0.1 Symbols

$S^n \equiv$ set of all symmetric matrices

$M_+^n \equiv$ set of all positive definite matrices (PD)

$S_+^n \equiv$ set of all symmetric positive semidefinite matrices (SPSD)

$S_{++}^n \equiv$ set of all symmetric positive definite matrices (SPD)

$cl(X) \equiv$ closure of X

$relint(X) \equiv$ relative interior of X

$int(X) \equiv$ interior of X

$[x]^+ \equiv max(x, 0)$

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

$prox_{f,\lambda}(y) = f(x) + \frac{1}{2\lambda}\|x - v\|_2^2$

SoftThresholding$_\lambda(y) = prox_{\|*\|_1,\lambda}(y)$

0.2 Preliminary

Consider $f : \mathbb{R}^n \to \mathbb{R}$
Gradient of f: $\nabla f(x) = \begin{bmatrix} \partial f / \partial x_i \\ .. \end{bmatrix}$
$f(x) = a^T x \implies \nabla f(x) = a$
$f(x) = x^T P x, P = P^T \implies \nabla f(x) = 2Px$
$f(x) = x^T P x \implies \nabla f(x) = 2(\frac{P^T+P}{2})x = (P^T+P)x$

Taylor expansion approximation:
$f(x) \approx f(x_0) + \nabla^T f(x_0)(x - x_0) + o((x-x_0)^2)$
$f(x + \delta x) \approx f(x_0) + \nabla^T f(x)\delta x + o((\delta x)^2)$

Chain rule:
$f : \mathbb{R} \to \mathbb{R}, g : \mathbb{R} \to \mathbb{R}, h(x) = f(g(x))$
$\nabla h(x) = g'(f(x))\nabla f(x)$

$g : \mathbb{R}^m \to \mathbb{R}, g(x) = f(Ax + b)$
$\nabla g(x) = A^T \nabla f(Ax + b)$

2nd derivative:
$\nabla^2 f(x) = \begin{bmatrix} \partial^2 f / \partial x_1 \partial x_1 & ... \\ .. & \partial^2 f / \partial x_n \partial x_n \end{bmatrix}$
$\nabla f(x) = Px + g$
$\nabla^2 f(x) = P$

Hessian gives the 2nd order approximation:
$f(x) \approx f(x_0) + \nabla^T f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$

Matrices:
$A \in \mathbb{R}^{m \times n}$: set of all real matrices
inner product: $\sum_i \sum_j x_{ij} y_{ij} = trace(XY^T) = trace(Y^T X) = \sum_i (XY)_{ii}$
note trace has cyclic property
frobenius norm: $\|X\|_F = (\sum_i \sum_j X_{ij}^2)^{\frac{1}{2}}$
range: $R(A) = \{Ax : x \in \mathbb{R}^n\} = \sum_i a_i x_i$, where a_i is ith column (column space of A)
null space: $N(A) = \{x : Ax = 0\}$

SVD:
$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$
U and V are left and right eigenvector matrixes
U and V are orthogonal matrixes($BB^T = B^T B = I$)
Σ is rectangular diagonal matrix of eigenvalues
$A_{m \times n} x_n$
linear transformation: $U \Sigma V^T x$
rotation - scaling - rotation

PSD matrix:
$A\ PSD \iff (\forall x) x^T A x \geq 0 \iff (\forall i) \lambda_i(A) \geq 0$
$A\ PSD \implies A^{1/2}$ exists

Real symmetric matrices have real eigenvalues:

$$Av = \lambda v$$
$$v^* A v = v^* \lambda v = \lambda \|v\|_2^2$$
$$(v^* A v)^* = v^* A^* v = \lambda^* \|v\|_2^2 \implies \lambda = \lambda^*$$

Affine sets
A set $C \subseteq \mathbb{R}^n$ is affine if $(\forall x_1, x_2 \in C)(\forall \theta \in \mathbb{R}) \implies \theta x_1 + (1 - \theta)x_2 \in C$

