Intro and background

2022年9月13日 13:55

Data fitting problem

- Given a set of data points (x_i, y_i), $i \in \{1,2,..,n\} = [n]$, Find a, b that defines a line $y = ax + b$ that best matches the data.
  - $(a, b) \in \mathbb{R}^2$ are optimization variables.
- Define an error function
  - $z_i = y_i - (ax_i + b)$, $i \in [n]$.
- Aim to minimize squared error.
  - $\min_{a,b} \sum_{i=1}^n (y_i - ax_i - b)^2$.
  - $\frac{\partial f}{\partial a} = \sum_{i=1}^n 2(y_i - ax_i - b)(-x_i) = 0$.
    - Simplify: $\sum_{i=1}^n x_i y_i = (\sum_{i=1}^n x_i^2)a + (\sum_{i=1}^n x_i)b$.
  - $\frac{\partial f}{\partial b} = \sum_{i=1}^n 2(y_i - ax_i - b)(-1) = 0$.
    - Simplify: $\sum_{i=1}^n y_i = (\sum_{i=1}^n x_i)a + (\sum_{i=1}^n 1)b$.
  - In matrix form
    - $\begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$.
    - If invertible, we get a unique (a^*, b^*).
- Least squares
  - Has an analytic solution
  - Convex problem
  - Quadratic form in terms of (a, b).
- Linear algebraic approach
  - $\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \dots & 1 \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} z_1 \\ \dots \\ z_n \end{pmatrix}$.
    - $y = Hv + z$.
  - We want to minimize $|z|^2 = |y - Hv|^2$.
  - $\min_v |y - Hv|^2 = \min_v (y^T y - 2y^T Hv + v^T H^T Hv)$.
  - Take derivative with respect to v.
    - $-2y^T H + 2v^T H^T H = 0$.
    - $H^T Hv = H^T y$.
    - $v^* = (H^T H)^{-1} H^T y$.
  - $(H^T H)^{-1}$ is a pseudo inverse of H.

MLE (maximum likelihood estimation) gaussian

- Gaussian noise model
  - $y_i = ax_i + b + z_i$.
  - $z_i = y_i - ax_i - b \sim iid\ N(0, \sigma^2)$.
    - i.e. $z_i \sim P_z(\zeta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\zeta^2}{2\sigma^2}\right)$.
- Problem:
  - Pick (a, b) to maximize probability of observed data.
  - $(a^*, b^*) = \operatorname{argmax} P(x, y; a, b) = \operatorname{argmax} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - ax_i - b)^2\right)$.
    - $= \operatorname{argmax} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - ax_i - b)^2\right)$.
    - $= \operatorname{argmax} \exp\left(-\frac{1}{2\sigma^2} |y - Hv|^2\right)$ (here $|y - Hv|^2$ is the l2-norm).
    - i.e. minimizing $|y - Hv|^2$.
- Unconstrained QP

MLE exp

- Model
  - $y_i = ax_i + b + z_i$.
  - $z_i = y_i - ax_i - b \sim iid$ double-sided exponentials.
    - i.e. $z_i \sim P_z(\zeta) = \frac{1}{2c}\exp\left(-\frac{1}{c}|\zeta|\right)$.
- Problem:
  - Pick (a, b) to maximize probability of observed data.
  - $(a^*, b^*) = \operatorname{argmax} P(x, y; a, b) = \operatorname{argmax} \prod_{i=1}^{n} \frac{1}{2c}\exp\left(-\frac{1}{c}|y_i - ax_i - b|\right)$.
    - $= \operatorname{argmax} \left(\frac{1}{2c}\right)^n \exp\left(-\frac{1}{c}\sum_{i=1}^{n}|y_i - ax_i - b|\right)$.
    - $= \operatorname{argmax} \exp\left(-\frac{1}{c}|y - Hv|\right)$ (here $|y - Hv|$ is the l1-norm).
- To express l1-norm as an LP, introduce auxiliary variables $(t_1, \ldots, t_n)$.
  - $\min \sum_{i=1}^{n} t_i$, such that $|y_i - ax_i - b| \le t_i$, $i \in [n]$.
  - Equivalent to $\min \sum_{i=1}^{n} t_i$, such that $y_i - ax_i - b \le t_i$, $y_i - ax_i - b \ge -t_i$.
- Single-sided exp noise
  - $P(\zeta) = \begin{cases} \frac{1}{c}\exp\left(-\frac{\zeta}{c}\right), \zeta \ge 0 \\ 0, \zeta < 0 \end{cases}$.
  - Log-likelihood $\log\left(P(\zeta)\right) = \begin{cases} const\ -\frac{\zeta}{c}, \zeta \ge 0 \\ -\infty, \zeta < 0 \end{cases}$.
  - MLE for 1-sided exp noise
    - $\min \sum_{i=1}^{n} y_i - ax_i - b$, such that $y_i - ax_i - b \ge 0$, $i \in [n]$.

MLE uniform

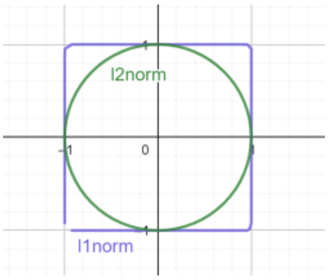
- Uniform noise
  - $P(\zeta) = \begin{cases} \frac{1}{2c}, |\zeta| \le c \\ 0, otherwise \end{cases}$.
  - $\log\left(P(\zeta)\right) = \begin{cases} const, |\zeta| \le c \\ -\infty, otherwise \end{cases}$.
- Problem
  - $\max \log\left(\prod_{i=1}^{n} P(y_i - ax_i - b)\right) = \max \sum_{i=1}^{n} \log P(y_i - ax_i - b)$.
- An ML solution is any solution that satisfies $|y_i - ax_i - b| \le c$, $\forall i \in [n]$.
- LP-feasibility

Feasibility problem

- $\min d$, such that $y_i - ax_i - b \le d$, $y_i - ax_i - b \ge -d$, $\forall i \in [n]$.
- If $d^* \le c$, then feasible. If $d^* > c$, infeasible.
- Prior on (a, b): $(a, b) \sim N\left((\mu_a, \mu_b), \Sigma\right)$.
  - Where μ are the means, Σ is the 2×2 covariance matrix.
  - Instead of $\max P(x, y; a, b)$, will $\max P(a, b \mid x, y)$.
    - Bayes: $P(a, b \mid x, y) = \frac{P(x,y \mid a,b)P(a,b)}{P(x,y)}$.
    - $P(x, y)$ is fixed by data.
    - $P(a, b)$ is the prior.
    - $P(x, y \mid a, b)$ is the likelihood of the given model.
- Reduce the problem to $\max P(a, b)$ such that (a, b) feasible.
  - $(a, b) \sim \frac{1}{2\pi \det \Sigma}\exp\left(-\frac{1}{2}(a - \mu_a, b - \mu_b)\Sigma^{-1}\begin{pmatrix} a - \mu_a \\ b - \mu_b \end{pmatrix}\right)$.
  - So, we want to minimize $(a - \mu_a, b - \mu_b)\Sigma^{-1}\begin{pmatrix} a - \mu_a \\ b - \mu_b \end{pmatrix}$.
    - Such that $y_i - ax_i - b \le c$, $y_i - ax_i - b \ge -c$.
  - This a quadratic program (QP)

Vector space

- Def: A set of elements (vectors) closed under addition and scalar multiplication.
- Normed vector space: a vector space with a notion of length of any particular vector and a measure of length or norm
- Inner product space: a normed vector space with a notion of angle between any pair of vectors specifics an inner product space
- Norm: a norm is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ such that $\forall x, y \in \mathbb{R}^n$.
  - positivity: $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$ (add identity).
  - Scaling property: $\|tx\| = |t|\|x\|$, $t \in \mathbb{R}$, $x \in \mathbb{R}^n$.
  - Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.
- examples
  - Euclidean norm: $\|x\| = \sqrt{\sum_{i=1}^{n} x_n^2}$.
  - l_p-norms, $p \geq 1$: $\|x\|_p = \left(\sum_{i=1}^{n} |x_n|^p\right)^{\frac{1}{p}}$.
    - l_1-norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$.
    - l_2-norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_n^2}$.
      - It is not only familiar from Euclidean space, but can also be induced by an inner product
    - l_∞-norm: $\|x\|_\infty = \max_{i \in [n]} |x_i|$.
    - Unit norm balls
      - Norm balls must be convex sets



- Inner product
  - $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$.
  - Angle: $\langle x, y \rangle = \|x\| \|y\| \cos\theta$.
  - x and y are orthogonal ($x \perp y$) if $\langle x, y \rangle = 0$.
- Cauchy-Schwartz inequality: $|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$.

Matrices

- Set of $m \times n$ matrices with elements from $\mathbb{R}$ is denoted as $\mathbb{R}^{m \times n}$.
- Rank of the matrix: $rank(A) = \min\{m, n\}$.
- Inner product of matrices: $X, Y \in \mathbb{R}^{m \times n}$, $\langle X, Y \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij} = tr(X^T Y)$.
  - Induces the Frobenius norm: $\|X\|_F = \sqrt{\langle X, X \rangle} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2} = \sqrt{tr(X^T X)}$.
- Matrices as transformations $\mathbb{R}^n \to \mathbb{R}^m$.
  - Range of A: $R(A) = \{Ax : x \in \mathbb{R}^n\} \subset \mathbb{R}^m$.
  - Nullspace of A: $N(A) = \{x : Ax = 0\} \subset \mathbb{R}^n$.
- Singular value decomposition (SVD)
  - $A = U \Sigma V^T$.
  - $A \in \mathbb{R}^{m \times n}$.
  - $U \in \mathbb{R}^{m \times m}$, orthogonal.
    - $U^T U = U U^T = I_m$.
    - Orthogonal means that $U^T x$ preserves the length of x.
      - $\|U^T x\|^2 = x^T U U^T x = x^T x = \|x\|^2$.
  - $\Sigma \in \mathbb{R}^{m \times n}$.
    - Rectangular matrix with singular values along the diagonal
    - Number of singular values = $rank(A) = r$.

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.
- $\begin{pmatrix} \sigma_1 & 0 & 0 & \\ 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma_r & \\ & 0 & & 0 \end{pmatrix}$.

  - $V \in \mathbb{R}^{n \times n}$, orthogonal.
    - $V^T V = V V^T = I_n$.
- Operation of A on any $x \in \mathbb{R}^n$.
  - $Ax = U \Sigma V^T x$.
  - $V^T x$ is a length-preserving rotation
  - Σ is a scaling: scale each of the first r components of $(V^T x)$ by σ_i.
  - U is again a rotation.

Symmetric matrices

- A matrix A is symmetric if $A = A^T$.
- Let S^n be the set of real symmetric matrices, $S^n \subset \mathbb{R}^{n \times n}$.
- If $A \in S^n$, can diagonalize (spectral decomposition), $A = Q \Lambda Q^T$.
  - Q is $n \times n$ orthogonal matrix.
  - $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$, where λ_i are eigenvalues of A.
  - Note: real symmetric matrices have purely real eigenvalues.
- A real symmetric matrix $A \in S^n$ is positive semi-definite (PSD) if $v^T A v \geq 0$ for all $v \in \mathbb{R}^n$ and is positve definite (PD) if $v^T A v > 0$ for all $v \in \mathbb{R}^n - \{0\}$.
  - Set of PSD: S_+^n.
  - Set of PD: S_{++}^n.
- Consider $A \in S_+^n$, can write $A = Q \Lambda Q^T$.
  - Thus $v^T A v = v^T Q \Lambda Q^T v = w^T \Lambda w = \sum_{i=1}^n \lambda_i w_i^2$.
  - Since Q is invertible, $v^T A v \geq 0$ means $w^T \Lambda w \geq 0$ for all $w \in \mathbb{R}^n$.
  - So $\sum_{i=1}^n \lambda_i w_i^2$ means all $\lambda_i \geq 0$.
  - A symmetric matrix A is PSD if and only if all its eigenvalues are non-negative.
  - A symmetric matrix A is PD if and only if all its eigenvalues are positive.

Square-root matrix (of a PSD matrix)

- $A \in S_+^n$, so $A = Q \Lambda Q^T$.
- $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$, $\Lambda^{\frac{1}{2}} = \begin{pmatrix} {\lambda_1}^{1/2} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n^{1/2} \end{pmatrix} = diag\left(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}\right)$.
- Then $A^{\frac{1}{2}} = Q \Lambda^{\frac{1}{2}} Q^T$.
  - Since $A^{\frac{1}{2}} A^{\frac{1}{2}} = Q \Lambda^{\frac{1}{2}} Q^T Q \Lambda^{\frac{1}{2}} Q^T = Q \Lambda Q^T = A$.

Partial derivative and gradients

- Let $f : \mathbb{R}^n \to \mathbb{R}$ and fix a point $x \in \mathbb{R}^n$, consider $\lim_{\alpha \to 0} \frac{f(x + \alpha e_i) - f(x)}{\alpha}$ where e_i is the ith unit vector. If the limit exists, it is called the partial derivative of f at x and is denoted $\frac{\partial f}{\partial x_i}(x)$.
- If all partial derivative exists, the gradient of f at x is $\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \dots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}$.
- Directional derivative
  - For any $y \in \mathbb{R}^n$, the one-sided directional derivative of f at $x \in \mathbb{R}^n$ is $f'(x, y) = \lim_{\alpha \to 0} \frac{f(x + \alpha y) - f(x)}{\alpha}$.
  - Gateaux differentiability: If $f'(x, y)$ exists for directions $y \in \mathbb{R}^n$ and is a linear function of y, then

f is differentiable at x.
  - A function f is differentiable at x if and only if the gradient $\nabla f(x)$ exists and satisfies $\nabla f(x)^T y = f'(x, y)$ for all $y \in \mathbb{R}^n$.
- Terminology
  - f is differentiable over a subset $U \subset \mathbb{R}^n$ if f is differentiable at all $x \in U$.
  - f is differentiable if differentiable at all $x \in \mathbb{R}^n$.
  - f is continuously differentiable over $U \subset \mathbb{R}^n$, if it is differentiable over U and ∇f is continuous over U.
  - f is smooth if it is continuous differentiable over all $\mathbb{R}^n$.
- If f is continuously differentiable over $U \subset \mathbb{R}^n$, then $\lim_{y \to 0} \frac{f(x+y) - f(x) - \nabla f(x)^T y}{\|y\|} = 0$.
  - What it means is that f can be arbitrarily well approximated by an affine function of y as $y \to 0$.
  - An alternate definition of differentiability: can you approximate the function arbitrarily well with some affine approximation (Frechet differentiability)

Little o notation
- Given 2 semi-infinite sequences $\{x_k\}, \{y_k\}$, write $x_k = o(y_k)$ if $\lim_{k \to \infty} \frac{x_k}{y_k} = 0$.
- For functions, $h(y) = o(\|y\|)$ if $\lim_{y \to 0} \frac{h(y)}{\|y\|} = 0$ for al sequences $\{y_k\}$ such that $y_k \to 0$.
- For any sequence $\{y_1, y_2, ..\}$ such that $\lim_{k \to \infty} y_k = 0$, $\lim_{y \to 0} \frac{f(x+y) - f(x) - \nabla f(x)^T y}{\|y\|} = 0$.
  - $\forall \epsilon > 0, \exists k_0$ such that $\forall k > k_0$, $\left| \frac{f(x+y_k) - f(x) - \nabla f(x)^T y_k}{\|y_k\|} \right| < \epsilon$.
  - i.e. $|f(x + y_k) - f(x) - \nabla f(x)^T y_k| < \epsilon \|y_k\|$.
  - $f(x + y_k) = f(x) + \nabla f(x)^T y_k + o(\|y_k\|)$ is the affine approximation.
  - Drop the index: $f(x + y) \approx f(x) + \nabla f(x)^T y$.

Scalar function approximation
- 1st order: $f(u) = f(u_0) + f'(u_0)(u - u_0) + o(u - u_0)$.
- 2nd order: $f(u) = f(u_0) + f'(u_0)(u - u_0) + \frac{1}{2} f''(u_0)(u - u_0)^2 + o\left((u - u_0)^2\right)$.

1st order approximation for $f : \mathbb{R}^2 \to \mathbb{R}$:
- $f(u, v) = f(u_0, v_0) + f_u(u_0, v_0)(u - u_0) + f_v(u_0, v_0)(v - v_0) + o(\|(u, v) - (u_0, v_0)\|)$.
- $f(u, v) \approx f(u_0, v_0) + f_u(u_0, v_0)(u - u_0) + f_v(u_0, v_0)(v - v_0) = f(u_0, v_0) + \nabla f(u_0, v_0)^T \begin{pmatrix} u - u_0 \\ v - v_0 \end{pmatrix}$.
  - $= f(u_0, v_0) + \left\langle \nabla f(u_0, v_0), \begin{pmatrix} u - u_0 \\ v - v_0 \end{pmatrix} \right\rangle$.

1st order approximation for $f : \mathbb{R}^n \to \mathbb{R}$:
- $f(x) = f(x_0) + \sum_{i=1}^n f_{x_i}(x_0)(x - x_0) = f(x_0) + \nabla f(x_0)^T (x - x_0) = f(x_0) + \langle \nabla f(x_0), (x - x_0) \rangle$.
  - Affine approximation $f : \mathbb{R}^n \to \mathbb{R}$.
- Rewrite $\Delta x = x - x_0$, it defines a direction, we can scale by λ to get a line $x_0 + \lambda \Delta x$, $\lambda \in \mathbb{R}$.
  - $f(x_0 + \lambda \Delta x) = f(x_0) + \lambda \left(\nabla f(x_0)^T \Delta x \right)$.
  - For any centers x_0 and directions Δx, same form as first order Taylor approximation for a scalar function, i.e. some affine $f : \mathbb{R} \to \mathbb{R}$.
  - Inner product $\nabla f(x_0)^T \Delta x$ plays role of slope.
  - λ parametrizes distance from x_0.
  - Often take Δx to be a unit vector so $||\Delta x|| = 1$.

