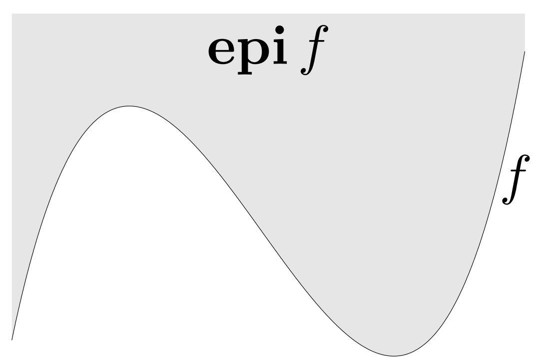
**log-sum-exp**: $f(x) = \log \sum_{k=1}^n \exp x_k$ is convex

can generalize to $\log \int \exp$

# Relationship of convex sets and functions

**epigraph** of $f : \mathbf{R}^n \to \mathbf{R}$:

$$\mathbf{epi}\, f = \{(x,t) \in \mathbf{R}^{n+1} \mid x \in \mathbf{dom}\, f,\ f(x) \le t\}$$




$f$ is convex if and only if $\mathbf{epi}\, f$ is a convex set

# Jensen's inequality

**basic inequality:** if $f$ is convex, then for $0 \leq \theta \leq 1$,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

**extension:** if $f$ is convex, then

$$f(\mathbf{E}\, z) \leq \mathbf{E}\, f(z)$$

for any random variable $z$

useful source of lower bounds

basic inequality is special case with discrete distribution

$$\mathbf{prob}(z = x) = \theta, \qquad \mathbf{prob}(z = y) = 1 - \theta$$

# Verifying convexity

practical methods for establishing convexity of a function

1. verify definition

2. for twice differentiable functions, show $\nabla^2 f(x) \succeq 0$

3. show that $f$ is obtained from simple convex functions by operations that preserve convexity

   - nonnegative weighted sum
   - composition with affine function
   - pointwise maximum and supremum
   - composition

# Operations that preserve convexity

nonnegative multiple: $\alpha f$ is convex if $f$ is convex, $\alpha \geq 0$

sum: $f_1 + f_2$ convex if $f_1, f_2$ convex (extends to infinite sums, integrals)

composition with affine function: $f(Ax + b)$ is convex if $f$ is convex

if $f_1$, . . . , $f_m$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex

if $f(x, y)$ is convex in $(x, y)$ and $C$ is a convex set, then

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex

*e.g.*, distance to a set: $\mathbf{dist}(x, S) = \inf_{y \in S} \|x - y\|$ is convex if $S$ is convex

# Optimization problem in standard form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

- $x \in \mathbf{R}^n$ is the optimization variable
- $f_0 : \mathbf{R}^n \to \mathbf{R}$ is the objective or cost function
- $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \ldots, m$, are the inequality constraint functions
- $h_i : \mathbf{R}^n \to \mathbf{R}$ are the equality constraint functions

**optimal value:**

$$p^\star = \inf\{f_0(x) \mid f_i(x) \leq 0,\ i = 1, \ldots, m,\ h_i(x) = 0,\ i = 1, \ldots, p\}$$

- $p^\star = \infty$ if problem is infeasible (no $x$ satisfies the constraints)
- $p^\star = -\infty$ if problem is unbounded below

# Optimal and locally optimal points

$x$ is **feasible** if $x \in \mathbf{dom}\, f_0$ and it satisfies the constraints

a feasible $x$ is **optimal** if $f_0(x) = p^\star$; $X_{\text{opt}}$ is the set of optimal points

$x$ is **locally optimal** if there is an $R > 0$ such that $x$ is optimal for

$$
\begin{array}{ll}
\text{minimize (over } z) & f_0(z) \\
\text{subject to} & f_i(z) \leq 0, \quad i = 1, \ldots, m, \quad h_i(z) = 0, \quad i = 1, \ldots, p \\
& \|z - x\|_2 \leq R
\end{array}
$$