Convex sets
A set $C \subseteq \mathbb{R}^n$ is convex if $(\forall x_1, x_2 \in C)(\forall \theta \in \mathbb{R}) 0 \leq \theta \leq 1 \implies \theta x_1 + (1 - \theta)x_2 \in C$

Operations preserving convex sets:

- partial sum
- sum
- coordinate projection
- scaling
- translation
- intersection between any convex sets

Separating Hyperplanes: if $S, T \subset \mathbb{R}^n$ are convex and disjoint, then $\exists a \neq 0, b$ such that:

$$a^T x \geq b, \forall x \in S$$
$$a^T x \leq b, \forall x \in T$$

Supporting Hyperplane:
if S is convex, $\forall x_0 \in \partial S$ (boundary of S), then $\exists a \neq 0$ such that $a^T x \leq a^T x_0, \forall x \in S$

Convex combination:
$\sum_i \theta_i x_i, \forall \theta_i \in \mathbb{R}, \sum_i \theta_i = 1, \theta_i \geq 0$

Convex hull:
The set of all convex combinations of points in C, the hull is convex

Hyperplane

$$C = \{x : a^T x = b\}, a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$$

Halfspaces

$$\begin{aligned}
&C = \{x : a^T x \leq b\}, a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R} \\
&\text{ let } a^T x_c = b \\
&C = \{x : a^T (x - x_c) \leq 0\}, a \in \mathbb{R}^n, a \neq 0
\end{aligned}$$

Elipse

$$\begin{aligned}
E(x_c, P) = \{x : (x - x_c)^T P^{-1} (x - x_c) \leq 1\}, P > 0 \\
P = r^2 I \implies \text{ Euclidean Ball} \\
P = Q \begin{bmatrix} \lambda_1 & .. \\ .. & \lambda_n \end{bmatrix} Q^T \\
(x - x_c)^T (Q \begin{bmatrix} \lambda_1 & .. \\ .. & \lambda_n \end{bmatrix} Q^T)^{-1} (x - x_c) \leq 1 \\
\tilde{x}^T \begin{bmatrix} \frac{1}{\lambda_1} & .. \\ .. & \frac{1}{\lambda_n} \end{bmatrix} \tilde{x} \leq 1 \\
\tilde{x}^T \begin{bmatrix} \frac{1}{\lambda_1} & .. \\ .. & \frac{1}{\lambda_n} \end{bmatrix} \tilde{x} = \frac{\tilde{x_1}^2}{\lambda_1} + .. + \frac{\tilde{x_n}^2}{\lambda_n} \leq 1
\end{aligned}$$

volum of elipsoid proportional to $\sqrt{det(P)} = \sqrt{\Pi_i \lambda_i}$

0.3 Problem Types

LP

standard, inequality, general forms

$$\begin{aligned} \min_x c^T x \ s.t. : \\ Ax = b \\ x \succeq 0 \end{aligned}$$

$$\begin{aligned} \min_x c^T x \ s.t. : \\ Ax \preceq b \end{aligned}$$

$$\begin{aligned} \min_x c^T x + d \ s.t. : \\ Gx \preceq h \\ Ax = b \end{aligned}$$

QP

$$\begin{aligned} \min_x \frac{1}{2} x^T P x + q^T x + r \ \ s.t. : \\ Gx \leq h \\ Ax = b \end{aligned}$$

QCQP

$$\begin{aligned} \min_x \frac{1}{2} x^T P_0 x + q_0^T x + r_0, s.t. : \\ \frac{1}{2} x^T P_i x + q_i^T x + r_i \leq 0, \forall i \\ Ax = b \end{aligned}$$

SOCP

$$\begin{aligned} \min_x f^T x \ s.t. : \\ \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \forall i \\ Fx = g \end{aligned}$$

$$\begin{aligned} (\forall i) b_i = 0 \implies LP \\ (\forall i) c_i = 0 \implies QCQP \end{aligned}$$

GP

$$\begin{aligned} \min_x f_0(x) \ s.t. : \\ f_i(x) \leq 1, \forall i \\ h_i(x) = 1, \forall i \\ f_i \ is \ a \ posynomial := \sum_i h_i \\ h_i \ is \ a \ monomial := c x_1^{a_1} x_2^{a_2} .., c > 0, a_i \in \mathbb{R} \end{aligned}$$

Use transform of objective and constraint functions:
$y_i = log x_i, x_i = e^{y_i}$
$\tilde{h_i}$ becomes exponential of affine function
$\tilde{f_i} = log(f_i)$ becomes log sum exp (convex)
If all constraints and objective are monomials, reduces to LP after transform.