1st order approximation for $f : \mathbb{R}^n \to \mathbb{R}^m$:
- Can see if as m mapping $f_i : \mathbb{R}^n \to \mathbb{R}$.
- $f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \cdots \\ f_n(x) \end{pmatrix} = \begin{pmatrix} f_1(x_0) \\ f_2(x_0) \\ \cdots \\ f_n(x_0) \end{pmatrix} + \begin{pmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \cdots \\ \nabla f_n(x)^T \end{pmatrix} (x - x_0) + \begin{pmatrix} o(\|x - x_0\|) \\ o(\|x - x_0\|) \\ \cdots \\ o(\|x - x_0\|) \end{pmatrix}$.

- $\begin{pmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \cdots \\ \nabla f_n(x)^T \end{pmatrix} = \begin{pmatrix} f_{1x_1}(x_0) & f_{1x_2}(x_0) & \dots & f_{1x_n}(x_0) \\ \dots & \dots & \dots & \dots \\ f_{nx_1}(x_0) & f_{nx_2}(x_0) & \dots & f_{nx_n}(x_0) \end{pmatrix} = Df(x_0) = J(x_0)$ is the derivative (Jacobian) matrix.
- $f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$.

2nd order approximation for $f : \mathbb{R}^n \to \mathbb{R}$:

- $f(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0) + o\left(\|x - x_0\|^2\right)$.
- Hessian of f at $x_0 \in \mathbb{R}^n$: $\nabla^2 f(x_0) = \begin{pmatrix} f_{x_1x_2}(x_0) & f_{x_1x_2}(x_0) & \dots & f_{x_1x_n}(x_0) \\ \dots & \dots & \dots & \dots \\ f_{x_nx_1}(x_0) & f_{x_nx_2}(x_0) & \dots & f_{x_nx_n}(x_0) \end{pmatrix}$.
  - When $n = 1$, $\nabla^2 f(x_0) = f''(x_0)$.
  - Symmetry: $\left(\nabla^2 f(x_0)\right)^T = \nabla^2 f(x_0)$, because $f_{x_ix_j}(x) = f_{x_jx_i}(x)$.
- Approximation along a line $l = \{x : x = x_0 + \lambda u, \lambda \in \mathbb{R}\}$.
  - $f(x_0 + \lambda u) = f(x_0) + \nabla f(x_0)^T (\lambda u) + \frac{1}{2}(\lambda u)^T \nabla^2 f(x_0)(\lambda u) + o\left(\|x - x_0\|^2\right)$.
  - $= f(x_0) + \lambda(\nabla f(x)^T u) + \frac{1}{2}\lambda^2 u^T \nabla^2 f(x_0) u$.
  - Along any line (choice of (x_0, u)), get familiar 2nd order Taylor.
  - Offset, the slope and the curvature all depend on x_0 and u.
  - 1st order approximation is a plane and 2nd order gives a quadratic surface

Examples of gradients

- $f(x) = \langle a, x \rangle = a^T x$, $\nabla f(x) = a$.
- $f(x) = x^T P x = \sum_{i=1}^n \sum_{j=1}^n x_i x_j P_{ij} = \sum_{i=1}^n x_i^2 P_{ii} + \dots + x_i x_j \left(P_{ij} + P_{ji}\right)$,
  - $\frac{\partial}{\partial x_k} x^T P x = 2x_k P_{kk} + 2\sum_{i<k} x_i \frac{(P_{ik}+P_{ki})}{2} = 2x_i \frac{(P_{ii}+P_{ii})}{2} + 2\sum_{i<k} x_i \frac{(P_{ik}+P_{ki})}{2}$.
  - $= \sum_{i=1}^n x_i \left(P_{ki} + (P^T)_{ki}\right)$.
  - $\nabla(x^T P x) = x^T (P + P^T) = (P + P^T) x$.
  - If P is symmetric, $\nabla(x^T P x) = 2Px$.

Chain rules:

- Gradients for compositions of functions
- $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, $h(x) = g(f(x))$.
  - $\nabla h(x) = g'(f(x)) \nabla f(x)$.
- $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$, $h(x) = g(f(x))$.
  - $\frac{\partial h(x)}{\partial x_k} = \frac{\partial g}{\partial f_1}\frac{\partial f_1}{\partial x_k} + \frac{\partial g}{\partial f_2}\frac{\partial f_2}{\partial x_k} + \dots + \frac{\partial g}{\partial f_m}\frac{\partial f_m}{\partial x_k}$.
  - $Df(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$ (Jacobian).
  - $\nabla h(x)^T = Df(x)^T \nabla g(f(x))$.
  - Function of affine function of x:
    - $f : \mathbb{R}^n \to \mathbb{R}^m$, $f(x) = Ax + b$.
    - $h(x) = g(Ax + b)$.
    - $\nabla h(x)^T = Df(x)^T \nabla g(f(x)) = A^T \nabla g(Ax + b)$.

Gradient of log det function

- $f : S^n \to \mathbb{R}$, $f(x) = \log\det X$, $dom(f) = S_{++}^n$(positive definite $\det X > 0$).
- $\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \dots & \frac{\partial f}{\partial x_{1n}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f}{\partial x_{n1}} & \frac{\partial f}{\partial x_{n2}} & \dots & \frac{\partial f}{\partial x_{nn}} \end{pmatrix}$.

- Consider $\log\det(X+\Delta X)$, $X \in S_{++}^n$, $\Delta X \in S$, $X+\Delta X \in S_{++}^n$.
- $\log\det(X+\Delta X) = \log\det\left(X^{\frac{1}{2}}\left(I + X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}\right)X^{\frac{1}{2}}\right) = \log\det\left(\left(I + X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}\right)X\right)$.
- $= \log\det X + \log\det\left(I + X^{-\frac{1}{2}}\Delta X X^{\frac{1}{2}}\right) = \log\det X + \log\left(\prod_{i=1}^n (1+\lambda_i)\right) = \log\det X + \sum_{i=1}^n \log(1+\lambda_i)$.
    - $M = X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}$, λ_i are eigenvalues of M.
- If Δx is small, then all the λ_i are small and $\log(1+\lambda_i) \approx \lambda_i$.
- Then $\log\det(X+\Delta X) \approx \log\det X + \sum_{i=1}^n \lambda_i = \log\det X + tr(M) = \log\det X + tr\left(X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}\right)$.
    - $= \log\det X + tr\left(X^{-1}\Delta X\right) = \log\det X + \langle X^{-1}, \Delta X\rangle$ (since $tr(AB) = tr(BA)$).
- This means that $\nabla f(x) = X^{-1}$.

For 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $A^{-1} = \frac{1}{\det A}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Basic concepts

September 9, 2022 8:21 PM

Mathematical program (optimization)

- Objective function $f_0 : \mathbb{R}^n \to \mathbb{R}$.
- Optimization variable $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.
- Constraint $f_i : \mathbb{R}^n \to \mathbb{R}$, $i \in \{1,2,\dots,m\} = [m]$ is the index set.
- Constrained problem
    - $\min_{x \in \mathbb{R}^n} f_0(x)$.
    - Such that $f_i(x) \le 0$, $i \in [m]$.

Solving a problem

- An optimal x denoted x^* is an x that yields smallest $f_0(x)$ among all x that satisfies constraints
- Could be unique, not unique or does not exist

Convex problems

- f_0 and f_i will be convex functions

Affine sets

- A set $C \subset \mathbb{R}^n$ is affine if $\forall x_1, x_2 \in C$, $\theta x_1 + (1-\theta)x_2 \in C$ for any $\theta \in \mathbb{R}$
- Note: can rewrite as $x_2 + \theta(x_1 - x_2)$.
    - x_2 is the offset.
    - θ is scaling.
    - $x_1 - x_2$ is direction in $\mathbb{R}^n$.
    - Is a subspace + offset
- e.g.
    - A line is an affine set
    - Solution to a set of linear equations $\{x : Ax = b\}$ is affine
- An affine combination of points $x_1, \dots, x_m$ is $\sum_{i=1}^m \theta_i x_i$ where $\theta_i \in \mathbb{R}$ and $\sum_{i=1}^m \theta_i = 1$.
- Affine hull contains all affine combinations of points in the set

Convex sets

- A set $C \in \mathbb{R}^n$ is convex if $\forall x_1, x_2 \in C$, $\forall \theta \in [0,1]$, $\theta x_1 + (1-\theta)x_2 \in C$.
- Convex combination of $x_1, \dots, x_m$ is $\sum_{i=1}^m \theta_i x_i$ where $\theta_i \ge 0$ and $\sum_{i=1}^m \theta_i = 1$.
- Convex hull of a set C is the set of all convex combinations of points in C.
    - Notation: $conv(C) = \{\sum_{i=1}^m \theta_i x_i : x_i \in C, \theta_i \ge 0, \forall i \in [m], \sum_{i=1}^m \theta_i = 1\ \forall m \in \mathbb{Z}^+\}$.

Conic sets

- A set C is a cone if $\forall x \in C$, $\theta x \in C$, $\forall \theta \ge 0$.
- Conic combination of points $x_1, \dots, x_m$ is $\sum_{i=1}^m \theta_i x_i$ with $\theta_i \ge 0$.

Hyperplanes and half-spaces

- Hyperplanes: $H = \{x : a^T x = b, a \ne 0\}$.
    - b is the offset of the subspace from origin.
    - Solution to set of linear constraint
    - Convex and affine
    - Dimension $n - 1$.
    - Other reps: $H = \{x : a^T(x - x_0) = 0\} = x_0 + a^\perp$ where $a^\perp = \{v : a^T v = 0\}$.
        - With $a^T x_0 = b$.
- Half-space: $\{x : a^T x \le b\} = \{x : a^T(x - x_0) \le 0\} = \{x : \langle a, x - x_0 \rangle \le 0\}$.
    - $a^T x_0 = b$.

Polyhedral

- $P = \left\{x : a_j^T x \leq b_j, j \in [m], c_k^T x = d_k, k \in [L]\right\} = \{x : Ax \leq b, Cx = d\}$.
- Polyhedral are convex

Balls and ellipsoids

- Euclidean balls: $B(x_c, r) = \left\{x : \|x - x_c\|_2 \leq r\right\} = \left\{x : (x - x_c)^T (x - x_c) \leq r^2\right\}$.
  - Convex
- Ellipsoids: $E(x_c, P) = \left\{x : (x - x_c)^T P^{-1} (x - x_c) \leq 1, P \in S_{++}^n\right\}$.
  - S_{++}^n is positive definite (symmetric and has spectral decomposition $P = Q\Lambda Q^T$, with $\Lambda = diag(\lambda_1, \dots, \lambda_n)$, $\lambda_i > 0$).
  - l_2-ball is ellipsoid with $P = r^2 I$.
  - $E(x_c, P)$ is the image of unit l_2-ball $B(x_c, 1)$ under affine map $f(u) = P^{\frac{1}{2}} u + x_c$.
  - Geometries
    - Consider $P = Q \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix} Q^T$,
    - ellipse is defined by $(x - x_c)^T Q \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix} Q^T (x - x_c)$.
    - $(x - x_c)^T Q$ is a projection of $x - x_c$ onto each orthonormal eigenvector of Q.
    - Let $\tilde{x} = Q^T(x - x_c)$, then $\tilde{x}\Lambda^{-1}\mathrm{x} = \sum_{i=1}^n \frac{\tilde{x}_i^2}{\lambda_i} \leq 1$.
    - Volume of the ellipsoid: $\sqrt{\det P}$.
- Unit norm ball: $\{x : \|x - x_c\| \leq 1\}$.

Cone of PSD matrices

- PSD: $S_+^n = \{x \in S^n : v^T X v \geq 0, \forall v \in \mathbb{R}^n\}$ eigenvalues are real and non-negative.
- S_+^n is a cone because if $X \in S_+^n$, $\theta X \in S_+^n$, $\forall \theta \geq 0$.
- Shorthand: $X \in S_+^n \Leftrightarrow X \geq 0, X \in S_{++}^n \Leftrightarrow X > 0$.
- S_+^n is a convex cone. Let $A, B \in S_+^n$, $\theta_1, \theta_2 \geq 0$, $\theta_1 + \theta_2 = 1$, $\theta_1 A + \theta_2 B \in S_+^n$.

Generalized inequalities

- A proper cone $K \subset \mathbb{R}^n$
  - is a closed, convex set.
  - Has a non-empty interior
  - Contains no lines (pointed)
  - e.g. half-space is a not-pointed cone
- A proper cone K defines a generalized inequalities denoted $\leq_K$ (less than or equal to w.r.t. K).
  - $x \leq_K y \Leftrightarrow (y - x) \in K$, $x <_K y \Leftrightarrow (y - x) \in int(K)$.
- For standard scalar inequality, the cone K is $K = \mathbb{R}_+ = \{x : x \geq 0\}$.

Operations that preserve convexity

- Take the (possibly infinite) intersections of sets S_α.
  - If S_α is affine for all α, then $\cap_\alpha S_\alpha$ is affine.
  - If S_α is convex for all α, then $\cap_\alpha S_\alpha$ is convex.
  - If S_α is conic for all α, then $\cap_\alpha S_\alpha$ is conic.
- Affine functions preserve convexity
  - Affine function: $f(x) = Ax + b$, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^n \to \mathbb{R}^m$.
  - If $S \subset \mathbb{R}^n$ is convex, then $f(S) = \left\{f(x) : x \in S\right\}$ is convex.
  - If $S \subset \mathbb{R}^m$ is convex, then $f^{-1}(S) = \left\{x : f(x) \in S\right\}$ is convex.
- Examples
  - A polyhedron is a convex set $P = \{x : Ax \leq b\}$ as intersections of m half spaces.
  - $\left\{y : y = Ax + b, \|x\| \leq 1\right\}$ is convex, because $\|x\| \leq 1$ is convex and $y = Ax + b$ is affine.
  - $\{x : \|Ax + b\| \leq 1\}$ is convex as pre-image of norm ball under affine map.
  - Linear matrix inequality - LMI is convex.

- $\{x \in \mathbb{R}^n : x_1 A_1 + \cdots + x_n A_n \leq B, A_i \in S^m, i \in [m], B \in S^m\}$.
- $f : \mathbb{R}^n \to S^m$ s.t. $f(x) = B - \sum_{i=1}^{n} x_i A_i$ is an affine map.
- $\{x : B - Ax \geq 0\} = \{x : B - Ax \in S_+^m\}$ and RHS is convex.
- Then pre image is convex.

Separating & hyperplanes

- Separating: if $S, T \subset \mathbb{R}^n$ are convex and disjoint ($S \cap T = \emptyset$), then there exists $a \in \mathbb{R}^n$, $a \neq 0$, $b \in \mathbb{R}$, such that $a^T x \geq b$, $\forall x \in T$, $a^T x \leq b$, $\forall x \in S$.
    - If inequalities are strict, it is a strict separating hyperplane
- Supporting: if S is convex, $\forall x_0 \in \partial S$, then there exists $a \in \mathbb{R}^n$, $a \neq 0$ such that $a^T x \leq a^T x_0$, $\forall x \in S$.

Convex functions

October 31, 2022 11:23 AM

Convex functions

- Suppose a function $f:\ \mathbb{R}^n \to \mathbb{R}$ is defined on a convex domain ($dom(F)$ is convex set), then f is a convex function if $\forall x, y \in dom(F)$, $\forall \theta \in [0,1]$, $f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$.
  - f is concave if $f(\theta x + (1-\theta)y) \ge \theta f(x) + (1-\theta) f(y)$.
  - f is strictly convex if $\forall \theta \in (0,1)$, $x \ne y$, $f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$.
  - f is strictly concave if $\forall \theta \in (0,1)$, $x \ne y$, $f(\theta x + (1-\theta)y) > \theta f(x) + (1-\theta) f(y)$.
- Remark:
  - Extended value function of a convex function is $\tilde{f}(x) = \begin{cases} f(x), if\ x \in dom(f) \\ \infty, otherwise \end{cases}$.
- Example
  - Linear and affine functions are both convex and concave
  - Parabola is convex
  - $\log x$ with $dom = \mathbb{R}_{++}$ is concave
  - $\|x\|$ is convex since $\|\theta x + (1-\theta)y\| \le \theta \|x\| + (1-\theta)\|y\|$.
  - $\frac{1}{x}$ is convex on $\mathbb{R}_{++}$, concave on $\mathbb{R}_{--}$.
- Useful facts
  - f is convex $\Rightarrow \alpha f$ is convex, for all $\alpha \ge 0$.
  - f_1, f_2 convex $\Rightarrow f_1 + f_2$ is convex over $dom(f_1) \cap dom(f_2)$.
  - If f is convex, $g(x) = f(Ax + b)$ is convex $\forall x$ such that $Ax + b \in dom(f)$, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.
  - f_1, f_2 convex $\Rightarrow \max(f_1, f_2)$ is convex.

The epigraph of a function $f: \mathbb{R}^n \to \mathbb{R}$ is $epi(f) = \{(x,t) \in \mathbb{R}^{n+1} : x \in dom(f), t \ge f(x)\}$.

- f is convex if and only if $epi(f)$ is a convex set

Sublevel set

- The sub-level set of a function $f:\ \mathbb{R}^n \to \mathbb{R}$ at level α is $C(\alpha) = \{x \in dom(f) : f(x) \le \alpha\}$.
- If f is convex, then all its sublevel sets are convex sets
  - $C(\alpha)$ is a convex set for all α.
  - Let $epi(f) = \{(x,t): x \in dom(f), t \ge f(x)\}$, $H = \{(x,t): t = \alpha\}$, $C(\alpha) = \Pi_x(epi(f) \cap H)$.
    - i.e. $C(\alpha)$ is the projection of a convex set to x.
- A function is quasi-convex if all its sublevel sets are convex sets

Super level set:

- $C(\alpha) = \{x : f(x) \ge \alpha\}$.
- A function is quasi-concave if all super level sets are convex sets.
- If f is concave, then all its super level sets are convex.

Convexity along lines

- f is convex if and only if $g(x_0 + tv)$ is convex in $t \in \mathbb{R}$, $\forall x_0 \in dom(f)$, direction $v \in \mathbb{R}^n$.
- $f(x_0 + tv)$ can be seen as $g_{x_0,v}(t)$, where x_0, v are fixed parameters.

Differentiable functions & convexity

- 1st order condition: a differentiable function f ($dom(f)$ is open and gradient exists everywhere) is convex if and only if $dom(f)$ is convex and $\forall x, y \in dom(f)$, $f(y) \ge f(x) + \nabla f(x)^T (y - x)$.
  - f is strictly convex if the inequality holds strictly $\forall x \ne y$.