# Implicit constraints

the standard form optimization problem has an **implicit constraint**

$$x \in \mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \;\cap\; \bigcap_{i=1}^{p} \mathbf{dom}\, h_i,$$

- we call $\mathcal{D}$ the **domain** of the problem
- the constraints $f_i(x) \leq 0$, $h_i(x) = 0$ are the explicit constraints
- a problem is **unconstrained** if it has no explicit constraints ($m = p = 0$)

**example**:

$$\text{minimize} \quad f_0(x) = -\textstyle\sum_{i=1}^{k} \log(b_i - a_i^T x)$$

is an unconstrained problem with implicit constraints $a_i^T x < b_i$

# Convex optimization problem

**standard form convex optimization problem**

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_i^T x = b_i, \quad i = 1, \ldots, p \end{array}$$

$f_0$, $f_1$, . . . , $f_m$ are convex; equality constraints are affine

often written as

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

feasible set of a convex optimization problem is convex

any locally optimal point of a convex problem is (globally) optimal

# Optimality criterion for differentiable $f_0$

$x$ is optimal if and only if it is feasible and

$$\nabla f_0(x)^T(y - x) \geq 0 \quad \text{for all feasible } y$$

if nonzero, $\nabla f_0(x)$ defines a supporting hyperplane to feasible set $X$ at $x$

**unconstrained problem**: $x$ is optimal if and only if

$$x \in \mathbf{dom}\, f_0, \qquad \nabla f_0(x) = 0$$

# Equivalent convex problems

two problems are (informally) **equivalent** if the solution of one is readily obtained from the solution of the other, and vice-versa

some common transformations that preserve convexity:

- **introducing slack variables for linear inequalities**

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & a_i^T x \leq b_i, \quad i = 1, \ldots, m \end{array}$$

  is equivalent to

$$\begin{array}{ll} \text{minimize (over } x\text{, } s\text{)} & f_0(x) \\ \text{subject to} & a_i^T x + s_i = b_i, \quad i = 1, \ldots, m \\ & s_i \geq 0, \quad i = 1, \ldots m \end{array}$$

- **minimizing over some variables**

$$\begin{array}{ll} \text{minimize} & f_0(x_1, x_2) \\ \text{subject to} & f_i(x_1) \leq 0, \quad i = 1, \ldots, m \end{array}$$

is equivalent to

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(x_1) \\ \text{subject to} & f_i(x_1) \leq 0, \quad i = 1, \ldots, m \end{array}$$

where $\tilde{f}_0(x_1) = \inf_{x_2} f_0(x_1, x_2)$

- **consensus**

$$\text{minimize} \quad f_1(x) + f_2(x) + \cdots + f_k(x)$$

is equivalent to

$$\begin{array}{ll} \text{minimize} & f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k) \\ \text{subject to} & x_i = x, \quad i = 1, \ldots, k \end{array}$$

# Examples of convex optimization problems

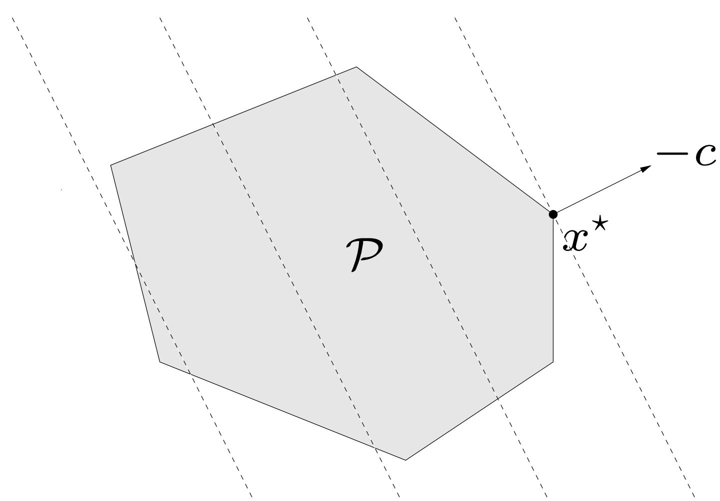
- maximum entropy
- maximum likelihood estimation in exponential families
- projection onto a convex set
  - Euclidean projection (measure distance to set in $\ell_2$ norm)
  - Bregman projection (measure via *Bregman divergence*)