Generlized Inequality, Conic

$$\begin{aligned} \min_x \ c^T x \ s.t. : \\ Fx + g \preceq_K 0 \\ Ax = b \end{aligned}$$

Gneralized Inequality, SDP

general, standard, inequality forms

$$\begin{aligned} \min_x \ c^T x \ s.t. : \\ LMI : \ \sum_i^n x_i F_i + G \preceq_K 0 \\ Ax = b \\ x \in \mathbb{R}^n \\ F_i, G \in S^m, K \in S_+^m \end{aligned}$$

$$\begin{aligned} \min_X \ tr(CX) s.t. : \\ tr(A_i X) = b_i, \forall i \\ X \succeq 0 \end{aligned}$$

$$\begin{aligned} \min_x \ c^T x \ s.t. : \\ \sum_i^n x_i A_i \preceq_K B \\ Ax = b \\ B, A_i \in S^m, K \in S_+^m \end{aligned}$$

concatenating constraints:

$$\begin{aligned} F^{(i)}(x) = \sum_j x_j F_i^{(i)} + G^{(i)} \preceq 0 \\ Gx \preceq h \\ \implies \\ diag(Gx - h, F^{(1)}(x), .., F^{(m)}(x)) \preceq 0 \end{aligned}$$

if all matrices are diagonal, reduces to LP

0.4 Convex/Concave Functions

- Affine
- Pointwise supremum of convex functions
  - distance to farthest point in arbitrary set
  - support function of set
- Partial minimization:
  $g(x,y)\ convex\ in\ x,y,C\ convex \implies$
  $min_{y\in C}\ g(x,y)\ convex$
- shortest distance to a convex set
- Any type of norm
- Non-negative weighted sum of convex functions
- indicator function $I_C(x)$

0.4.1 log det X, concave

$$\begin{aligned}
let\ X = Z + tV &\succ 0\\
f &= logdet(Z+tV)\\
f &= logdet(Z^{-0.5}(I+tZ^{-0.5}VZ^{0.5})Z^{0.5})\\
f &= log(det(Z^{-0.5})det(I+tZ^{-0.5}VZ^{0.5})det(Z^{0.5}))\\
f &= log(det(Z^0)det(I+tZ^{-0.5}VZ^{0.5}))\\
f &= logdet(I+tZ^{-0.5}VZ^{0.5})\\
f &= log\Pi_i(1+\lambda_i t)\\
f &= \sum_i log(1+\lambda_i t)\\
\frac{\partial f}{\partial t} &= \sum_i \frac{\lambda_i}{1+\lambda_i t}\\
\frac{\partial^2 f}{\partial t^2} = \sum_i \frac{-\lambda_i^2}{(1+\lambda_i t)^2} &= -\sum_i \frac{\lambda_i^2}{(1+\lambda_i t)^2} \leq 0\\
\nabla^2 f \leq 0 &\iff f\ concave
\end{aligned}$$

0.4.2 log $\sum_i exp(x_i)$, convex

$$\begin{aligned}
\nabla^2 f &= \frac{1}{(1^T z)^2}(1^T z diag(z) - zz^T)\\
v^T zz^T v &= det(v^T zz^T v) = det(vv^T zz^T)\\
v^T zz^T v &= \sum_j \sum_i z_j z_i v_j v_i\\
v^T zz^T v &= (\sum_j z_j z_j)(\sum_i z_i v_i)\\
v^T zz^T v &= (\sum_i z_i v_i)^2\\
&use\ Holder's\ Inequality:\\
&\|a\|_2^2\|b\|_2^2 \geq |a^T b|^2\\
&let\ a = z_i^{0.5}, b = v_i z_i^{0.5}\\
1^T z(\sum_i v_i^2 z_i) &- (\sum_i z_i v_i)^2 \geq 0\\
v^T \nabla^2 f v &= \frac{1}{(1^T z)^2}\left(1^T z(\sum_i v_i^2 z_i) - (\sum_i z_i v_i)^2\right) \geq 0\\
\nabla^2 f \geq 0 &\iff f\ convex
\end{aligned}$$