- Scalar case: $f(y) \geq f(x) + f'(x)(y-x)$.
- Connection to epigraphs
  - The epigraphs must lie in the same side of a hyperplane H.
  - $H = \left\{ \begin{pmatrix} u \\ v \end{pmatrix} : (\nabla f(x)^T, -1) \begin{pmatrix} u \\ v \end{pmatrix} + \left(f(x) - \nabla f(x)^T x\right) = 0 \right\}$.
- Second order condition: a continuously twice differentiable function f is convex if and only if $dom(f)$ is convex and $\nabla^2 f(x) \geq 0$ (PSD) for all $x \in dom(f)$.
  - If $\nabla^2 f(x) > 0$ (PD), then f is strictly convex. Reverse doesn't hold
    - $f(x) = x^4$ convex, but $f''(0) = 0$.
  - Scalar case: $f''(x) \geq 0$.
- e.g.
  - $f(x) = x^\alpha$ is convex on $\mathbb{R}_+$ for $\alpha \geq 1$ or $\alpha \leq 0$.
  - $\log x$ is concave on $\mathbb{R}_{++}$.
  - $x \log x$ is convex on $\mathbb{R}_{++}$.
  - $e^{\alpha x}$ is convex $\forall \alpha$.
  - $f(x) = x^T P x + 2q^T x + r$, $P \in S^n$ is convex if $P \geq 0$, concave if $P \leq 0$.
    - $\nabla f = (P + P^T)x + 2q$, note: $x^T A x = x^T \left(\frac{1}{2}(A + A^T)\right) x$, $A \in \mathbb{R}^{n \times n}$.
    - $\nabla^2 f = 2P$.
  - $f(x,y) = x^2 + y^2 + 3xy$ is convex along any horizontal/vertical line, but not convex in general.
  - $f(x) = \sqrt{x_1 x_2}$, $\nabla^2 f \leq 0$ negative semi-definite, concave.
  - $f(x) = \max_i x_i$ is convex.
  - $f(x) = \max_{(i,j,k)} x_{[i]} + x_{[j]} + x_{[k]}$ is convex.
  - $f(x) = \sum_{i=1}^n -\log(b_i - a_i^T x)$ is convex.
  - $f(x) = \sup_{y \in C} \|x - y\|$ is convex (C doesn't have to be convex).
  - $f(x) = \inf_{y \in C} \|x - y\|$ projection onto C, not convex in general.
    - f is convex if C is convex.
  - $f(x) = \log(e^{x_1} + \cdots + e^{x_n})$ is convex on $\mathbb{R}^n$.
    - $\nabla^2 f(x) = \frac{1}{(1^T z)^2}\left((1^T z) diag(z) - z z^T\right)$ where $z = (e^{x_1}, \dots e^{x_n})$.
    - $v^T \nabla^2 f(x) v = \frac{1}{(1^T z)^2}\left(\left(\sum z_i\right)\left(\sum v_i^2 z_i\right) - \left(\sum v_i z_i\right)^2\right) \geq 0$ by Cauchy schwarz.
      - With $a_i = v_i \sqrt{z_i}$, $b_i = \sqrt{z_i}$.
      - Cauchy-schwarz: $(a^T a)(b^T b) \geq (a^T b)^2$.
    - Or from the basic definition:
      - $\theta f(x) + (1-\theta) f(y) = \theta \log \sum e^{x_i} + (1-\theta) \log \sum e^{y_i}$,
        $= \log\left(\left(\sum e^{x_i}\right)^\theta \left(\sum e^{y_i}\right)^{1-\theta}\right)$,
        $= \log\left(\left(\sum (e^{\theta x_i})^{\frac{1}{\theta}}\right)^\theta \left(\sum (e^{(1-\theta) y_i})^{\frac{1}{1-\theta}}\right)^{1-\theta}\right)$,
        $\geq \log\left(\sum e^{\theta x_i + (1-\theta) y_i}\right)$ (by Holder's inequality).
  - $f(x) = \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}$ is concave on $\mathbb{R}_{++}^n$.
    - $\nabla^2 f(x) = -\frac{\prod_{i=1}^n x_i^{\frac{1}{n}}}{n^2}\left(n\, diag(x_1^{-2}, \dots, x_n^{-2}) - q q^T\right)$ where $q_i = x_i^{-1}$.
    - $v^T \nabla^2 f(x) v = -\frac{\prod_{i=1}^n x_i^{\frac{1}{n}}}{n^2}\left(n \sum \frac{v_i^2}{x_i^2} - \left(\sum \frac{v_i}{x_i}\right)^2\right) \leq 0$.
      - $a = 1$, $b_i = v_i / x_i$.

Consequences of convexity for differentiable functions

- From 1st order condition, if $\exists x^* \in dom(f)$, such that $\nabla f(x^*) = 0$, then $f(y) \geq f(x^*)$ for any y.
  - i.e. if f convex and $\exists x^* \in dom(f)$ such that $\nabla f(x^*) = 0$, then x^* is a global minimum.
  - Converse: if x^* is a global minimizer of f and f is differentiable, then $\nabla f(x^*) = 0$.
  - Can be used for unconstrained optimization

Local optimum

- Def: x^* is a local optimum of f if $\exists \epsilon > 0$ such that $\forall x$ such that $\|x - x^*\| < \epsilon$, we have $f(x^*) \le f(x)$.
- Thm: suppose f is a twice differentiable function, then
  - If x^* is a local optimum, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x) \ge 0$.
  - If $\nabla f(x^*) = 0$ and $\nabla^2 f(x) > 0$, then x^* is a local optimum.
- e.g. $f(x) = x^3$, $f''(0) = 0$, 0 is not an optimum.

Summary

- For continuously twice differentiable functions, if x^* is a local optimum, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x) \ge 0$.
- If in addition, f is convex, i.e. $\nabla^2 f \ge 0\ \forall x \in dom(f)$, then $\nabla f(x^*) = 0$ gives x^* a global optimum.
- For convex and C^2 functions, local optimum is global optimum.

Projection

- If $h(x, y)$ is convex in $(x, y) \in \mathbb{R}^{n+p}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, then $f(x) = \inf_y h(x, y)$ is convex in x.
- e.g. $f(x) = \inf_{y \in C} \|x - y\|$ is convex if C is a convex set.

Composition of functions

- $f(x) = g(h(x))$, $h: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R} \to \mathbb{R}$, $dom(f) = \{x : g(x) \in dom(h)\}$. Then f is convex if
  - g and h are convex and h is non-decreasing.
  - g concave, h convex and h is non-increasing.
- $f(x) = g(h(x))$, $h: \mathbb{R}^n \to \mathbb{R}^k$, $g: \mathbb{R}^k \to \mathbb{R}$. Then f is convex if h_i is convex for each $i \in [k]$ (or affine), g is convex and non-decreasing in each argument.
- $f(Ax + b)$ is convex if f is convex.

Examples

- $f(x) = \exp(g(x))$ is convex if $g(x)$ is convex.
- $f(x) = \frac{1}{g(x)}$ is convex if g is concave and positive.
  - $h(w) = \frac{1}{w}$ is convex and non increasing on $\mathbb{R}_{++}$.
- $f(x) = (g(x))^p$ is convex if $p \ge 1$ and $g(x)$ is convex and positive.
  - $h(w) = w^p$ is convex and nondecreasing.
- $f(x) = -\sum_{i=1}^k \log(-f_i(x))$ is convex on $\{x : f_i(x) < 0, \forall i \in [k]\}$ if all f_i are convex.
  - $dom(f)$ is convex as intersection of convex sublevel sets.
  - $\log x$ is concave, so $-\log x$ is convex.
  - Each term in the sum is $-\log(-f_i(x))$, $g(x) = -f_i(x)$ is concave and $h(x) = -\log x$ is convex, non-increasing, thus convex.
  - Sum of convex functions is convex.
- $f(X) = \log\det(X^{-1})$ is convex where $dom(f) = S_{++}^n$, $f : S_{++}^n \to \mathbb{R}$.
  - Check along a line, let $X_0 \in S_{++}^n$, $V \in S^n$, consider $X_0 + tV$, $t \in \mathbb{R}$.
  - $\tilde{f}(t) = \log\det\left((X_0 + tV)^{-1}\right)$ is well defined as long as $X_0 + tV \in S_{++}^n$.
  - $= \log\det\left(X_0^{\frac{1}{2}}\left(I + tX_0^{-\frac{1}{2}}VX_0^{-\frac{1}{2}}\right)X_0^{\frac{1}{2}}\right)^{-1} = \log\det\left(X_0^{-\frac{1}{2}}\left(I + tX_0^{-\frac{1}{2}}VX_0^{-\frac{1}{2}}\right)^{-1}X_0^{-\frac{1}{2}}\right)$.
  - $= \log\det(X_0^{-1}) + \log\det\left(I + tX_0^{-\frac{1}{2}}VX_0^{-\frac{1}{2}}\right)^{-1}$.
    - Let $M = X_0^{-\frac{1}{2}}VX_0^{-\frac{1}{2}}$, with eigenvalues λ_i, then $I + tM$ has eigenvalues $1 + t\lambda_i$.
    - Proof: let u_i be eigenvector of M, then $(I + tM)u_i = u_i + t\lambda_i u_i = (1 + t\lambda_i)u_i$.
  - $= \log\det(X_0^{-1}) + \log\left(\prod_{i=1}^n (1 + \lambda_i t)\right)^{-1}$.
    - Since $\det A^{-1} = \frac{1}{\det A}$, $\det(X) = \prod_{i=1}^n \lambda_i$.
  - $= \log\det(X_0^{-1}) - \sum_{i=1}^n \log(1 + \lambda_i t)$.

- $1 + \lambda_i t$ is linear in t, $-\log x$ is convex, sum of convex functions is convex.

- $f(X) = (\det X)^{1/n}$ is concave on S^n_{++}.
  - $g(t) = (\det X)^{\frac{1}{n}} = (\det(Z + tV))^{\frac{1}{n}} = \left(\det\left(Z^{\frac{1}{2}}\left(I + tZ^{-\frac{1}{2}}VZ^{-\frac{1}{2}}\right)Z^{\frac{1}{2}}\right)\right)^{\frac{1}{n}}$,

    $= \left(\det Z^{\frac{1}{2}}\left(\det\left(I + tZ^{-\frac{1}{2}}VZ^{-\frac{1}{2}}\right)\right)\det Z^{\frac{1}{2}}\right)^{\frac{1}{n}}$,

    $= (\det Z)^{\frac{1}{n}}\left(\prod_{i=1}^{n}(1 + t\lambda_i)\right)^{\frac{1}{n}}$ where λ_i are eigen values of $Z^{-\frac{1}{2}}VZ^{-\frac{1}{2}}$.

- $f(X) = \lambda_{max}(X)$ (max eigenvalue of X) is convex on S^n.
  - $\lambda_{max}(X) = \sup_{\|v\| \le 1} v^T X v$.
    - By spectral decomposition $X = Q\Lambda Q^T$ with $QQ^T = Q^TQ = I$.
    - Then $v^TXv = v^TQ\Lambda Q^Tv = \tilde{v}^T\Lambda\tilde{v}$ (with $\tilde{v} = Q^Tv$, $\|\tilde{v}\| = \|v\|$).
    - $v^TXv = \tilde{v}^T\Lambda\tilde{v} = \sum_{i=1}^{n}\lambda_i(\tilde{v}_i)^2 \le \sum_{i=1}^{n}\lambda_{max}(\tilde{v}_i)^2 = \lambda_{max}$.
      - Inequality is tight, proof by checking $\tilde{v} = e_k$.
  - v^TXv is linear in X, $\sup(v^TXv)$ is convex as supremum over set of convex functions.

- $f(X) = \sigma_{max}(X)$ (largest singular value of X) is convex on $dom(f) = \mathbb{R}^{n\times m}$.
  - $\sigma_{max}(X) = \sup_{\|w\| \le 1}\|Xw\|$.
    - Consider single value decomposition $X = u\Sigma v^T = u\begin{pmatrix}\sigma_1 & 0 & 0\\ 0 & \dots & 0\\ 0 & 0 & \sigma_r\end{pmatrix}v^T$, with $r = rank(X)$, $u \in \mathbb{R}^{n\times r}$, $v \in \mathbb{R}^{m\times r}$, $u^Tu = v^Tv = I_r$.
    - $\|Xw\| = \|u\Sigma v^Tw\| = (w^Tv\Sigma^Tu^Tu\Sigma v^Tw)^{\frac{1}{2}} = (\widetilde{w}^T\Sigma^2\widetilde{w})^{\frac{1}{2}}$.
      - $u^Tu = I$, $\Sigma^T\Sigma = \Sigma^2$, let $v^Tw = \widetilde{w}$.
    - Since $\Sigma \in S^n$, $(\widetilde{w}^T\Sigma^2\widetilde{w})^{\frac{1}{2}} = \left(\sum_{i=1}^{r}\sigma_i\widetilde{w_i}^2\right)^{\frac{1}{2}} \le \left(\sum_{i=1}^{r}\sigma_{max}\widetilde{w_i}^2\right)^{\frac{1}{2}}$.
    - $= \sigma_{max}\|w\| \le \sigma_{max}$.
    - Equality can be achieved by setting w equal to max right singular vector.
      - e.g. let $\sigma_1 = \sigma_{max}$, set $w = v_1$ where $v = (v_1, v_2, \dots, v_r)$.
      - Then $\widetilde{w} = v^Tw = \begin{pmatrix}v_1^T\\ \dots\\ v_r^T\end{pmatrix}w = \begin{pmatrix}1\\ \dots\\ 0\end{pmatrix}$.
  - Since $\|\cdot\|$ is a norm, $\|(\theta X_1 + (1-\theta)X_2)w\| = \|\theta X_1w + (1-\theta)X_2w\|$.
    - $\le \theta\|X_1w\| + (1-\theta)\|X_2w\|$.
    - So $\|Xw\|$ is convex in X.
  - Supremum of a set of convex functions is convex.

Convex optimization problems

October 31, 2022 11:23 AM

Optimization problem

- Let $f_i: \mathbb{R}^{n_i} \to \mathbb{R}, i \in \{0,1,2,\dots,m\}$, $h_i: \mathbb{R}^{n_i} \to \mathbb{R}, i \in [p]$.
- Objective function: $\min_x f_0(x)$.
- Such that (Under the constraints):
  - Inequality: $f_i(x) \le 0, i \in [m]$.
  - Equality: $h_i(x) = 0, i \in [p]$.
  - Each f_i has $dom(f_i)$ and h_i has $dom(h_i)$, $x \in \cap_{i=0}^{m} dom(f_i) \cap_{j=1}^{p} dom(h_j)$.
- Feasible set: $C = \{x : f_i(x) \le 0, \forall i \in [m], h_i(x) = 0, \forall i \in [p]\}$.
- Optimal value: $f^* = \inf_{x\in C} f_0(x)$.
  - If $C = \emptyset$, $f^* = \infty$.
- Optimal point is an x^* such that $f^* = f_0(x^*)$ and $x^* \in C$.

Feasibility problems

- $f_0(x) = \begin{cases} 0, if\ x \in C \\ \infty, if\ x \notin C \end{cases}$.

Convex optimization problem

- $\min f_0(x)$.
- s.t. $f_i(x) \le 0, i \in [m]$,
- $a_i^T x + b_i = 0, i \in [p]$, equivalently $Ax = b$.
  - They are affine and not generally convex, since level sets of convex functions are generally not convex set
  - e.g. $\{x : x^2 - 1 = 0\} = \{1, -1\}$ is a level set, not convex.
- And $f_0, f_i, i \in [m]$ are all convex functions.
- $C = \cap_{i=1}^{m} \{x : f_i(x) \le 0\} \cap \cap_{i=1}^{p} \{x : Ax = b\}$ is convex.
- e.g.
  - Linear program: $\min c_0^T x + d_0$,
    - such that $c_i^T x + d_i \le 0, i \in [m]$
    - $Ax = b$.
  - $\min\|Ax - b\|$.
    - Such that $l_i \le x_i \le u_i, i \in [n]$ (box constraint).
    - $Cx = d$.

Local optimality and constrained optimality

- Def: $x \in C$ is locally optimal if $\exists \epsilon > 0$ such that $\forall y \in C$ and $\|x - y\| < \epsilon$, we have $f_0(y) \ge f_0(x)$.
- For a convex optimization problem, a local min is a global min.

Differentiable functions with constraints

- For unconstrained optimization, if can find point where $\nabla f_0(x) = 0$, then x is global minimum.
- For constrained convex optimization, if f_0 is differentiable, then $x^* \in C$ is optimal if and only if $\nabla f_0(x^*)^T (y - x^*) \ge 0$.

Quasi-convex minimization

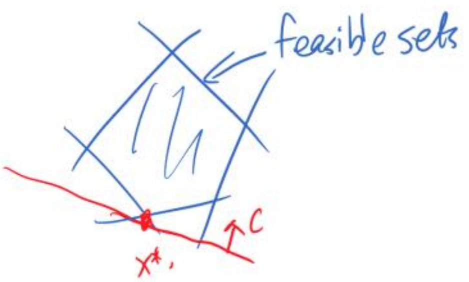
- $\min f_0(x)$ (quasi-convex, all sublevel sets are convex sets)
  - Such that $f_i(x) \le 0, i \in [m]$ (convex functions).
  - $h_i(x) = 0, i \in [p]$ (affine, $Ax = b$).
- Basic idea: introduce a surrogate function $\theta_t(x)$, such that $f_0(x) \le t \Leftrightarrow \theta_t(x) \le 0$.
- Solve a sequence (int t) of convex feasibility problems.

- $\phi_t(x) \le 0$ (for $\theta_t(x) \le 0$).
- $f_i(x) \le 0, i \in [m]$.
- $h_i(x) = 0, i \in [p]$.