    $e.g.$, minimum KL divergence to a convex set of distributions

# Linear program (LP)

$$\begin{array}{ll}\text{minimize} & c^T x + d \\ \text{subject to} & Gx \preceq h \\ & Ax = b\end{array}$$

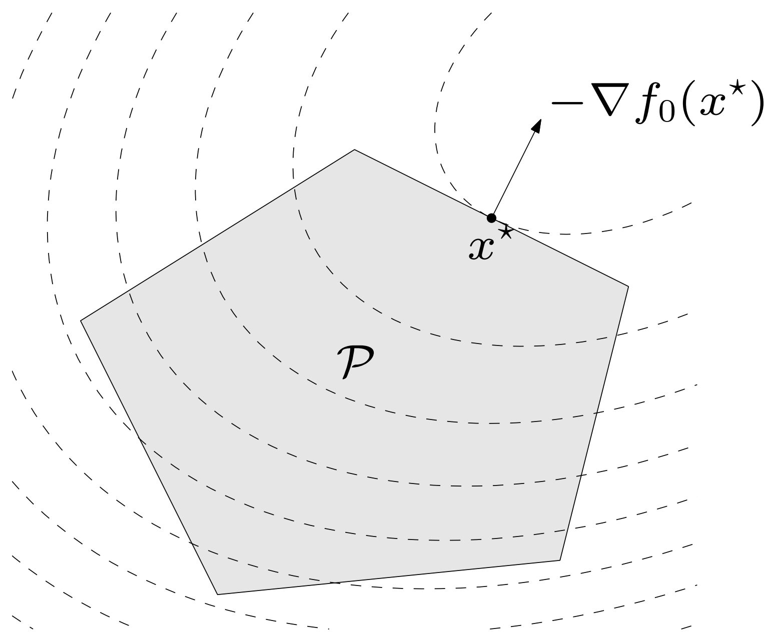
- convex problem with affine objective and constraint functions
- feasible set is a polyhedron



# Quadratic program (QP)

$$\begin{array}{ll}\text{minimize} & (1/2)x^T P x + q^T x + r \\ \text{subject to} & Gx \preceq h \\ & Ax = b\end{array}$$

- $P \in \mathbf{S}^n_+$, so objective is convex quadratic
- minimize a convex quadratic function over a polyhedron



# Euclidean projection

$\Pi_C(x_0)$: the point in $C$ closest to point $x_0$

can be computed in closed form for many useful examples

- affine set
- nonnegative orthant
- halfspace
- box
- consensus set $C = \{x \in \mathbf{R}^{Nn} \mid x_1 = x_2 = \cdots = x_N\}$

# Lagrangian

**standard form problem** (not necessarily convex)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value $p^\star$

**Lagrangian:** $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with $\mathbf{dom}\, L = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$,

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- weighted sum of objective and constraint functions
- $\lambda_i$ is Lagrange multiplier associated with $f_i(x) \leq 0$
- $\nu_i$ is Lagrange multiplier associated with $h_i(x) = 0$

# Lagrange dual function

**Lagrange dual function:** $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\
&= \inf_{x \in \mathcal{D}} \left( f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\end{aligned}
$$

$g$ is concave, can be $-\infty$ for some $\lambda$, $\nu$

**lower bound property:** if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$

# Least-norm solution of linear equations

$$\begin{array}{ll} \text{minimize} & x^T x \\ \text{subject to} & Ax = b \end{array}$$

**dual function**

- Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$
- to minimize $L$ over $x$, set gradient equal to zero:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \quad \Longrightarrow \quad x = -(1/2) A^T \nu$$