0.4.3 geometric mean on R^n_{++}, concave

$$\begin{aligned}
f &= (\Pi_i x_i)^{\frac{1}{n}}\\
\frac{\partial}{\partial x_i} f &= \frac{1}{n}(\Pi_i x_i)^{\frac{1}{n}-1}\Pi_{j\neq i} x_j\\
\frac{\partial^2}{\partial x_i^2} f &= \frac{1}{n}(\frac{1}{n}-1)(\Pi_i x_i)^{\frac{1}{n}-2}(\Pi_{j\neq i} x_j)^2\\
\frac{\partial^2}{\partial x_i^2} f &= \frac{1}{n}(\frac{1}{n}-1)\frac{(\Pi_i x_i)^{\frac{1}{n}}}{x_i^2}\\
\frac{\partial^2}{\partial x_i x_k} f &= \frac{1}{n^2}\frac{(\Pi_i x_i)^{\frac{1}{n}}}{x_i x_k}, i \neq k\\
\frac{\partial^2}{\partial x_i x_k} f &= \frac{1}{n^2}\frac{(\Pi_i x_i)^{\frac{1}{n}}}{x_i x_k} - \delta_{ik}\frac{1}{n}\frac{(\Pi_i x_i)^{\frac{1}{n}}}{x_i^2}\\
v^T \nabla^2 f v &= \frac{-(\Pi_i x_i)^{\frac{1}{n}}}{n^2}(n\sum_i \frac{v_i^2}{x_i^2} - (\sum_i \frac{v_i}{x_i})^2)\\
&apply\ Cauchy\ Schwartz\ Inequality:\\
&let\ a = \mathbf{1}, b_i = \frac{v_i}{x_i}\\
\|\mathbf{1}\|_2^2(\sum_i \frac{v_i^2}{x_i}) &\geq (\sum_i \frac{v_i}{x_i})^2\\
n\sum_i \frac{v_i^2}{x_i^2} - (\sum_i \frac{v_i}{x_i})^2 &\geq 0\\
v^T \nabla^2 f v \leq 0 &\iff f\ concave
\end{aligned}$$

0.4.4 quadratic over linear, convex

$$
\begin{aligned}
f(x,y) = \frac{h(x)}{g(y)}, g(y)\ linear, g(y) \in R_+ \\
\nabla^2 f = vv^T\ is\ PSD \iff f\ convex
\end{aligned}
$$

0.5 Composition of functions

Mnemonic derivation from scalar composite function

$$
\begin{aligned}
f = h(g(x)) \\
f' = g'(x)h'(g(x)) \\
f'' = g''(x)h'(g(x)) + (g'(x))^2 h''(g(x))
\end{aligned}
$$

$$
\begin{aligned}
\text{h convex \& non-decreasing, g convex} \implies \text{f convex} \\
h'' \geq 0, g''(x) \geq 0, h'(g(x)) \geq 0 \implies f'' \geq 0
\end{aligned}
$$

$$
\begin{aligned}
\text{h convex \& non-increasing, g concave} \implies \text{f convex} \\
h'' \geq 0, g''(x) \leq 0, h'(g(x)) \leq 0 \implies f'' \geq 0
\end{aligned}
$$

$$
\begin{aligned}
\text{h concave \& non-decreasing, g concave} \implies \text{f concave} \\
h'' \leq 0, g''(x) \leq 0, h'(g(x)) \geq 0 \implies f'' \leq 0
\end{aligned}
$$

$$
\begin{aligned}
\text{h concave \& non-increasing, g convex} \implies \text{f concave} \\
h'' \leq 0, g''(x) \geq 0, h'(g(x)) \leq 0 \implies f'' \leq 0
\end{aligned}
$$

0.6 Convexity Preservation of Sets

0.6.1 Intersection

$$\begin{aligned}(\forall \alpha \in A) S_\alpha \text{ is convex cone } \implies \\ \cap_{\alpha \in A} S_\alpha \text{ is convex cone}\end{aligned}$$

Any closed convex set can be represented by possibly infinitely many half spaces.