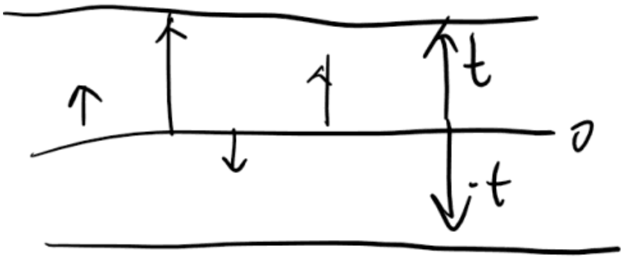
- E.g. $f_0(x) = \frac{p(x)}{q(x)}$, $p(x) \ge 0$ convex, $q(x) > 0$ concave.
    - Level sets: $\{x : f_0(x) \le t\} = \left\{x : \frac{p(x)}{q(x)} \le t\right\} = \{x : p(x) - tq(x) \le 0\}$.
    - $\phi_t(x) = p(x) - tq(x)$ is convex with $t \ge 0$.
- Linear fractional programming
    - A special case of the example above
    - $f_0(x) = \frac{a^T x + b}{c^T x + d}$, $dom(f_0) = \{x : c^T x + d > 0\}$.
    - Here $p(x) - tq(x)$ is linear in x and always convex.
- Norm optimization
    - $\min \|x\|$ s.t. $Ax = b$ (min of a convex problem, easy to solve).
    - $\max \|x\|$ s.t. $Ax = b$ (min of a concave problem, harder).
- Linear object with quadratic constraints
    - $\min c^T x$,
      s.t. $x^T P x + q^T x + r \le 0, P \ge 0$
      Is convex.
    - $\min c^T x$,
      s.t. $x^T P x + q^T x + r = 0, P \ge 0$
      Is not convex.
- Linear program
    - $\min c^T x$,
      s.t. $Ax = b$
      Is convex.
    - $\min c^T x$,
      s.t. $Ax = b, x \in \{\pm 1\}$
      Is not convex (integer program).

Linear programs (LPs)

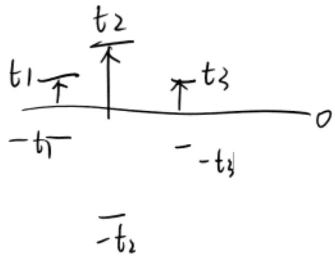
- $\min c^T x + d$, (d doesn't affect the program)
  s.t. $Gx \le h$,
  $Ax = b$.
- Affine objective, affine equality and inequality constraints
- Feasible sets are polytopes
- Level sets of objective functions are hyperplanes $\{x : c^T x + d = 0\}$.



- Problems that can be formulated as LPs
    - $\min_x \|Ax - b\|_\infty$,
      s.t. $Fx \le g$.
        - Recall $\|w\|_\infty = \max_{i \in [n]} |w_i|$.
        - $\Leftrightarrow \min_{x,t} t$,
          $Ax - b \le 1t$,
          $Ax - b \ge -1t$,
          $Fx \le g$,
          $(x, t) \in \mathbb{R}^{n+1}$, with $x \in \mathbb{R}^n$.



- $\min_x \|Ax - b\|_1$,
  s.t. $Fx \leq g$.
  - Recall $\|w\|_1 = \sum_{i=1}^{n} |w_i|$.
  - $\Leftrightarrow \min_{x,t} \sum_{i=1}^{n} t_i$,
    $Ax - b \leq t$,
    $Ax - b \geq -t$,
    $Fx \leq g$,
    $(x, t) \in \mathbb{R}^{n+m}$, with $x \in \mathbb{R}^n$, $t \in \mathbb{R}^m$.



- Fitting the largest sphere in a polytope
  - Let $P = \{x : a_i^T x \leq b_i, i \in [m]\}$, x_c =center of the sphere, r = radius of the sphere.
  - $x_C + u \in P$ means $a_i^T(x_c + u) \leq b_i$, $i \in [m]$, $\forall u$ such that $\|u\| \leq r$.
  - Look at a single constraint $a_i^T x_c + a_i^T u \leq b_i$.
    - Solve for value of case u that just satisfies the inequality.
    - Direction $\frac{a_i}{\|a_i\|}$, $u_i^* = \frac{a_i}{\|a_i\|} r$, $\|u_i^*\| = r$.
    - Need to satisfy: $a_i^T x_c + a_i^T u_i^* \leq b_i$.
    - Note: $a_i^T u_i^* = \|a_i\| r$, so the constraint is $a_i^T x_c + \|a_i\| r \leq b_i$, $i \in [m]$.
  - $\max_{x_c, r} r$,
    s.t. $a_i^T x_c + \|a_i\| r \leq b_i$, $i \in [m]$,
    $(x_c, r) \in \mathbb{R}^{n+1}$.

Quadratic program (QP)

- $\min \frac{1}{2} x^T P x + q^T x + r$,
  s.t. $Gx \leq h$,
  $Ax = b$.
- Convex if $P \geq 0$.
- Feasible set is a polytope
- 
- e.g.
  - $\min \|Ax - b\|_2^2$,
    s.t. $l_i \leq x_i \leq u_i$, $i \in [n]$ (box constraint).

- $\|Ax-b\|_2^2 = (Ax-b)^T(Ax-b) = x^TA^TAx - 2b^TAx + b^Tb$.

Linear programs with random costs (Portfolio optimization)

- Let $x = (x_1, x_2, \dots, x_n)$, where $1^Tx = 1$, x_i is partition (fraction) of portfolio invested in ith stock.
- Let $c = (c_1, c_2, \dots, c_n)$ with c_i being the return of ith stock after 1 investment period.
- Total return c^Tx.
- Don't know c_i ahead of time, but you have some idea of the distribution $c \sim N(\bar{c}, \Sigma)$.
  - $\bar{c}$ is the vector of expected returns.
  - $\Sigma = \mathrm{E}\left((c-\bar{c})(c-\bar{c})^T\right)$ is the covariance matrix.
- Expected return: $E(c^Tx) = \bar{c}^Tx$.
- Variance: $Var(c^Tx) = E\left(\left(c^Tx - \bar{c}^Tx\right)^2\right) = E\left(\left((c^T - \bar{c}^T)x\right)^2\right) = E\left(x^T(c-\bar{c})(c-\bar{c})^Tx\right) = x^TE\left((c-\bar{c})(c-\bar{c})^T\right)x = x^T\Sigma x$.
- $\min_x -\bar{c}^Tx + \gamma x^T\Sigma x$, $(\gamma \in \mathbb{R}, \gamma \geq 0)$
  s.t. $Gx \leq h$,
  $Ax = b$. (other constraints on portfolio allocation)
- $\gamma = 0$ means risk doesn't matter, larger γ means avoiding some risk.

Quadratically constrained quadratic program (QCQP)

- $\min \frac{1}{2}x^TP_0x + q_0^Tx + r_0$,
  s.t. $\frac{1}{2}x^TP_ix + q_i^Tx + r_i \leq 0, i \in [m]$,
  $Ax = b$.
- If P_0, P_i, $i \in [m]$ are PSD, then the problem is convex.

Second order cone program (SOCP)

- $\min f^Tx$,
  s.t. $\|A_ix + b_i\|_2 \leq c_i^Tx + d_i, i \in [m], A_i \in \mathbb{R}^{n_i \times n}, b_i \in \mathbb{R}^n, c_i \in \mathbb{R}^n, d_i \in \mathbb{R}$
  $Fx = g, F \in \mathbb{R}^{p \times n}, g \in \mathbb{R}^p$.
- Norm cone
  - $K = \{(x,t) : \|x\| \leq t\} \subset \mathbb{R}^{n+1}$, K is convex (from homogenity/scaling property and triangular inequality).
- Consider $f_i(x) = \begin{pmatrix} A_i \\ c_i^T \end{pmatrix} x + \begin{pmatrix} b_i \\ d_i \end{pmatrix} = \begin{pmatrix} A_ix + b_i \\ c_i^Tx + d_i \end{pmatrix} \in \mathbb{R}^{n_i+1}$, f_i is affine.
- To satisfy ith constraint, $\{x : f_i(x) \in K_i\} = f_i^{-1}(K_i)$ with K_i the norm cone with l_2 norm and $n = n_i$.
- If $A_i = 0$, LP.
- If $c_i = 0$, QCQP.

QCQP/SOCP with an analytic solution

- $\min c^Tx$,
  s.t. $x^TAx \leq 1, A \succ 0$.
- Let $y = A^{\frac{1}{2}}x$, $A = Q\Lambda Q^T$, $A^{\frac{1}{2}} = Q\Lambda^{\frac{1}{2}}Q^T$.
- $c^Tx = c^TA^{-\frac{1}{2}}A^{\frac{1}{2}}x = \tilde{c}^Ty$, with $\tilde{c} = A^{-\frac{1}{2}}c$.
- $x^TAx = x^TA^{\frac{1}{2}}A^{\frac{1}{2}}x = y^Ty = \|y\|^2 \leq 1$.
- The equivalent problem is: $\min \tilde{c}^Ty$ s.t. $\|y\|^2 \leq 1$.
- $y^* = -\frac{\tilde{c}}{\|\tilde{c}\|}$, $x^* = A^{-\frac{1}{2}}y^* = -\frac{A^{-1}c}{\|A^{-1}c\|}$.
- when $A \notin S_+^n$
  - For x^TAx to be valid, $A \in S^n$, it can be decomposed into $A = V\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}V^T = \sum_{i=1}^n \lambda_i V_i V_i^T$

- $x^T Ax = \sum_{i=1}^n \lambda_i x^T V_i V_i^T x = \sum_{i=1}^n \lambda_i y_i^T y_i$, with $y_i = V_i^T x, y = V^T x$.
- The constraint is then $\sum_{i=1}^n \lambda_i y_i^2 \leq 1$
- The objective is $\min c^T (V^T)^{-1} y = \min (V^{-1}c)^T y$

Unconstrained QPs

- $\min \frac{1}{2} x^T Px + q^T x + r$, where $P \in S^n$.
- P not PSD, objective is unbounded below
  - Take v an eigenvector of P such that $\lambda_v < 0$.
  - Look along line tv as $t \to \infty$, $\frac{1}{2} x^T Ax = \frac{1}{2}(tv)^T \lambda_v (tv) = \frac{1}{2} t^2 \lambda_v < 0$.
  - $\frac{1}{2} x^T Px + q^T x + r = \frac{t^2}{2}\lambda_v + tq^T v + r \to -\infty$.
- $P \geq 0$, problem is convex.
  - $\nabla \left(\frac{1}{2} x^T Px + q^T x + r\right) = Px + q$, if can find x^* such that $Px^* = -q$, x^* is optimal.
  - $P > 0$, P is invertible, $x^* = -P^{-1} q$ unique.
  - $P \geq 0$, but P has some zero eigen values.
    - If $q \in R(P)$ (column space of P), then can find x^* to write $Px^* = -q$, x^* is zero slope and global min, not unique.
    - If $q \notin R(P)$, unbounded below.
      - Let $q = q_\parallel + q_\perp$ with $q_\parallel \in R(P)$, $q_\perp \perp q_\parallel$.
      - Take $x = -tq_\perp$ with $t \geq 0$, $\frac{1}{2} x^T Px + q^T x + r = -t\|q_\perp\|^2 + r \to -\infty$.

Robust LP

- $\min c^T x$,
  s.t. $a_i^T \leq b_i, i \in [m]$,
  Don't know (a_i, b_i) exactly, have some uncertainty.
- Worst case (uncertainty ellipse)
  - $a_i \in E_i = \{\bar{a_i} + P_i u : \bar{a_i} \in \mathbb{R}^n, P_i \in \mathbb{R}^{n \times n}, \|u\|_2 \leq 1\}$.
  - $\Rightarrow \min c^T x, a_i^T \leq b_i, a_i \in E_i$.
  - $\Rightarrow \min c^T x, \sup_{\|u\|_2 \leq 1} (\bar{a_i} + P_i u)^T x \leq b_i$.
  - $(\bar{a_i} + P_i u)^T x = \bar{a_i}^T x + u^T P_i^T x \leq \bar{a_i}^T x + \left(\frac{P_i^T x}{\|P_i^T x\|}\right)^T P_i^T x = \bar{a_i}^T x + \|P_i^T x\|$.
  - Equivalently: $\min c^T x$, s.t. $\bar{a_i}^T x + \|P_i^T x\| \leq b_i$ (SOCP).
- Statistical approach
  - $a_i \sim N(\bar{a_i}, \Sigma)$.
  - $\min c^T x$, s.t. $\Pr(a_i^T x \leq b_i) \geq \eta$, with $\eta > \frac{1}{2}$ the level of confidence. Take $\eta = 0.95$.
  - $E[a_i^T x - b_i] = \overline{a_i^T} x - b_i \to \mu_i$.
  - $E\left[\left((a_i^T x - b_i) - E(a_i^T x - b_i)\right)^2\right] = E\left[\left(a_i^T x - \overline{a_i^T} x\right)^2\right] = x^T E\left[(a_i - \bar{a_i})(a_i - \bar{a_i})^T\right] x = x^T \Sigma x \to \sigma_i^2$.
  - With $\Phi(\gamma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\gamma} e^{-t^2/2} dt$, $\Pr\left[\frac{(a_i^T x - b_i) - \mu_i}{\sigma_i} \leq -\frac{\mu_i}{\sigma_i}\right] = \Phi\left(-\frac{\mu_i}{\sigma_i}\right) = \Phi\left(\frac{b_i - \overline{a_i^T} x}{\|\Sigma^{\frac{1}{2}} x\|}\right)$.
  - Invert Φ, $\frac{b_i - \overline{a_i^T} x}{\|\Sigma^{\frac{1}{2}} x\|} \geq \Phi^{-1}(\eta)$.
  - This gives $\min c^T x$, s.t. $b_i - \overline{a_i^T} x \geq \Phi^{-1}(\eta) \|\Sigma^{\frac{1}{2}} x\|$ (SOCP).

Least square problems

- Setup: solve system of linear equations $Ax = b$.
- If A is square invertible, $x = A^{-1} b$.
- otherwise, 2 cases for $A \in \mathbb{R}^{m \times n}$.

- Overdetermined $m > n$.
  - More constraints, fewer parameters.
  - No vector x exactly satisfies $Ax = b$.
  - Idea: find best x that most closely matches the constraints, $\min\|Ax - b\|_2^2$.
  - $f = \|Ax - b\|_2^2 = x^T A^T Ax - 2b^T Ax + b^T b$.
  - $\nabla f = 2A^T Ax - 2A^T b = 0$.
  - If $A^T A$ is invertible (A is full-rank), then $x^* = \left(A^T A\right)^{-1} A^T b$.
  - If not, we have linearly dependent columns.
- Underdetermined $m < n$.
  - More parameters, fewer constraints.
  - In general, many x satisfy $Ax = b$.
  - Assume A is full rank.
  - Idea: $\min\|x\|^2$, s.t. $Ax = b$.
  - Note: set of x that satisfy $Ax = b$, is $\{x : Ax = b\} = x_0 + N(A)$
    - x_0 is one solution $Ax_0 = b$.
    - $N(A) = \{x : Ax = 0\}$ is the null space of A.
  - Claim: $x^* = A^T\left(AA^T\right)^{-1} b$.
    - $Ax^* = b$.
    - Orthogonality: $\langle x - x^*, x^* \rangle = 0$ for $Ax = b$.
  - To calculate it, $(b - Ax)^T A = 0$ gives $A^T Ax = A^T b$.

Optimal control example

- Goals: move mass M from 0 to D in KT seconds (discretized time steps).
  - Block initially at rest, surface is frictionless
  - Want block at rest at position D at time KT.
  - $u[k]$ is a constant force applied from $t = KT$ to $t = (K + 1)T$.
  - Suppose fuel consumption is proportional to $(u[k])^2$.
- Total consumption: $\sum_{i=0}^{K-1}(u[i])^2$.
- System state: $\begin{pmatrix} x[k] \\ \dot{x}[k] \end{pmatrix}$.
  - $\begin{pmatrix} x[0] \\ \dot{x}[0] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.
  - $\begin{pmatrix} x[K] \\ \dot{x}[K] \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix}$.
- Transitions
  - $\dot{x}[k+1] = \dot{x}[k] + \ddot{x}[k]T$.
  - $x[k+1] = x[k] + \dot{x}[k]T + \frac{1}{2}\ddot{x}[k]T^2$.
  - $\ddot{x}[k] = \frac{u[k]}{M}$.
- $\begin{pmatrix} x[k+1] \\ \dot{x}[k+1] \end{pmatrix} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x[k] \\ \dot{x}[k] \end{pmatrix} + \begin{pmatrix} \frac{T^2}{2M} \\ \frac{T}{M} \end{pmatrix} u[k]$.
- So $X[k+1] = AX[k] + Bu[k]$, with $A = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} \frac{T^2}{2M} \\ \frac{T}{M} \end{pmatrix}$.
  - Using recursion, $X[K] = A^K X[0] + CU$,
  - with $C = \left[B, AB, A^2B, \ldots A^{K-1}B\right]$ and $U = \begin{pmatrix} u[K-1] \\ \ldots \\ u[0] \end{pmatrix}$.
- Problem formulation
  - $\min \sum_{i=0}^{K-1}(u[i])^2 = \|U\|^2$,
    s.t. $X = A^K X[0] + CU$.
- Optimal solution.
  - $U_{LS}^* = C^T\left(CC^T\right)^{-1}\left(X[k] - A^K X[0]\right)$.

- $C^T(CC^T)^{-1} = \begin{pmatrix} B^T \\ \vdots \\ B^T(A^T)^{K-1} \end{pmatrix} \left(\sum_{j=0}^{K-1} A^j BB^T (A^T)^j\right)^{-1}$.
- $\sum_{j=0}^{K-1} A^j BB^T (A^T)^j$ is the discrete time controllability Gremmian matrix.
- If C is not full rank, no optimal.

Geometric programs(GP)

- Monomial: $h(x) = cx_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}$, $c \geq 0$, $\alpha_j \in \mathbb{R}$, $dom(h) = \{x : x_i > 0\} = \mathbb{R}_{++}^n$.
- Posynomial: $f(x) = \sum_k c_k x_1^{\alpha_{1k}} x_2^{\alpha_{2k}} \dots x_n^{\alpha_{nk}}$, $c_k \geq 0$ is the sum of monomials.
- Problem:
  - $\min f_0(x)$,
    s.t. $f_i(x) \leq 1$, $i \in [m]$,
    $h_i(x) = 1$, $i \in [p]$.
    $f_0, f_1, \dots f_m$ all posynomials, h_i all monomials.
- Get the convex form
  - Let $y_i = \log x_i$.
  - For monomials: $\log h(x) = \log c + \alpha_1 y_1 + \dots + \alpha_n y_n$. (affine in y)
  - For posynomials: $\log f(x) = \log\left(\sum c_k e^{y_1 \alpha_{1k}} \dots e^{y_n \alpha_{nk}}\right) = \log\left(\sum e^{\sum_{i=1}^n y_i \alpha_{ik} + \beta_k}\right)$.
    - $\beta_k = \log c_k$.
    - Convex in y.
  - The problem becomes: $\min \log f_0(e^{y_1}, \dots, e^{y_n})$
    s.t. $\log f_i(e^{y_1}, \dots, e^{y_n}) \leq 0$, $i \in [m]$,
    $\log h_i(e^{y_1}, \dots, e^{y_n}) = 0$, $i \in [p]$.