- plug in in $L$ to obtain $g$:

$$g(\nu) = L((-1/2)A^T \nu, \nu) = -\frac{1}{4} \nu^T A A^T \nu - b^T \nu$$

  a concave function of $\nu$

**lower bound property**: $p^\star \geq -(1/4)\nu^T A A^T \nu - b^T \nu$ for all $\nu$

# The dual problem

**Lagrange dual problem**

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on $p^\star$, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted $d^\star$
- $\lambda$, $\nu$ are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \textbf{dom}\, g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \textbf{dom}\, g$ explicit

# Weak and strong duality

**weak duality:** $d^\star \leq p^\star$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

**strong duality:** $d^\star = p^\star$

- does not hold in general
- (usually) holds for convex problems
- Slater's constraint qualification
  - strong duality holds for a convex problem if it's strictly feasible
  - guarantees that the dual optimum is attained (if $p^\star > -\infty$)

# Karush-Kuhn-Tucker (KKT) conditions

the following four conditions are called KKT conditions (for a problem with differentiable $f_i$, $h_i$):

1. primal feasible: $f_i(x) \leq 0$, $i = 1, \ldots, m$, $h_i(x) = 0$, $i = 1, \ldots, p$
2. dual feasible: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1, \ldots, m$
4. gradient of Lagrangian with respect to $x$ vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

# KKT conditions for convex problem

if $\tilde{x}$, $\tilde{\lambda}$, $\tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal

if **Slater's condition** is satisfied:

$x$ is optimal if and only if there exist $\lambda$, $\nu$ that satisfy KKT conditions

- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(x) = 0$ for unconstrained problem

# Duality and problem reformulations

- equivalent formulations of a problem can lead to very different duals
- reformulating the primal problem can be useful when the dual is difficult to derive, or uninteresting

**common reformulations**

- introduce new variables and equality constraints
  - consensus
- make explicit constraints implicit or vice-versa
- transform objective or constraint functions

# Unconstrained minimization

$$\text{minimize} \quad f(x)$$

- $f$ convex, continuously differentiable (hence $\mathbf{dom}\, f$ open)
- we assume optimal value $p^\star = \inf_x f(x)$ is attained (and finite)

**unconstrained minimization methods**

- produce sequence of points $x^{(k)} \in \mathbf{dom}\, f$, $k = 0, 1, \ldots$ with

$$f(x^{(k)}) \to p^\star$$

- can be interpreted as iterative methods for solving optimality condition

$$\nabla f(x^\star) = 0$$

# Gradient descent method

$$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)}), \quad \text{where } t \text{ is the step size}$$

---

**given** a starting point $x \in \mathbf{dom}\, f$.
**repeat**
1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size $t$ via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

**until** stopping criterion is satisfied.

---

- a descent method (objective decreases each iteration)
- very simple, but often very slow; rarely used in practice
- in the constrained case, can use 'projected gradient', which wraps the righthand side with Euclidean projection onto feasible set

# Optimization algorithms

- many algorithms available for different classes of problems
- reformulating the problem may make different algorithms applicable
- important to distinguish between the problem formulation and the algorithm used to solve it
- specialized vs general-purpose algorithms
  - belief propagation (inference in graphical models)
  - expectation-maximization (MLE with latent variables)
- can decide whether to solve the problem directly or via the dual
  - can make available additional problem structure
  - but the dual function is generally *nonsmooth*

# Variational methods

- the term *variational* refers generically to optimization-based methods for doing something
  - historically, comes from 'calculus of variations'
  - 'variational inference' refers to optimization-based methods to carry out inference in graphical models

- a *variational characterization* of an object is one that expresses the object as the solution to an optimization problem

  often based on this principle: a closed convex function is the pointwise supremum of all its affine underestimators

- related but different task: given an algorithm, figure out what optimization problem it is implicitly solving (if any)
  - can give a deeper understanding of the algorithm
  - $e.g.$, loopy BP, EM, boosting

# Variational methods

once a variational representation of an object is available

- design/apply different algorithms to compute the object
- approximate the object by *relaxing* the optimization problem (simplify objective/constraints)
- get bounds on the object ($e.g.$, via duality)
  - Jensen's inequality
  - Fenchel's inequality