0.6.2 Affine functions

let $f(x) = Ax + b, f : \mathbb{R}^n \to \mathbb{R}^m$
then if S is a convex set we have:

- project forward: $f(S) = \{f(X) : X \in S\}$ is convex
- project back: $f^{-1}(S) = \{X : f(X) \in S\}$ is convex

Example:

$$C = \{y : y = Ax + b, \|x\| \leq 1\}$$

$\|x\| \leq 1$ is convex, $Ax + b$ is affine $\implies$ C is convex
Example:

$$C = \{x : \|Ax + b\| \leq 1\}$$

$\{y : \|y\| \leq 1\}$ is convex $\wedge$ y is an affine function of $x \implies$ C is convex

0.7 Constraint Qualifications

0.7.1 Slater's Constraint Qual.

Optimal solution is in relative interior: $x^* \in relint(S)$

Inequalities $(\forall i) f_i(x)$ convex $\wedge$ $f_i(x) < 0 \implies$ Slater's constraint satisfied.

Inequalities $(\forall i) f_i(x)$ affine $\implies$ $(\forall i) f_i(x) \leq 0 \wedge (\exists i) f_i(x) < 0 \implies$ Slater's constraint satisfied.

Achieving Slater's constraint implies 0 duality gap.

0.7.2 KKT

Assumes optimality achieved with 0 duality gap: $\nabla L(x^*, \lambda^*, v^*) = 0$

$$L(x^*, \lambda^*, v^*) = f_0(x^*) + \sum_i \lambda_i^* f_i(x^*) + \sum_i v_i h_i(x^*)$$
$$\nabla L(x^*, \lambda^*, v^*) = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_i v_i \nabla h_i(x^*)$$

We have the constraints:

$$f_i(x^*) \leq 0$$
$$\lambda_i^* \geq 0$$
$$h_i(x^*) = 0$$
$$\lambda_i^* f_i(x^*) = 0$$
$$\nabla L(x^*, \lambda^*, v^*) = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_i v_i \nabla h_i(x^*)$$

Primal inequality constraints convex and equality constraints affine and KKT satisfied $\implies$ 0 duality gap with specified points for primal and dual. (sufficient).

If Slater's constraint satisfied then the above is sufficient and necessary:
Primal inequality constraints convex and equality constraints affine and KKT satisfied $\iff$ 0 duality gap with specified points for primal and dual.

0.8 Definitions

0.8.1 Convex Function

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \forall \theta = [0,1]$$

For convenience we sometimes define an extended value function:

$$\tilde{f}(x) = \begin{cases} f(x), & x \in dom(f) \\ \infty, & other\ wise \end{cases}$$

if $f(x)$ convex, then $\tilde{f}$ is also convex

Sublevel set of a function

$$C(\alpha) = \{x \in dom(f) : f(x) \leq \alpha\}$$

For convex function, all sublevel sets are convex ($\forall \alpha$). Converse is not true.

Quasi-convex function: if its sublevel sets are all convex.

Epigraph of functions:
$epi(f) = \{(x,t) : x \in dom(f), f(x) \leq t\} \in \mathbb{R}^{n+1}$,
$f \in \mathbb{R}^n \to \mathbb{R}$.

f is convex function $\iff$ $epi(f)$ is convex set

0.8.2 First order condition

Suppose f is differentiable and domain of f is convex. Then:
f convex $\iff$
$(\forall x, x_0 \in dom(f)) f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0)$

rough proof:
suppose $f(x)$ is convex but $(\exists x, x_0) f(x) < f(x) + \nabla f(x_0)^T(x - x_0)$
then this means the function should bend across the tangent line which violates the convexity

proof for converse direction:
suppose that $(\exists x, x_0) f(x) \geq f(x) + \nabla f(x_0)^T(x - x_0)$
to show that $f(x)$ is convex lets take $x, y \in dom(f), z = \theta x + (1-\theta)y$
$\theta f(x) + (1-\theta)f(y) \geq f(z) + \nabla f(z)^T(\theta x - \theta z + (1-\theta)y - (1-\theta)z)$
$\theta f(x) + (1-\theta)f(y) \geq f(\theta x + (1-\theta)y)$
$f(x)$ is convex