Example: wireless transmission

- n transmitters $TX_1, \dots, TX_n$, n receivers $RX_1, \dots, RX_n$, mutally interferring, G_{ij} the gain between TX_i, RX_j, σ^2 receiver noise.
- Signal to interference and noise ratio: $SINR_i = \frac{P_i G_{ii}}{\sum_{j \neq i} P_j G_{ji} + \sigma^2}$.
- Rate of communication: $R_i = \log(1 + SINR_i)$.
- Type 1: $\max_{P_1, \dots, P_n} \min_i SINR_i$, s.t. $P_i \leq P_{max}$.
  - Equivalently, $\max t$ (same as $\min \frac{1}{t}$), s.t. $SINR_i \geq t$, $\forall i \in [n]$, $\frac{P_1}{P_{max}} \leq 1$, $i \in [n]$.
  - $\frac{P_i G_{ii}}{\sum_{j \neq i} P_j G_{ji} + \sigma^2} \geq t \Leftrightarrow 1 \geq \frac{\left(\sum_{j \neq i} P_j G_{ji} + \sigma^2\right) t}{P_i G_{ii}} \Leftrightarrow \left[\left(\sum_{j \neq i} P_j G_{ji}\right) (P_i G_{ii})^{-1} + \sigma^2 (P_i G_{ii})^{-1}\right] t \leq 1$.
  - The GP is: $\min \frac{1}{t}$,
    s.t. $\frac{P_1}{P_{max}} \leq 1$, $i \in [n]$,
    $\left[\left(\sum_{j \neq i} P_j G_{ji}\right) (P_i G_{ii})^{-1} + \sigma^2 (P_i G_{ii})^{-1}\right] t \leq 1$.
- Type 2: $\max \sum_{i=1}^n R_i$, s.t. $P_i \leq P_{max}$.
  - Assume high power ratio, $SINR_i > 1$, $R_i \approx \log(SINR_i)$.
  - $\arg\max \sum_{i=1}^n \log SINR_i = \arg\max \log\left(\prod_{i=1}^n \frac{P_i G_{ii}}{\sum_{j \neq i} P_j G_{ji} + \sigma^2}\right) =$
    $\arg\min \log\left(\prod_{i=1}^n \frac{\sum_{j \neq i} P_j G_{ji} + \sigma^2}{P_i G_{ii}}\right) = \arg\min \log\left[\left(\sum_{j \neq i} P_j G_{ji}\right) (P_i G_{ii})^{-1} + \sigma^2 (P_i G_{ii})^{-1}\right]$.
  - And product of posynomials is a posynomial.
  - When $R_i = \log(1 + SINR_i) = \log\left(\frac{\sum_j P_j G_{ji} + \sigma^2}{\sum_{j \neq i} P_j G_{ji} + \sigma^2}\right)$, not a GP.

Optimization with generalized inequalities

- $\min f_0(x)$, ($f_0 : \mathbb{R}^n \to \mathbb{R}$)
  s.t. $f_i(x) \preceq_{K_i} 0$, $i \in [m]$, ($f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$, K_i is a proper cone in $\mathbb{R}^{k_i}$)
  $h_i(x) = 0$, $i \in [p]$ (h_i are affine).

- f_i are K_i-convex, i.e. $f_i(\theta x + (1-\theta)y) \leq_{K_i} \theta f_i(x) + (1-\theta)f_i(y)$, $\forall \theta \in [0,1]$, $x, y \in dom(f_i)$.
  - A function is K_i-convex iff it is K_i-convex along all lines.
  - Sublevel sets are convex, hence feasible sets are convex.
  - Local optimum=global optimum.
  - Optimality condition: objective non-decreasing as move into feasible set from x^*.

Semidefinite programs (SDP)
- Special case of generalized inequalities
- $\min c^T x$,
  s.t. $f_0 + x_1 f_1 + \cdots x_n f_n \leq_{PSD} 0$, (can have many of them, $f_i \in S^m$)
  $Gx = h$.
  - Note: $f(x) = f_0 + f_1 x_1 + \cdots + f_n x_n$ is an affine function of x.
  - $\{x : f(x) \leq 0\} = \{x : -f(x) \in S_+^m\}$ the preimage of S_+^m under an affine map, thus convex.
- Standard form: $\min Tr(CZ)$,
  s.t. $Tr(A_i Z) = b_i$, $i \in [m]$
  $Z \geq 0$,
  $Z \in S^m$, $C, A_1, \ldots A_m \in S^m$.
  - To transform the above into standard form
    - Introduce slack variables, to turn $\leq$ into $=$.
    - Write each x in initial form as $x = x^+ - x^-$ where $x^+ \geq 0$, $x^- \geq 0$.
    - $\tilde{Z} = -F_0 - \sum x_i F_i$, $Z = \begin{pmatrix} \tilde{Z} & 0 & 0 \\ 0 & diag(x^+) & 0 \\ 0 & 0 & diag(x^-) \end{pmatrix}$.
    - $A_i = \begin{pmatrix} 0 & 0 & 0 \\ 0 & G_i & 0 \\ 0 & 0 & -G_i \end{pmatrix}$.
    - $C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathrm{diag}(c) & 0 \\ 0 & 0 & -\mathrm{diag}(c) \end{pmatrix}$.

Portfolio design
- $x = (x_1, \ldots, x_n)$ allocations of stocks.
- $P = (P_1, \ldots, P_n)$ expected returns.
- $\Sigma = E((x - \bar{x})(x - \bar{x})^T)$.
- If we don't know Σ exactly, what is the worst Σ for fixed investment strategy x?
- Maybe $V_{kl} \leq \Sigma_{kl} \leq U_{kl}$, $k \in [n]$, $l \in [n]$.
- $\max x^T \Sigma x$,
  s.t. $V_{kl} \leq \Sigma_{kl} \leq U_{kl}$, $k, l \in [n]$,
  $\Sigma \geq 0$.
- $x^T \Sigma x = tr(x^T \Sigma x) = tr(\Sigma x x^T)$, so this is a SDP.

Relaxation of homogeneous QCQPs
- $\min x^T P_0 x + q_0^T x + r_0$,
  s.t. $\frac{1}{2} x^T P_i x + q_i^T x + r_i \leq 0$, $i \in [m]$.
- Convex if $P_i \geq 0$, $\forall i$.
- Homogeneous means $q_i = 0$, $\forall i$.
- Problem non-convex if any P_i not PSD, or if replace $\leq$ with $=$.

e.g. $\min x^T C x$,
s.t. $x^T F_i x \geq g_i$, $i \in [m]$, if F_i not negative semidefinite, then not convex.
$x^T H_i x = l_i$, $i \in [p]$, not convex.
- $x^T C x = tr(x^T C x) = tr(C x x^T) = tr(CX)$, with $X = x x^T$.
  - $rank(X) = 1$, $X \geq 0$.

- Equivalently, $\min tr(CX)$,
  s.t. $tr(F_i X) \geq g_i$, $i \in [m]$,
  $tr(H_i X) = l_i$, $i \in [p]$,
  $rank(X) = 1$, $X \geq 0$.
  - Linear constraints
  - The only non-convex constraint is $rank(X) = 1$.

SDP relaxation:

- drop the only non-convex constraint ($rank(X) = 1$) to get a convex optimization problem
- Objective value may be lower
- Now can compute some X^* for relaxed problem. Hope it tells something about solution to original problem
- Calculate the $rank(1)$ approximation to X^* using SVD

e.g. two way partitioning problem

- Setup: n items, partition into 2 sets
- Costs: W_{ij} cost/utility of $i, j \in [n]$ being in the same partition.
  - $-W_{ij}$ is the cost if they are in different partition.
  - $W_{ij} = W_{ji}$.
- Problem: $\min x^T W x$,
  s.t. $x_i \in \{-1,1\}$, $i \in [n] \Leftrightarrow x_i^2 = 1$ (non-convex).
  $x^T W x = \sum_{i,j} x_i x_j W_{ij}$.
- Equivalently:
  - $\min tr(WX)$ s.t. $X_{ii} = 1$, $i \in [n]$, $X \geq 0$, $rank(X) = 1$.
- Relax $rank(X) = 1$ to get SDP

Duality Theory

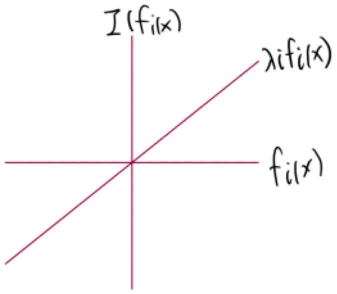
September 12, 2022 12:56 PM

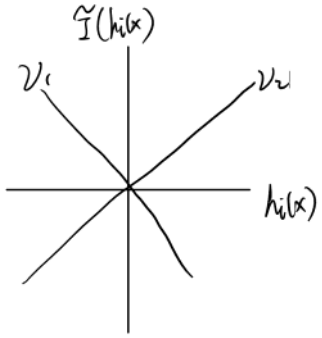
Start with a (not necessarily convex) optimization problem in standard form
$\min f_0(x)$,
s.t. $f_i(x) \leq 0, i \in [m]$,
$h_i(x) = 0, i \in [p]$.
With optimal value p^*, optimal variables x^*, x is called the primal variables.
Domain $D = \left(\cap_{i=0}^{m} dom(f_i)\right) \cap \left(\cap_{i=0}^{p} dom(h_i)\right)$.

The Lagrangian function: $L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$, where λ_i, ν_i are the Lagrange multipliers or dual variables, $dom(L) = D \times \mathbb{R}^m \times \mathbb{R}^p$.

The dual function $g(\lambda, \nu) = \min_x L(x, \lambda, \nu)$
- x may be feasible or infeasible.
- Minimization removes dependency on x.

Dual optimal problem
- $\max_{\lambda,\nu} g(\lambda, \nu)$,
  s.t. $\lambda \geq 0$.
- Dual optimum: d^*, optimal variables λ^*, ν^*, (λ, ν) are dual variables.
- $g(\lambda, \nu)$ is concave in λ, ν even if the original f is not convex and h_i is not affine.
  - $g(\lambda, \nu) = \min_x\{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\} = \min\{affine\ functions\}$ is thus concave.
- $g(\lambda, \nu) \leq f_0(x)$ if (a) x is primal feasible and (b) (λ, ν) is dual feasible.
  - Set of points satisfying (a)(b) are $\{x \in D: f_i(x) \leq 0, h_i = 0\} \times \{\lambda, \nu: \lambda_i \geq 0\}$.
  - $f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) = f_0(x) + negative < f_0(x)$ for x primal feasible, $\lambda > 0$.
  - $g(\lambda, \nu) = \min\{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\} \leq f_0(x)$.
  - Remarks
    - $f_0(x) \geq g(\lambda, \nu)$ for primal feasible x and dual feasible (λ, ν). i.e. dual problem provides a lower bound.
    - Best lower bound is to $\max g(\lambda, \nu)$, s.t. $\lambda \geq 0$.
    - Bound holds for x^*, i.e. $p^* = f_0(x^*) \geq g(\lambda^*, \nu^*) = d^*$.
- For primal and dual feasible (x, λ, ν), $f_0(x) - g(\lambda, \nu)$ is the duality gap.
  - Weak duality: $p^* - d^* \geq 0$.
- Strong duality: for convex optimization problems (f_i convex, h_i affine) and under certain constriant qualification conditions (not all possible constraints are allowed), then $p^* - d^* = 0$.
  - Convexity + constraint qualification is sufficient condition for duality to hold, but not necessary conditions
- Pricing interpretation
  - $\min f_0(x)$,
    s.t. $f_i(x) \leq 0, i \in [m]$,
    $h_i(x) = 0, i \in [p]$.
  - Reformulate as an unconstrained problem using two penalty functions I and $\tilde{I}$.
    - $I(x) = \begin{cases} 0, if\ x \leq 0 \\ \infty, else \end{cases}$.
    - $\tilde{I}(x) = \begin{cases} 0, if\ x = 0 \\ \infty, else \end{cases}$.
    - $\min f_0(x) + \sum_{i=1}^{m} I\left(f_i(x)\right) + \sum_{i=1}^{p} \tilde{I}\left(h_i(x)\right)$.
    - Note: this is not nice mathematically.
  - Basic idea in Lagrange duality is to relax I and $\tilde{I}$ to make it mathematically nice.

- 

- 

    - $\lambda_i f_i(x)$ gives a lower bound for $I\big(f_i(x)\big)$, $\nu_i h_i(x)$ gives a lower bound for $\tilde{I}\big(h_i(x)\big)$.
    - So $\min f_0(x) + \sum_{i=1}^m I\big(f_i(x)\big) + \sum_{i=1}^p \tilde{I}\big(h_i(x)\big) \geq \min f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)$.

Dual problem

- Start with original problem
  $\min f_0(x)$,
  s.t. $f_i(x) \leq 0$, $h_i(x) = 0$.
- Replace with lower bound, $\min_x \big(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)\big) = L(x, \lambda, \nu)$.
- Solve for dual function $g(\lambda, \nu) = \min_{x \in D} L(x, \lambda, \nu)$,
    - Provides lower bound on any primal feasible x if (λ, ν) dual feasible
- Maximize lower bound for all dual feasible (λ, ν).
    - $\max g(\lambda, \nu)$,
      s.t. $\lambda \geq 0$.

Remarks

- Can consider λ_i and ν_i for violating constraints (cost per unit violation)
- In $L(x, \lambda, \nu)$ are allowed to consider non-primal feasible $x \in D$ and pay linearly
- In problem for which strong duality holds, can replace I and $\tilde{I}$ with linear bounds as long as set λ_i^* and ν_i^* correctly

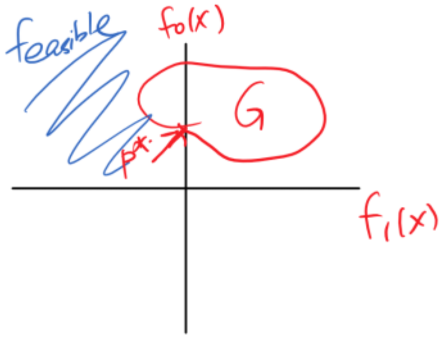
Slater's conditions

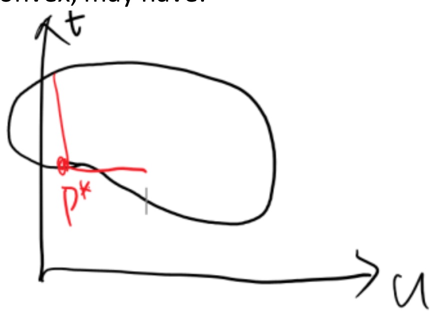
- Thm: a set of constraints $f_i(x) \leq 0$, $i \in [m]$, $Ax = b$ satisfies Slater's conditions if $\exists x \in D$ such that $f_i(x) < 0$, $i \in [m]$ and $Ax = b$.
- e.g. convex constraints not satisfying Slater
    - $(x_1 + 1, y_1) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 + 1 \\ y_1 \end{pmatrix} \leq 1$.
    - $(x_1 - 2, y_1) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ y_1 \end{pmatrix} \leq 4$.
    - Intersection has no interior
- If we have affine inequality constraints $f_i(x) = g_i^T x + h_i \leq 0$, we only need to satisfy with equality, not necessarily strict (not part of Slaters)

Strong duality

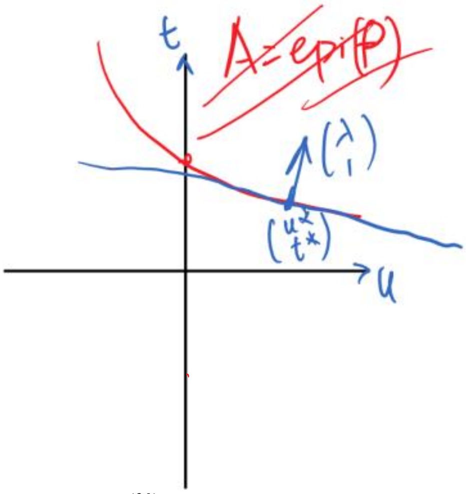
- Thm: if primal optimization problem is convex and Slater's conditions are satisfied, then $p^* = d^*$. (i.e. duality gap is 0)
- Consider only a single inequality constraint
    - Primal: $\min f_0(x)$, s.t. $f_1(x) \leq 0$ with optimal p^*.

- Lagrangian: $L(x,\lambda) = f_0(x) + \lambda f_1(x)$.
- Dual function: $g(\lambda) = \min_{x\in D} L(x,\lambda) = \min_{x\in D}\big(f_0(x) + \lambda f_1(x)\big)$.
- Dual problem: $\max g(\lambda)$, s.t. $\lambda \geq 0$, with optimal d^*.

- Resource tradeoff: $G = \cup_{x\in D}\{(f_1(x), f_0(x))\}$.
- Shadow of G (solutions dominated by G): $A = G + \mathbb{R}_+^2 = \{(u,t) : u \geq f_1(x), t \geq f_0(x), x \in D\}$.
  - 

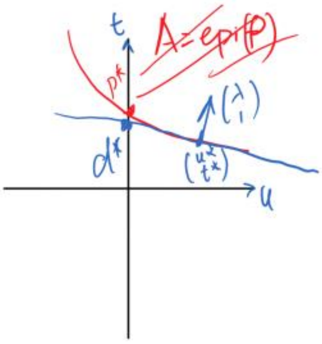
  - Boundary of corresponds to set of interesting designs
  - A contains both feasible and infeasible designs.
  - Boundary of A is some function $p(u)$
    - $p(u) = \min f_0(x)$,
      s.t. $f_1(x) \leq u$.
  - Note: $p^* = p(0)$ by definition
  - Will show if f_0, f_1 convex,
    - p is non-increasing in u.
    - p is convex, implying $A = epi(p)$ is convex.
  - If nonconvex, may have:
    - 

- Prove convexity of p.
  - Thm: p is convex, i.e. $\forall u_1, u_2 \in dom(p)$, $\lambda \in [0,1]$, $p(\lambda u_1 + (1-\lambda)u_2) \leq \lambda p(u_1) + (1-\lambda)p(u_2)$.
  - Setup:
    - $p(u_1) = \min_x f_0(x)$, s.t. $f_1(x) \leq u$.
    - $x_1 = \arg\min_x f_0(x)$, s.t. $f_1(x) \leq u$, i.e. $f_0(x_1) = p(u_1)$.
    - Similarly, let x_2 be $f_0(x_2) = p(u_2)$.
    - Look at $\tilde{x} = \lambda x_1 + (1-\lambda)x_2$, $\lambda \in [0,1]$.
    - $f_1(\tilde{x}) = f_1(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f_1(x) + (1-\lambda) f_2(x)$ (convexity of f_1),
      $\leq \lambda u_1 + (1-\lambda)u_2$ (since x_1 feasible for $p(u_1)$).
  - $p(\lambda u_1 + (1-\lambda)u_2) = f_0(\tilde{x}) \leq \lambda f_0(x) + (1-\lambda) f_0(x) = \lambda p(u_1) + (1-\lambda)p(u_2)$.
- Consider the following optimization problem with $\lambda \geq 0$.

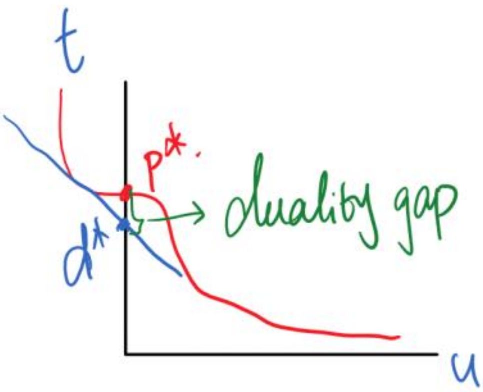
- 

- $\min_{(u,t)} (\lambda, 1) \binom{u}{t}$, s.t. $(u, t) \in A$, A convex.
- Let $(\lambda, 1) \binom{u}{t} = const$, then $t = const - \lambda u$.
- Optimum point is on boundary, corresponds to some $x^*(\lambda)$ s.t. $(u^*, t^*) = \big(f_1(x^*(\lambda)), f_0(x^*(\lambda))\big)$.
- All other points are no better
  - $(\lambda, 1) \binom{u^*}{t^*} \le (\lambda, 1) \binom{u}{t}$ for all $\binom{u}{t} \in A$.
  - $(\lambda, 1) \binom{u - u^*}{u - t^*} \ge 0$.
- i.e. (u^*, t^*) defines a supporting hyperplane of $epi(p) = t$, touches at point (u^*, t^*).
- This is non-vertical, since $(\lambda, 1)$ cannot be horizontal, unless $\lambda \to \infty$.
- Tangent point: $\big(f_1(x^*(\lambda)), f_0(x^*(\lambda))\big)$.
- Extrapolate back to get y-intercept, $\big(0, f_0(x^*(\lambda)) + \lambda f_1(x^*(\lambda))\big)$.

- Connection to dual
  - $\min_{t,u} (\lambda, 1) \binom{u}{t}$, s.t. $(u, t) \in A$.
  - $= t^* + \lambda u^* = f_0(x^*) + \lambda f_1(x^*) = \min_{x \in D} f_0(x) + \lambda f_1(x) = g(\lambda)$ (dual function).
  - Dual optimal: $d^* = \max_\lambda g(\lambda)$, $\lambda \ge 0$.
    - Maximize y-intercept to get as close to p^* as possible



  - If non-convex, A might not be a convex set.



- If Slater's condition doesn't hold, the supporting hyperplane at p^* may be vertical.

Sensitivity analysis

- Consider the problem $p^*(u,v) = \min f_0(x)$,
  s.t. $f_i(x) \le u_i$, $i \in [m]$, ($u_i < 0$ tighten constraint, $u_i > 0$ relax constraint)
  $h_i(x) = v_i$, $i \in [p]$ ($v_i \neq 0$, switch operating point).
- This is the generalization of $p(u)$ function.
  - $p^*(0,0) = p^*$ is the primal optimal value for unperturbed problem.
- Assume convex optimization satisfying Slater's.
  - $p^*(0,0) = g(\lambda^*, v^*)$ by strong duality.
    $= \min_x L(x, \lambda^*, v^*)$, p^* achieved at some x^*, λ^*, v^*,
    $\le L(x, \lambda^*, v^*)$ for any $x \in D$. Furthermore, pick x primal feasible for perturbed problem.
    $= f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p v_i^* h_i(x)$.
    $\le f_0(x) + \sum_{i=1}^m \lambda_i^* u_i + \sum_{i=1}^p v_i^* v_i$, since x is feasible for perturbed problem and $\lambda_i^* \ge 0$.
    $= f_0(x) + (\lambda^*)^T u + (v^*)^T v$.
- Also holds for $x \in D$, optimal for perturbed problem for which $f_0(x) = p^*(u,v)$.
- $p^*(u,v) \ge p^*(0,0) - (\lambda^*)^T u - (v^*)^T v$.
- If $\lambda^* \gg 1$, a small change in constraint changes the optimality greatly.

Lagrange method

- $\min f_0(x)$,
  s.t. $f_i(x) \le 0, i \in [m]$.
- Steps
  - Form Lagrangian, $L(x,\lambda) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)$.
  - Find dual, $g(\lambda) = \min_x L(x,\lambda)$.
  - Find $\lambda^* = \arg\min_{\lambda \ge 0} g(\lambda)$.
  - Recover x^* (primal optimal) using $L(x,\lambda^*)$ by finding x to minimize $L(x,\lambda^*)$.
- Remarks
  - Attractive framework if there exists structure in dual problem that makes it easy to solve (λ^*, v^*) for numerically or analytically.
  - Given λ^*, the x that minimizes $L(x,\lambda^*)$ may not be unique when $p(u)$ is convex but not strictly convex.

Lagrange method for least squares.

- $\min \|x\|^2$,
  s.t. $Ax = b$, $A \in \mathbb{R}^{m \times n}$, $m < n$, underdetermined.
- $x^* = A^T(AA^T)^{-1}b$.
- $L(x,v) = \|x\|^2 + \sum_{i=1}^m v_i(a_i^T x - b_i) = \|x\|^2 + v^T(Ax - b)$.
- $g(v) = \min_x(\|x\|^2 + v^T(Ax - b))$,
  - $\frac{\partial L}{\partial x} = 2x + A^T v = 0$ gives $x = -\frac{1}{2}A^T v$.
- $g(v) = \frac{1}{4}\|A^T v\|^2 - \frac{1}{2}v^T AA^T v - v^T b = -\frac{1}{4}v^T AA^T v - v^T b$,
  - $g'(v) = -\frac{1}{2}AA^T v - b$, $v^* = \arg\max g(v) = -2(AA^T)^{-1}b$.
  - $x^* = -\frac{1}{2}A^T v^* = A^T(AA^T)^{-1}b$.
- Consider the dual problem
  - $\max_v g(v) = \min_v \left(\frac{1}{4}v^T AA^T v + v^T b\right)$.
  - Equivalently, $\min \left\|\frac{1}{2}A^T v + x_0\right\|_2^2$ where $Ax_0 = b$ (overdetermined).
  - $\left\|\frac{1}{2}A^T v + x_0\right\|_2^2 = \left(\frac{1}{2}A^T v + x_0\right)^T \left(\frac{1}{2}A^T v + x_0\right) = \frac{1}{4}v^T AA^T v + v^T b + const$.
  - Note: no constraints in over determined dual problem.
  - Re-express: $\min \|y\|^2$ s.t. $y = \frac{1}{2}A^T x + x_0$.
- Dual of the dual
  - Lagrangian $L(x,y,v) = y^T y + v^T\left(\frac{1}{2}A^T x + x_0 - y\right)$.

- $g(v) = \min_{x,y} L(x, y, v)$.
  - $\frac{\partial L}{\partial x} = \frac{1}{2}Av$, hence $g(v) = -\infty$, unless $\frac{1}{2}Av = 0$.
  - If $\frac{1}{2}Av = 0$, $L(x, y, v) = y^T y + v^T x_0 - v^T y$, $\frac{\partial L}{\partial y} = 2y - v$, so $y = \frac{1}{2}v$.
- $g(v) = \begin{cases} -\infty, if \frac{1}{2}Av \neq 0 \\ -\frac{1}{4}v^T v + v^T x_0, if \frac{1}{2}Av = 0 \end{cases}$.
- Dual problem: $\max_v -\frac{1}{4}v^T v + v^T x_0$ s.t. $Av = 0$.
- $\Leftrightarrow \min_v \frac{1}{4}(v - 2x_0)^T (v - 2x_0) - x_0^T x_0$, s.t. $Av = 0$.
- $\Leftrightarrow \min_v \frac{1}{4}\|v - 2x_0\|^2$, s.t. $Av = 0$.
- Let $z = v - 2x_0$, $Av = Az + 2b$, so $\min \left\|\frac{z}{2}\right\|^2$, s.t. $Az = -2b$.
- Let $\tilde{z} = -\frac{z}{2}$, $\min\|\tilde{z}\|^2$, s.t. $Az = b$.
- Dual of the dual is the primal for convex problems

Duals of LPs

- $\min c^T x$, s.t. $Ax \leq b$.
- Lagrangian: $L(x, \lambda) = c^T x + \lambda^T (Ax - b) = -\lambda^T b + (c^T + \lambda^T A)x$.
- Dual function: $g(\lambda) = \begin{cases} -\infty, if\ c^T + \lambda^T A \neq 0 \\ -\lambda^T b, c^T + \lambda^T A = 0 \end{cases}$.
- Dual problem:

  $\max -\lambda^T b$,

  s.t. $\lambda \geq 0$, $c^T + \lambda^T A = 0$.
- Dual of an LP is an LP
- LP satisfies Slater's so strong duality holds
-

	#variables	# constraints
Primal	$\dim(x)$	$\dim(b)$
Dual	$\dim(b)$	$\dim(b) + \dim(x)$

- Dual of dual
  - Rewrite $\min \lambda^T b$, s.t. $A^T \lambda = -c$, $-\lambda \leq 0$.
  - $L(\lambda, z, y) = \lambda^T b - z^T \lambda + y^T (A^T \lambda + c) = (b^T - z^T + y^T A^T)\lambda + y^T c$.
  - $g(z, y) = \min_\lambda (b^T - z^T + y^T A^T)\lambda + y^T c = \begin{cases} -\infty, b^T - z^T + y^T A^T \neq 0 \\ y^T c, b^T - z^T + y^T A^T = 0 \end{cases}$.

    Dual: $\max c^T y$,

    s.t. $Ay + b - z = 0$, $z \geq 0$.
  - Equivalently: let $x = -y$

    $\min_x c^T x$,

    s.t. $Ax \leq b$.

Game theory

- zero sum game with linear payout
- Player 1 (P_1) plays $i \in [n]$, wants to minimize P_{ij}.
- Player 2 (P_2) plays $j \in [m]$, wants to maximize P_{ij}.
- Randomized strategies are allowed
  - P_1 plays i with probability u_i.
  - P_2 plays j with probability v_j.
  - Average payout: $\sum_i \sum_j u_i P_{ij} v_j = u^T P v$.
- Suppose P_1 goes first, its strategy u is known by P_2, what strategy should P_2 use?

  $\max_v u^T P v$,

  s.t. $1^T v = 1$, $v \geq 0$.
  - Note: $(u^T P)v$ is simply selecting j element for $P^T u$, $\max_{j \in [m]} [P^T u]_j$.
  - Knowing P_2 will do this, P_1 should choose u to minimize this.

$\min_u \max_{j\in[m]} [P^T u]_j$,
s.t. $1^T u = 1, u \geq 0$.

- Equivalently.
  $\min t$,
  $P^T u \leq t\mathbf{1}$,
  $1^T u = 1$,
  $u \geq 0$ (1).

- Conversely, P_2 goes first, P_1 wants to minimize the cost.
  $\min_u u^T P v$,
  s.t. $1^T u = 1, u \geq 0$.
  - Knowing P_1 will do this, P_2 should choose v to maximize this.
    $\max_v \min_{i\in[n]} [Pv]_i$,
    s.t. $1^T v = 1, v \geq 0$.
  - Equivalently.
    $\max t$,
    $Pv \geq t\mathbf{1}$,
    $1^T v = 1$,
    $v \geq 0$ (2).
- Note: $\min_u \max_v f(u,v) \geq \max_v \min_u f(u,v)$.
  - Always have 2nd mover (inner) advantage.
  - So (1)≥(2).
- Here (1)=(2) since (1) is the dual of (2).
  - Lagrangian of (1): $L(t,u,\lambda,\mu,\nu) = t + \lambda^T(P^T u - t1) - \mu^T u + \nu(1 - 1^T u)$.
  - Dual of (1): $g(\lambda,\mu,\nu) = \begin{cases} -\infty, 1 - \lambda^T 1 \neq 0 \ or \ (P\lambda - \mu - 1\nu) \neq 0 \\ \nu, else \end{cases}$.
  - Dual problem
    $\max \nu$,
    s.t. $\lambda \geq 0, \mu \geq 0$,
    $1^T \lambda = 1$,
    $P\lambda - \mu - 1\nu = 0$.
  - Equivalently,
    $\max \nu$,
    s.t. $\lambda \geq 0$,
    $1^T \lambda = 1$,
    $P\lambda \geq 1\nu$.
- Note: helped us that inner optimization had explicit solution (select largest/smallest entry)

Constrained game theory

- Strategy of P_1 constrained to $Au \leq b$.
- Strategy of P_2 constrained to $Fv \leq g$.
- If P_1 goes first, P_2 will $\max_v u^T P v$, s.t. $Fv \leq g$.
  P_1 solves: $\min_u \max_v u^T P v$,
  s.t. $Au \leq b, Fv \leq g$.
- If P_2 goes first, P_1 will $\min_u u^T P v$, s.t. $Au \leq b$.
  P_2 solves $\max_v \min_u u^T P v$,
  s.t. $Au \leq b, Fv \leq g$.
- Dualize $\max_v u^T P v$ to get a min problem for P_1.
- Dualize $\min_u u^T P v$ to get a min problem for P_2.
- Then show the min problem is the dual of the max problem

Dualize l_1-norm:

- $\min\|x\|$ s.t. $Ax = b$.
- Equivalently: $\min \sum_{i=1}^n t_i$, s.t. $x_i \leq t_i$, $x_i \geq -t_i$, $Ax = b$.
  - $\min[0^T, 1^T]\binom{x}{t}$,

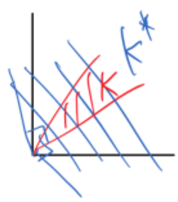
s.t. $\begin{pmatrix} I & -I \\ -I & -I \end{pmatrix} \le 0$,
$Ax = b$.

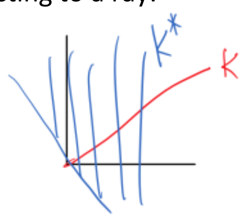
- $L(x,t,\lambda,\nu) = \sum_{i=1}^{n} t_i + \lambda^T \begin{pmatrix} I & -I \\ -I & -I \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} + \nu^T(Ax - b)$.
- $g(\lambda,\nu) = \inf_{x,t} \left(\left([0^T, 1^T] + \lambda^T \begin{pmatrix} I & -I \\ -I & -I \end{pmatrix} + (\nu^T A, 0^T) \right) \begin{pmatrix} x \\ t \end{pmatrix} - \nu^T b \right) = \begin{cases} -\nu^T b, if multiplier\ is\ 0 \\ \infty \end{cases}$.
- Let $\lambda^T = (\lambda_x^T, \lambda_t^T)$, $\lambda^T \begin{pmatrix} I & -I \\ -I & -I \end{pmatrix} = (\lambda_x^T - \lambda_t^T, -\lambda_x^T - \lambda_t^T)$.
- Dual problem
  - $\min \nu^T b$,
    s.t. $\lambda_x^T - \lambda_t^T + \nu^T A = 0$,
    $1^T - \lambda_x^T - \lambda_t^T = 0$,
    $\lambda_x \ge 0$, $\lambda_t \ge 0$.
  - Final two lines give $\lambda_{x_i} \in [0,1]$, $\lambda_{t_i} \in [0,1]$, box constraints.
  - Combining all constraints $\nu^T A = \lambda_t^T - \lambda_x^T = 2\lambda_t^T - 1 \in [-1,1]$ (l_∞norm).
  - $\min \nu^T b$,
    s.t. $\|\nu^T A\|_\infty \le 1$.
- Dual of l_p is l_q where $\frac{1}{p} + \frac{1}{q} = 1$.

Generalized inequalities

- $f_i(x) \le_K 0$ where $f_i: \mathbb{R}^n \to \mathbb{R}^m$ and $K \subset \mathbb{R}^m$.
  - $f_i(x) \le 0$ where $f_i: \mathbb{R}^n \to \mathbb{R}$ is a special case $\begin{pmatrix} f_i(x) \\ \vdots \\ f_m(x) \end{pmatrix} \le_{\mathbb{R}_+^m} 0$.
- K is a proper cone if it is pointed, convex, non-empty and closed.
- $X \le_K Y$ if $X - Y \in K$.
- For SDP, $K = S_+^m$, $x, y \in S^m$.
  - $\min c^T x$,
    s.t. $x_1 F_1 + x_2 F_2 + \cdots x_n F_n + G \le_{S_+^m} 0$,
    $Ax = b$.
  - $F_1, \ldots, F_n, G \in S^m$.

Dualizing generalized inequalities

- Key idea of dualization: $\sum \lambda_i f_i(\tilde{x}) = \langle \lambda, f(\tilde{x}) \rangle \le 0$, $\tilde{x}$ primal feasible, λ dual feasible.
- For generalized inequalities, need to identify some set that restricts dual variables to keep $\langle \lambda, f(x) \rangle \le 0$ for all x feasible ($f(x) \le_K 0$).
- Idea: if primal feasibility constraints defined by cone K, the dual variables will need to be constrained to dual cone K^*.
- Def: Let K be a cone. The set $K^* = \{Y : \langle X, Y \rangle \ge 0, \forall X \in K\}$ is the dual cone.
  - e.g.
    - 

    - Restricting to a ray:
      - 

    - $K = \mathbb{R}_+^2$, then $K^* = \mathbb{R}_+^2$ (self-dual).