0.8.3 Second order condition

Suppose f is twice differentiable and $dom(f)$ is convex,
then $f(x)$ is convex $\iff$ $\nabla^2 f(x) \geq 0$ (PSD, eg: wrt. S^n_+)

proof for scalar case:
suppose that $f(x)$ is convex, then the first-order condition holds
for $x, y \in dom(f) : f(x) \geq f(y) + f'(y)(x - y)$
for $y, x \in dom(f) : f(y) \geq f(x) + f'(x)(y - x)$
$f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x)$
$f'(x)(y - x) \leq f'(y)(y - x) \implies 0 \leq (y - x)(f'(y) - f'(x))$
take $y \to x : 0 \leq f''(x)$
$f''(x) \geq \frac{f'(x+\delta x) - f'(x)}{\delta x}$

conversely, suppose that $f'(z) \geq 0, \forall z \in dom(f)$, take $x, y \in dom(f)$ WLOG $x < y$
$\int_x^y f''(z)(y - z)dz \geq 0$
$f''(z) \geq 0, (y - z) \geq 0$
$I_1 = \int_x^y f''(z)ydz = y(f'(y) - f'(x))$
$I_2 = -\int_x^y zf''(z)dz$
$dv = f''(z)dz \implies v = f'(z)$
$u = z \implies du = dz$
$I_2 = -[zf'(z)]|_x^y + \int_x^y f'(z)dz = -yf'(y) + xf'(x) + f(y) - f(x)$
$I_1 + I_2 = yf'(y) - yf'(x) - yf'(y) + xf'(x) + f(y) - f(x) \geq 0$
$\implies f(y) \geq f(x) + f'(x)(y - x)$ first order condition: $x < y$
first order condition holds $\implies$ $f(x)$ convex

0.8.4 Inequalities

$x \preceq_K y \iff y - x \in K$

x is a minimum in S wrt. cone K:

$x \in S : (\forall y \in S) f(y) \succeq_K f(x) \iff f(y) - f(x) \in S$
$S \subseteq x + K$

x is a minimal in S wrt. cone K:

$$x \in S : (\forall y \in S) f(y) \preceq_K f(x) \implies x = y$$
$$x \in S : (\forall y \in S) f(x) - f(y) \in K \implies x = y$$
$$(x - K) \cap S = \{x\}$$

0.8.5 Cone

$$(\forall x \in C, \forall \theta \geq 0) \theta x \in C$$

0.8.6 Convex Cone

Eg: S^n, S^n_+ are convex cones
convexity check for S^N_+:

$$\begin{aligned} x_1 \in S^n_+ &\implies v^T x_1 v \geq 0 \\ x_2 \in S^n_+ &\implies v^T x_2 v \geq 0 \\ x = \theta x_1 + (1-\theta) x_2&, \theta \in [0,1] \\ v^T(\theta x_1 + (1-\theta) x_2) v &\geq 0 \\ v^T \theta x_1 v + (1-\theta) v^T x_2 v &\implies x \in S^n_+ \end{aligned}$$

convexity check for cone:

$$\begin{aligned} x_1 \in S^n_+ &\implies \theta x_1 \in S^n_+, \theta \geq 0 \\ (\forall v) v^T x v \geq 0 &\implies v^T(\theta x) v \geq 0 \implies \text{cone} \end{aligned}$$

0.8.7 Proper Cone

Definition:

- convex
- closed(contains all limit points)
- solid(non-empty interior)
- pointed(contains no line): $x \in K \implies -x \notin K$