        - When $\alpha > 90$, K^* reduces to 0.
    - To show K^* is cone.
        - Take $Y \in K^*$, $\forall \alpha \geq 0, X \in K^*$, $\langle X, Y\rangle \geq 0$, $\langle X, \alpha Y\rangle = \alpha\langle X, Y\rangle \geq 0$.
    - K^* is convex:
        - Let $Y, Z \in K^*$, $\lambda \in [0,1]$, $X \in K^*$.
        - $\langle X, \lambda Y + (1-\lambda)Z\rangle = \lambda\langle X, Y\rangle + (1-\lambda)\langle X, Z\rangle \geq 0$.
- For $K = S_+^m$, $K^* = K = S_+^m$ is self-dual.
    - Inner product for matrices: $X, Y \in \mathbb{R}^{n\times m}$, $\langle X, Y\rangle = tr(X^TY) = \sum_{i=1}^n \sum_{j=1}^m X_{ij}Y_{ij}$.
    - $K^* = \left(S_+^m\right)^* = \{Y : tr(XY) \geq 0, \forall X \in S_+^m\}$.
    - Any $Y \in S^m$, $Y \notin S_+^m$ is not in K^*.
        - To show, for each Y, find a single $X \in S_+^m$ s.t. $\langle X, Y\rangle < 0$.
        - If $Y \notin S_+^m$, then $\exists q \in \mathbb{R}^m$, s.t. $q^TYq < 0$.
        - Let $X = qq^T \in S_+^m$.
        - $\langle X, Y\rangle = tr(XY) = tr(qq^TY) = tr(q^TYq) = q^TYq < 0$.
        - So $Y \notin K^*$.
    - Any $Y \in S_+^m$ is in K^*.
        - To show, show that $\forall X \in S_+^m$, s.t. $\langle X, Y\rangle \geq 0$.
        - For $X \in S_+^m$, $X = Q\Lambda Q^T = \sum_{i=1}^m \lambda_i q_i q_i^T$, Q orthogonal, $\Lambda \geq 0$.
        - $\langle X, Y\rangle = tr(XY) = tr\left(\sum_{i=1}^m \lambda_i q_i q_i^T Y\right) = \sum_{i=1}^m \lambda_i tr\left(q_i q_i^T Y\right) \geq 0$, since $Y \in S_+^m$.

Dual of SDPs
- $\min c^Tx$,
  s.t. $x_1F_1 + \cdots + x_nF_n + G \leq 0$, $F_i, G \in S^n$.
- Primal variable: $x \in \mathbb{R}^n$.
- Dual variable: $Z \in S^n$.
- Lagrangian: $L(x, Z) = c^Tx + \langle Z, x_1F_1 + \cdots x_nF_n + G\rangle = c^Tx + \sum_{i=1}^n x_i\langle Z, F_i\rangle + \langle Z, G\rangle$.
    - $= \sum_{i=1}^n x_i\left(c_i + \langle Z, F_i\rangle\right) + \langle Z, G\rangle$.
- Dual function: $g(z) = \inf_x L(x, Z) = \begin{cases} tr(ZG), c_i + tr(ZF_i) = 0, \forall i \in [n] \\ -\infty, else \end{cases}$.
- Dual optimization problem:
  $\max tr(ZG)$,
  s.t. $c_i + tr(ZF_i) = 0$, $Z \geq 0$.
- Dual of SDP is SDP
- SDP can also satisfy strong duality if Slater's conditions are satisfied.

General approach to dualizing generalized inequalities
- If cone defining inequalities is K, find dual cone K^*.
- Constrain dual variables to K^*.
- Weak duality will follow from analogous step.
    - $g(z) = \inf_x L(x, z) = \inf_x\left(c^Tx + \langle Z, x_1F_1 + \cdots x_nF_n + G\rangle\right)$,
      $= \inf_x\left(c^Tx - \langle Z, -(x_1F_1 + \cdots x_nF_n + G)\rangle\right)$,
      $\leq c^Tx$.
    - If x primal feasible, $-\left(x_1F_1 + \cdots x_nF_n + G\right) \geq 0$.
    - If z dual feasible, then $Z \in \left(S_+^m\right)^* = S_+^m$.
- For 2 cases of interest, the cones are self-dual.
    - $\left(\mathbb{R}_+^m\right)^* = \mathbb{R}_+^m$.
    - $\left(S_+^m\right)^* = S_+^m$.

Motivation: SDP relaxations
- Original problem
    - $\min x^TAx$,
      s.t. $x_i \in \{-1,1\}$, $i \in [n]$ or $x_i^2 = 1$.
- 1st relaxation
    - $\min x^TAx$,

s.t. $-1 \le x \le 1$.
- If $A \in S_{++}^n$, get $x = 0$, not helpful.
- If $A \notin S_+^n$, still not convex.

- 2nd relaxation
  - $\min tr(XA)$, $X = xx^T$,
    $X_{ii} = 1$,
    $X \ge 0$,
    $rank(X) = 1$ (dropped to get SDP).
- Dualizing original problem:
  - $L(x, \nu) = x^T Ax + \sum_{i=1}^n \nu_i (x_i^2 - 1) = x^T (A + diag(\nu))x - 1^T \nu$.
  - $g(\nu) = \inf_x L(x, \nu) = \begin{cases} -1^T \nu, A + diag(\nu) \ge 0 \\ -\infty, else \end{cases}$.
  - Dual problem (SDP):
    $\max -1^T \nu$,
    s.t. $A + diag(\nu) \ge 0$.
- Dualizing 2nd relaxation:
  - $L(X, Z, \nu) = tr(XA) + \sum_{i=1}^n \nu_i (X_{ii} - 1) + \langle Z, -X \rangle = tr\left(X\big(A + diag(\nu) - Z\big)\right) - 1^T \nu$.
  - $g(Z, \nu) = \min_x L(X, Z, \nu) = \begin{cases} -1^T \nu, A + diag(\nu) - Z = 0 \\ -\infty, else \end{cases}$.
  - Dual problem:
    $\max -1^T \nu$,
    $A + diag(\nu) - Z = 0, Z \ge 0$.
  - Equivalent to dualizing the original problem

Non-convex problem satisfying strong duality

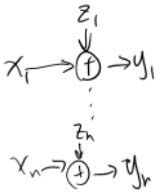
- $\min x^T Ax$,
  s.t. $x^T x \le 1$, $A \in S^n$.
- If $A \in S^n$, $A = Q\Lambda Q^T$, $Q \in \mathbb{R}^{n \times n}$ is orthonormal, rows/cols provide basis for $\mathbb{R}^n$.
  - Can write any $x \in \mathbb{R}^n$ as $x = \sum_{i=1}^n \alpha_i v_i = Q\alpha$.
- Rewrite the problem:
  - $x^T Ax = (\alpha^T Q^T) Q\Lambda Q^T Q\alpha = \alpha^T \Lambda \alpha = \sum_{i=1}^n \alpha_i^2 \lambda_i$.
  - $x^T x = \alpha^T Q^T Q\alpha = \alpha^T \alpha = \sum_{i=1}^n \alpha_i^2 \le 1$.
- $A \in S_+^n$, so $\lambda_i \ge 0$, then $p^* = 0$ with $\alpha_i = 0$, $x = 0$.
- $A \notin S_+^n$, so $\exists i$ s.t. $\lambda_i < 0$, then $p^* \ge \lambda_{min} \sum_{i=1}^n \alpha_i^2 \ge \lambda_{\min}$ achieved at $\alpha = e_j$, j corresponding to λ_{min}.
- Dualize problem
  - $L(x, \lambda) = x^T Ax + \lambda(x^T x - 1) = x^T (A + \lambda I)x - \lambda$.
  - $g(\lambda) = \min L(x, \lambda) = \begin{cases} -\lambda, A + \lambda I \ge 0 \\ -\infty, else \end{cases}$.
  - $\max -\lambda$,
    s.t. $A + \lambda I \ge 0, \lambda \ge 0$.
  - $A + \lambda I = Q\Lambda Q^T + \lambda QQ^T = Q(\Lambda + \lambda I)Q^T$, so $A + \lambda I \ge 0$ gives $\Lambda + \lambda I \ge 0$ or $\lambda \ge -\lambda_{min}(A)$.
  - When $A \in S_+^m$, we get $d^* = 0$.
  - When $A \notin S_+^m$, we get $d^* = \lambda_{min}$.
  - Strong duality holds
- Dual of the dual
  - $\min \lambda$,
    s.t. $A + \lambda I \ge 0, \lambda \ge 0$.
  - $L(x, Z, \nu) = \lambda - \langle Z, A + \lambda I \rangle - \nu\lambda = \lambda\, tr\left(\frac{1}{n} I - \frac{\nu}{n} I - tr(Z)\right) - tr(ZA)$.
  - $g(Z, \nu) = \inf_x L(x, Z, \nu) = \begin{cases} -tr(ZA), tr\left(\frac{1}{n} I - \frac{\nu}{n} I - tr(Z)\right) = 0 \\ -\infty, else \end{cases}$.
  - $\max -tr(ZA)$,
    s.t. $\nu \ge 0, Z \ge 0, \ tr(Z) = 1 - \nu$.
  - Equivalently, $\min tr(ZA)$, s.t. $Z \ge 0$, $tr(Z) \le 1$.
  - Equivalent to the relaxed SDP of the initial problem, with $Z = xx^T$.

KKT conditions

- Consider an optimization problem, for which primal and dual optimal values are obtained (at x^*, λ^*, ν^*) and $p^* = d^*$ (strong duality holds)
- $\min f_0(x)$,
  s.t. $f_i(x) \leq 0$, $i \in [m]$,
  $h_i(x) = 0$, $i \in [p]$.
- $L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)$.
  - $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$.
  - If strong duality holds, then $f_0(x^*) = g(\lambda^*, \nu^*)$.
  - $f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x\left[f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x)\right]$,
    $\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*)$, (1)
    $\leq f_0(x^*)$. (2)
  - We must have all equalities
  - Consequences:
    - (1) $\Rightarrow x^*$ is a minimizer of $L(x, \lambda^*, \nu^*)$.
    - (2) $\Rightarrow \lambda_i^* f_i(x^*) = 0, \forall i \in [m]$.
- Complementary slackness
  - Condition that $\lambda_i^* f_i(x^*) = 0, \forall i \in [m]$.
  - If ith constraint is inactive, $f_i(x) < 0$, then $\lambda_i^* = 0$.
    - No more return if we use more resource (changing from $f_i(x) < 0$ to $f_i(x) = 0$).
  - If $\lambda_i^* > 0$, then $f_i(x) = 0$.
    - We have use up all resources, if we want to improve, we go out of feasible set.
- If problem is differentiable
  - conditions
    - $f_0(x)$, $f_i(x)$, $h_i(x)$ are all differentiable.
    - Strong duality still holds
    - Convexity is not considered
  - x^* minimizes $L(x, \lambda^*, \nu^*)$ without constraints, $\nabla_x L(x, \lambda^*, \nu^*)\big|_{x=x^*} = 0$.
    - First order/primal optimal condition.
- KKT conditions
  - $\nabla_x L(x, \lambda^*, \nu^*)\big|_{x=x^*} = \nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) = 0$.
  - $f_i(x^*) \leq 0, \forall i \in [m]$, $h_i(x^*) = 0, \forall i \in [p]$.
  - $\lambda_i^* \geq 0, \forall i \in [m]$.
  - $\lambda_i^* f_i(x^*) = 0, \forall i \in [m]$.
- Theorems (necessary and sufficient conditions)
  - Necessary: If (x^*, λ^*, ν^*) are primal and dual optimal variables for an optimization problem, for which f_i and h_i all differentiable and for which strong duality holds, then (x^*, λ^*, ν^*) satisfies KKT conditions.
  - Sufficient: start with an optimization problem, for which f_i and h_i all differentiable, f_i convex, h_i affine, then if any $(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ satisfies KKT, then.
    - Strong duality holds.
    - $\tilde{x}$ primal optimal.
    - $\tilde{\lambda}, \tilde{\nu}$ dual optimal.
  - Proof (sufficient)
    - $L(x, \tilde{\lambda}, \tilde{\nu}) = f_0(x) + \sum_{i=1}^m \tilde{\lambda_i} f_i(x) + \sum_{i=1}^p \tilde{\nu_i} h_i(x)$.
    - Since f_0, f_i convex, $h_i(x)$ affine by assumption, $\tilde{\lambda_i} \geq 0$ by KKT, L is convex in x.
    - Since f_i, h_i differentiable, L is differentiable in x.
    - So, any point of zero gradient is global minimum.
    - By KKT(1), $\nabla_x L(x, \tilde{\lambda}, \tilde{\nu})\big|_{x=\tilde{x}} = 0$.
    - $g(\tilde{\lambda}, \tilde{\nu}) = \inf L(x, \tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$.
    - By definition, $g(\tilde{\lambda}, \tilde{\nu}) = f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda_i} f_i(\tilde{x}) + \sum_{i=1}^p \tilde{\nu_i} h_i(\tilde{x}) = f_0(\tilde{x})$.
      - Since $\tilde{\lambda_i} f_i(\tilde{x}) = 0$ by CS, $h_i(\tilde{x}) = 0$.
    - So strong duality holds.
    - Note, $\tilde{x}$ is also primal feasible by KKT(2).

- Summary
  - $g(\tilde{\lambda}, \tilde{\nu})$ is a lower bound on $f_0(x)$, $\forall x$ primal feasible and $\tilde{x}$ meets bound with equality, so $\tilde{x}$ is primal optimal.
  - $f_0(\tilde{x})$ is an upper bound on $g(\lambda, \nu)$, $\forall \lambda, \nu$ dual feasible and $(\tilde{\lambda}, \tilde{\nu})$ meets bound with equality,s o dual optimal.
- Combine two theorems
  - Class of optimization problems that are differentiable, so KKT condition exists.
  - Convex, so have sufficiency via B
  - Strong duality holds, so necessity via A
  - If differentiable, convex, satisfies Slater's, then KKT is necessary and sufficient

Water-filling for additive white Gaussian noise channels

- 

- $z_i \sim N(0, N_i)$.
  - $N_i \geq 0$: noise variance of channel i.
- $P_i \geq 0$: power over channel i.
- Total power constraint: $P_T \geq \sum_{i=1}^n P_i$.
- Problem

  $\max_{P_i} \sum_{i=1}^n \log\left(1 + \frac{P_i}{N_i}\right)$ (equivalently, $\min - \sum_{i=1}^n \log\left(1 + \frac{P_i}{N_i}\right)$),
  s.t. $P_i \geq 0$, $i \in [n]$,
  $\sum_{i=1}^n P_i \leq P_T$.
- $L(P, \lambda, \mu) = \sum_{i=1}^n - \log\left(1 + \frac{P_i}{N_i}\right) + \lambda\left(\sum_{i=1}^n P_i - P_T\right) - \sum_{i=1}^n \mu_i P_i$.
- KKT conditions
  - $\frac{\partial L}{\partial P_i} = -\frac{1}{1 + \frac{P_i}{N_i}} \frac{1}{N_i} + \lambda - \mu_i = 0, \forall i \in [n]$.
  - $P_i \geq 0, \sum_{i=1}^n P_i = P_T$.
  - $\lambda \geq 0, \mu_i \geq 0, \forall i \in [n]$.
  - $\mu_i P_i = 0$, if $P_i > 0$, then $\mu_i = 0$.
  - $\lambda\left(\sum_{i=1}^n P_i - P_T\right) = 0$, if $\sum P_i < P_T$, then $\lambda = 0$, if $\sum P_i = P_T$, then $\lambda \geq 0$.
    - Since objective is monotone increasing in each P_i, will use total budget, $\sum P_i = P_T$.
- By 1, $P_i + N_i = \frac{1}{\lambda - \mu_i}$.
  - If $P_i > 0$, then $\mu_i = 0$, $P_i + N_i = \frac{1}{\lambda}$ (power+noise=const for active channels).
  - If $P_i = 0$, then $N_i = \frac{1}{\lambda - \mu_i} \geq \frac{1}{\lambda}$.
- $\frac{1}{\lambda}$ is water-filling parameter.
  - If $N_i < \frac{1}{\lambda}$, we add power to channel i.
  - If $N_i \geq \frac{1}{\lambda}$, no need to make it active.
- For any fixed λ, $P_i = \max\left\{\frac{1}{\lambda} - N_i, 0\right\}$.
- By sorting (by noise level), identify $n^* \leq n$ active channels.
  - $\sum_{i=1}^{n^*} (P_i + N_i) = \sum_{i=1}^{n^*} \frac{1}{\lambda^*}$.
  - $P_T + \sum_{i=1}^{n^*} N_i = \frac{n^*}{\lambda^*}$.
  - $\frac{1}{\lambda^*} = \frac{1}{n^*}\left(P_T + \sum_{i=1}^{n^*} N_i\right)$.
- Perturb power budget from P_T to $P_T + \epsilon$.
  - Assume n^* active channel, each gets $\frac{\epsilon}{n^*}$ extra power, what's the benefit?
  - $\log\left(1 + \frac{P_i^* + \epsilon/n^*}{N_i}\right) - \log\left(1 + \frac{P_i^*}{N_i}\right)$,

$= \log\left(\frac{P_i^*+N_i+\epsilon/n^*}{P_i^*+N_i}\right) = \log\left(1+\frac{\epsilon/n^*}{P_i^*+N_i}\right)$,
$= \log\left(1+\frac{\epsilon/n^*}{1/\lambda^*}\right)$,
$\approx \frac{\epsilon/n^*}{1/\lambda^*} = \frac{\epsilon\lambda^*}{n^*}$ is the rate increase for each channel i.

- Total rate is increased by $\epsilon\lambda^*$.

Geometric interpretation of KKT

- $\min f_0(x)$,
  s.t. $f_i(x) \le 0, i \in [m]$,
  $h_i(x) = 0, i \in [p]$.
- At optimum x^*, some $f_i(x) < 0$ inactive, consider the following problem only
  $\min f_0(x)$,
  s.t. $f_i(x) = 0, \{i: f_i \ active\}$,
  $h_i(x) = 0, i \in [p]$.
- For equality constraints
  $\min f_0(x)$,
  s.t. $Ax = b$.
  - Perturb x^* while staying feasible
    - $A(x^* + \Delta x) = Ax^* + A\Delta x = b$.
    - A feasible perturbation satisfies $A\Delta x = 0$.
    - e.g. $A = (2,1)$, $b = 1$, $2x_1 + x_2 = 1$.
      - $\Delta x = \left\{\alpha\begin{pmatrix}-1\\2\end{pmatrix}, \alpha \in \mathbb{R}\right\}$.
    - Generally, $A\Delta x = \begin{pmatrix}a_1^T\\ \vdots \\ a_p^T\end{pmatrix}\Delta x = 0$ gives $a_i \perp \Delta x, \forall i \in [p]$, $\Delta x \in N(A)$.
    - A point $x^* \in C$ for a convex opt problem is optimal iff $\forall y \in C$, $\nabla f_0(x^*)^T(y - x^*) \ge 0$.
      - If $\nabla f_0(x^*)^T \Delta x \ge 0$, $\forall \Delta x \in N(A)$, then $-\Delta x \in N(A)$.
      - For optimality, need $\nabla f_0(x^*)^T \Delta x = 0$, $\forall \Delta x \in N(A)$.
      - In other words, $\nabla f_0(x^*)^T \perp N(A)$, i.e. $\nabla f_0(x^*)^T \in N(A)^\perp = R(A^T)$.
      - Hence, can write $\nabla f_0(x^*) = A^T\alpha$.
  - Optimum criteria for equality constrained optimization problem
    - A point x is optimal iff $\nabla f_0(x)^T \Delta x = 0$, $\forall x$, s.t. $A\Delta x = 0$.
  - Connect to KKT
    - $L(x, \nu) = f_0(x) + \nu^T(Ax - b)$.
    - $\nabla_x L = \nabla f_0(x) + A^T\nu = 0$, $\nabla f_0(x) = A^T(-\nu) \in R(A^T)$.
    - e.g. $\min \frac{1}{2}(x_1^2 + x_2^2)$, s.t. $(2,1)\begin{pmatrix}x_1\\x_2\end{pmatrix} = 1$, $x^* = \begin{pmatrix}2/5\\1/5\end{pmatrix}$.
      - $\nabla f_0(x^*) = \begin{pmatrix}2/5\\1/5\end{pmatrix}$, $A^T = \begin{pmatrix}2\\1\end{pmatrix}$, $-\nu = \frac{1}{5}$.
- The KKT condition represent balance of force
  - $\nabla f_0(x) = -\sum_{i=1}^m \lambda_i \nabla f_i(x) - \sum_{i=1}^p \nu_i \nabla h_i(x)$.
- Why Slater's?
  - We need some $\{\lambda_i\}$ to make $\nabla f_0(x) = -\sum_{i=1}^m \lambda_i \nabla f_i(x)$.
  - e.g. $\min x_1 + x_2$, s.t. $(x_1+1)^2 + x_2^2 \le 1$, $(x_1-2)^2 + x_2^2 \le 4$.
    - Only one feasible point $x^* = (0,0)$, Slater's doesn't hold.
    - $\nabla f_0(x) = \begin{pmatrix}1\\1\end{pmatrix}$, $\nabla f_1(x) = \begin{pmatrix}2x_1+2\\2x_2\end{pmatrix} = \begin{pmatrix}2\\0\end{pmatrix}$, $\nabla f_2(x) = \begin{pmatrix}2x_1-4\\2x_2\end{pmatrix} = \begin{pmatrix}-4\\0\end{pmatrix}$.
    - Cannot pick λ to have $\nabla f_0(x) = -\lambda_1 \nabla f_1(x) - \lambda_2 \nabla f_2(x)$.

Algorithms

September 12, 2022 12:56 PM

Unconstrained optimization

- $\min f_0(x)$, f_0 is convex and twice differentiable
- Idea: produce a sequence x^k, $k = 1,2,3,\ldots$ such that cost decreases at each step and $f_0(x^k) \to p^* = \min f_0(x)$.
- Descent method:
  - $x^{k+1} = x^k + t^k \Delta x^k$, t is step size, Δx is direction.
  - Need $f_0(x^{k+1}) < f_0(x^k)$.
- Steepest/gradient descent:
  - Pick Δx^k to align with direction of most negative gradient $\Delta x^k = -\nabla f_0(x^k)$.
  - Since $f(x)$ is convex, $f_0(y) \geq f_0(x) + \nabla f_0(x)^T(y - x)$.
    - Set $f_0(y) = f_0(x^{k+1})$, $f_0(x) = f_0(x^k)$, $(y - x) = \Delta x^k$.
  - For choice in steep descent, $f_0(x^{k+1}) \geq f_0(x^k) - \|\nabla f_0(x^k)\|_2^2$.
  - But just picking direction as above and step size $t = 1$ does not guarantee progress
  - Algorithm: given $x \in dom\,(f_0)$.
    - Repeat:
      - Choose $\Delta x = -\nabla f_0(x)$.
      - Choose $t > 0$.
      - Update $x + t\Delta x$.
    - Until $\|\nabla f_0(x)\|^2 < \epsilon$.
- Choosing t.
  - Exact line search:
    - Set $t = \arg\min_{t>0} f_0(x + t\Delta x)$.
    - 1D convex optimization problem.
  - Backtracking line search:
    - Parameters:
      - $\alpha \in (0,0.5)$: used to identify a good step size.
      - $\beta \in (0,1)$: multiplicative step size search parameter.
    - Algorithm: start with $t = \frac{1}{\beta}$.
      - Repeat:
        - Set $t = \beta t$ (reduce step size).
      - Until $f_0(x + t\Delta \mathrm{x}) < f_0(x) + \alpha t \nabla f_0(x)^T \Delta x$.
- Newton's method:
  - Improved direction
    - In steepest descent, fit a hyperplane to $f_0(x)$, first order method.
    - In Newton's method, fit a second order approximation to determine direction
  - $f_0(x + \Delta x) \approx f_0(x) + \nabla f_0(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 f_0(x) \Delta x$.
  - Minimize $f_0(x)$ w.r.t. Δx to find direction.
    - $\frac{\partial}{\partial \Delta x}\left(f_0(x + \Delta x)\right) = \nabla f_0(x) + \nabla^2 f_0(x)\Delta x = 0$.
    - $\Delta x = -\left(\nabla^2 f_0(x)\right)^{-1} \nabla f_0(x)$ if Hessian is invertible.
  - Algorithm: given $x \in dom(f_0)$.
    - Repeat:
      - Choose $\Delta x_{nt} = -\left(\nabla^2 f_0(x)\right)^{-1} \nabla f_0(x)$.
      - Choose $t > 0$.
      - Update $x = x + t\Delta x_{nt}$.
    - Until $\sqrt{\nabla f_0(x)^T \left(\nabla^2 f_0(x)\right)^{-1} \nabla f_0(x)} < \epsilon$.
  - Exit condition: since $f_0(x + t\Delta x_{nt}) \approx f_0(x) - \left(t - \frac{t^2}{2}\right) \nabla f_0(x)^T \left(\nabla^2 f_0(x)\right)^{-1} \nabla f_0(x)$.

- Example:
  - $f_0(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2), \gamma > 1$.
  - $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $x^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\nabla f_0(x) = \begin{pmatrix} x_1 \\ \gamma x_2 \end{pmatrix}$, $\nabla^2 f_0(x) = \begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}$.
  - $\Delta x_{nt} = -(\nabla^2 f_0(x))^{-1} \nabla f_0(x) = -\begin{pmatrix} 1 & 0 \\ 0 & 1/\gamma \end{pmatrix} \begin{pmatrix} x_1 \\ \gamma x_2 \end{pmatrix} = -\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

Equality constrained minimization

- $\min f_0(x)$,
  s.t. $Ax = b$.
- KKT conditions:
  - $L(x, \nu) = f_0(x) + \nu^T(Ax - b)$.
  - $\nabla_x L(x, \nu) = \nabla f_0(x) + A^T \nu = 0$.
  - $Ax = b$.
- Idea is to solve sequentially while continually satisfying primal feasibility
  - $x^{k+1} = x^k + t\Delta x$, $Ax^{k+1} = b$.
  - $t\Delta x$ must be selected to satisfy primal feasibility.
- $\min \nabla f_0(x) v + \frac{1}{2} v^T \nabla^2 f_0(x) v$,
  s.t. $A(x + v) = b$ (since $Ax = b$, we simply need $Av = 0$).
- Solve for v.
  - $L(v, \mu) = \nabla f_0(x)^T v + \frac{1}{2} v^T \nabla^2 f_0(x) v + \mu^T(Av)$.
  - KKT gives: $\nabla_v L = \nabla f_0(x) + \nabla^2 f_0(x) v + A^T \mu = 0$, $Av = 0$.
  - $\begin{pmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} v \\ \mu \end{pmatrix} = \begin{pmatrix} -\nabla f_0(x) \\ 0 \end{pmatrix}$.
  - The matrix is called KKT matrix.
  - Solution:
    - Invert KKT matrix to find v.
    - Back substitution if only $\nabla^2 f_0(x)$ is invertible. If not invertible, can still deal with that by making it PSD. Now consider the invertible case
- Back substitution
  - $v + (\nabla^2 f_0(x))^{-1} A^T \mu = -(\nabla^2 f_0(x))^{-1} \nabla f_0(x)$.
  - $Av + A(\nabla^2 f_0(x))^{-1} A^T \mu = -A(\nabla^2 f_0(x))^{-1} \nabla f_0(x)$.
  - Since $Av = 0$, $\mu = -\left(A(\nabla^2 f_0(x))^{-1} A^T\right)^{-1} A(\nabla^2 f_0(x))^{-1} \nabla f_0(x)$.
  - Substitute μ back into v equation, $v = -\nabla^2 f_0(x)^{-1}(\nabla f_0(x) + A^T \mu)$.
    - Note $A^T \mu$ adds the constraint.
- Algorithm: given $x^0 \in dom(f_0)$ such that $Ax^0 = b$.
  - Repeat:
    - Compute v as above.
    - Set $\Delta x_{nt} = v$.
    - Line search for t.
    - $x^{t+1} = x^t + t\Delta x_{nt}$ (since $Av = 0$, $Atv = 0$, doesn't affect feasibility).
  - Until $\Delta x_{nt}^T (\nabla^2 f_0(x))^{-1} \Delta x_{nt} < \epsilon^2$.
- Infeasible start Newton
  - $\min f_0(x^0) + \nabla f_0(x^0)^T v + \frac{1}{2} v^T \nabla^2 f_0(x^0) v$,
    s.t. $A(x^0 + v) = b$.
  - $\begin{pmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} v \\ \mu \end{pmatrix} = \begin{pmatrix} -\nabla f_0(x) \\ -(Ax^0 - b) \end{pmatrix}$.
  - If use step size $t = 1$, get a feasible x^1.
  - Can be used in the algorithm above.
- Interpretation of infeasible start as a primal dual algorithm
  - Update both primal variable x and dual variable ν in order to approximately statisfy KKT.
  - $\min f_0(x)$, s.t. $Ax = b$.

- KKT: $\nabla f_0(x) + A^T\nu = 0$, $Ax = b$.
  - Let $y = \begin{pmatrix} x \\ \nu \end{pmatrix}$, residue $r(y) = \begin{pmatrix} \nabla f_0(x) + A^T\nu \\ Ax - b \end{pmatrix}$.
  - Goal: drive $\|r(y)\| \to 0$, stop when $\|r(y)\| < \epsilon$.
  - Start at $y = \begin{pmatrix} x \\ \nu \end{pmatrix}$, move to $y + \Delta y = \begin{pmatrix} x \\ \nu \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta \nu \end{pmatrix}$.
  - $r(y + \Delta y) = r(y) + Dr(y)\Delta y = r(y) + \begin{pmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta \nu \end{pmatrix} = 0$.
    - $\begin{pmatrix} \nabla f_0(x) + A^T\nu \\ Ax - b \end{pmatrix} + \begin{pmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta \nu \end{pmatrix} = 0$.
  - $\begin{pmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta \nu \end{pmatrix} = -\begin{pmatrix} \nabla f_0(x) + A^T\nu \\ Ax - b \end{pmatrix}$.
  - Equivalently, $\begin{pmatrix} \nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \Delta x \\ \nu + \Delta \nu \end{pmatrix} = -\begin{pmatrix} \nabla f_0(x) \\ Ax - b \end{pmatrix}$.

Inequality constrained problems

- $\min f_0(x)$,
  s.t. $Ax = b$,
  $f_i(x) \le 0, i \in [m]$.
- Idea (interior point): build a barrier at edge of feasible set so that always stay strictly feasible.
- Log barrier
  - Adds a parameter $t > 0$, $-\frac{1}{t}\log(-u)$.
  - As $t \to \infty$, get $\begin{cases} 0, u < 0 \\ \infty, u = 0 \end{cases}$.
- Modify problem using log barrier
  - $\min f_0(x) - \frac{1}{t}\sum_{i=1}^m \log(-f_i(x))$,
    s.t. $Ax = b$.
  - Often do $\min t f_0(x) - \sum_{i=1}^m \log(-f_i(x))$, s.t. $Ax = b$.
- Algorithm (Barrier method)
  - Initialize x^0 feasible, $t^0 = 10$.
  - Repeat:
    - Solve $\min t f_0(x) - \sum_{i=1}^m \log(-f_i(x))$, s.t. $Ax = b$ using equality constrained algorithms.
    - Update $x^{k+1} = x^*(t^k)$.
    - Increment $t^{k+1} = \gamma t^k$ (typically $\gamma = 10 \sim 20$).
  - Until $\frac{m}{t} < \epsilon$, where m is the number of inequality constraints.
- Note: 2 loops
  - Outer: update t.
  - Inner: solve an optimization problem.
    - Requires Newton's method, since both $t f_0(x)$ and $\sum_{i=1}^m \log(-f_i(x))$ are large.
- Central path:
  - Trajectory of x^k, stays in the feasible set, moving towards the boundary.

Log barrier cont

- $\phi(x) = -\sum \log(-f_i(x))$.
- $\nabla \phi = \sum -\frac{1}{f_i(x)}\nabla f_i(x)$.
- $\nabla^2 \phi = \sum \frac{1}{(f_i(x))^2}\nabla f_i(x) \nabla f_i(x)^T + \sum \frac{1}{-f_i(x)}\nabla^2 f_i(x)$.

Phase I: find a feasible x^0

- Solve a feasibility problem
  - $\min s$,
    s.t. $f_i(x) \le s, i \in [m]$,
    $Ax = b$.

- $s^* < 0$, x^* is in interior, use as x^0.
- $s^* > 0$, feasible set is empty.
- To initialize phase I, need a strictly feasible (s, x).
  - Pick any $x \in \mathbb{R}^n$ (actually $\cap_i^m dom(f_i)$).
  - Set $s = \max f_i(x) + \epsilon$.

Stopping criteria

- Consider a point $x^*(t)$ on central path $x^*(t) = \arg\min_{Ax=b} tf_0(x) - \sum \log(-f_i(x))$.
  - Any such $x^*(t)$ is strictly feasible.
  - $Ax^*(t) = b$.
  - $f_i(x^*(t)) < 0$.
- Lagrangian: $\tilde{L}(x, \mu) = tf_0(x) - \sum \log(-f_i(x)) + \mu^T(Ax - b)$.
- Since x^* is optimum, must satisfy KKT: $t\nabla f_0(x^*) + \sum \frac{1}{-f_i(x^*)} \nabla f_i(x^*) + A^T\mu = 0$.
  - $\nabla f_0(x^*) + \sum \frac{1}{-tf_i(x^*)} \nabla f_i(x^*) + A^T\left(\frac{\mu}{t}\right) = 0$.
- For original problem
  - $\min f_0(x)$,
    s.t. $f_i(x) \le 0$,
    $Ax = b$.
  - $L(x, \lambda, \nu) = f_0(x) + \sum \lambda_i f_i(x) + \nu^T(Ax - b)$.
- $\lambda_i^* = \frac{1}{-tf_i(x^*(t))} > 0$, $\nu^* = \frac{\mu}{t}$.
- $L(x, \lambda^*, \nu^*) = f_0(x) + \sum \lambda_i^* f_i(x) + \nu^{*T}(Ax - b)$.
- Note: $L(x, \lambda^*, \nu^*)$ is convex in x.
- $\arg\min_x L(x, \lambda^*, \nu^*) = x$ such that $\nabla_x L(x, \lambda^*, \nu^*) = 0$.
- $\nabla_x L(x, \lambda^*, \nu^*) = \nabla f_0(x) + \sum \lambda_i^* \nabla f_i(x) + A^T\nu^* = 0$.
- $x^*(t) = \arg\min_x L(x, \lambda^*, \nu^*)$.
- $g(\lambda^*, \nu^*) = \min_x L(x, \lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) \le \max_{\lambda \ge 0} g(\lambda, \nu) = d^* = p^*$.
- $p^* \ge g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f_0(x^*) + \sum \frac{1}{-tf_i(x^*)} f_i(x^*) + (\nu^*)^T(Ax^* - b) = f_0(x^*) - \frac{m}{t}$.
- $\frac{m}{t} \ge f_0(x^*) - p^* \ge 0$.
- To apply equality constrained Newton to P_1, solve
  - $\begin{pmatrix} t\nabla^2 f_0(x) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \nu \end{pmatrix} = -\begin{pmatrix} t\nabla f_0(x) + \nabla\phi(x) \\ Ax - b \end{pmatrix}$.

Inequality-constrained SDPs

- $\min c^T x$,
  s.t. $x_1F_1 + \cdots x_nF_n + G \le 0$.
- Let $F(x) = x_1F_1 + \cdots + x_nF_n + G$.
- $\phi(x) = -\sum \log(-f_i(x)) = -\log(-\prod f_i(x)) = -\log\det\left(diag\left(-f_i(x)\right)\right)$ for ordinary problems.
- Barriers for SDPs: $\phi(X) = -\log\det(-F(X))$.
  - $\nabla \log\det X = X^{-1}$.
- Start with an $F(X)$ in interior, $-F(X) \in S_+^m$.
  - $-F(X) > 0$, $\det(-F(X)) > 0$.
- As an eigenvalue approaches boundary, $eig(-F(X)) \to 0$, $\det(-F(X)) \to 0$, $-\log\det(-F(X)) \to \infty$.
- $\min c^T x + \phi(x) = \min c^T x - \frac{1}{t}\log\det(-F(X))$.