Then the proper cone K defines a generalized inequality ($\leq_K$) in $\mathbb{R}^n$

$$x \leq_K y \implies y - x \in K$$
$$x <_K y \implies y - x \in int(K)$$

Example: $K = R^n_+$ (non-negative orthant):

$$n = 2$$
$$x \leq_{R^2_+} y \implies y - x \in R^2_+$$

Cone provides partial ordering using difference of 2 objects.
$X \leq_{S^n_+} Y \iff Y - X \in S^n_+ \iff Y - X$ is PSD

0.8.8 Norm Cone

$$K = \{(x,t) \in \mathbb{R}^{n+1} : \|x\| \leq t\}, x \in \mathbb{R}^n$$

0.8.9 Dual norm

$\|z\|_* := \sup_x \{z^T x : \|x\|_p \leq 1\}$

Dual of L1-norm:
$\|z\|_* := \sup_x \{z^T x : \|x\|_1 \leq 1\}$
max $\sum_i z_i x_i$,
subject to : $\sum_i \|x_i\| \leq 1$
select x_i corresponding to z_i with maximum absolute value
equivalent to $\|z\|_* = \|z\|_\infty$

Dual of L-∞-norm:
$\|z\|_* := \sup_x \{z^T x : \|x\|_\infty \leq 1\}$
max $\sum_i z_i x_i$,
subject to : $\|x_i\| \leq 1, \forall i$
choose $x_i = 1$ if $z_i \geq 0$ and $x_i = -1$ if $z_i < 0$
equivalent to $\|z\|_* = \|z\|_1$

Dual norm of Lp-norm: Lq-norm where $1/p + 1/q = 1$

Properties of Dual Cone:

- K^* closed and convex
- $K_1 \subseteq K_2 \implies K_2^* \subseteq K_1^*$
- K has non-empty interior $\implies K^*$ pointed
- $cl(K)$ pointed $\implies K^*$ has non-tempty interior
- $K^{**} = cl(convhull(K))$, useful for relaxed optimization
- K convex and closed $\implies K = K^{**}$

0.8.10 Operator norm

$\|X\|_{a,b}= sup\{\|Xu\|_a: \|u\|_b\leq 1\}, X \in \mathbb{R}^{m\times n}$

0.8.11 Dual cone

$$\begin{aligned} K\ is\ a\ cone \\ K^* = \{y : x^T y \geq 0, \forall x \in K\} \end{aligned}$$

0.8.12 Dual norm cone

$$\begin{aligned} K^* = \{(u, v) : \|u\|_* \leq v\} \\ where\ K = \{(x, t) : \|x\| \leq t\} \end{aligned}$$

0.8.13 support function of a set

$$\begin{aligned} S_C(x) = \sup\{x^T y : y \in C\} \\ dom(S_C) = \{x : \sup_{y\in C} x^T y < \infty\} \end{aligned}$$

It is pointwise supremum of convex function, so it is convex.

0.9 Relaxation

projection onto the feasible set

take a larger feasible set and optimize in it instead, resulting optimal value is smaller or equal to the original

equality contraints that are convex but not affine, make them inequalities thus transforming to convex problem

0.10 Regularized Approximation

Noise sensitivity of different objectives:

- robust least squares / Huber penalty
- log barrier
- deadzone linear
- quadratic

Least norm problems

- L2 norm objective with equality constraint
- sparsity inducing norms (eg: L1)
- norm ball constraint
- probability distribution
  - convex comb. of columns of A to fit b
- variable constraints
  - box
  - one sidded bound

Multicriterion Formulation

Tikhonov Regularization
todo..

0.11 Descent Methods

Algorithm 1: Descent Overview

```
init $x_0 \in dom f$;
while stopping criterion is not satisfied
    $\Delta x \leftarrow$ Compute Descent Direction;
    $t \leftarrow$ Compute Descent Step Size;
    $x_{k+1} \leftarrow x_k + t\Delta x$
```

0.11.1 Search Step Size

Given Δx, step direction
Search step size:
Exact Line Search: $t = \operatorname{argmin}_{s \geq 0} f(x + s\Delta x)$

Algorithm 2: Backtracking Line Search

```
$\Delta x$: Search direction
t   : Step size
$\alpha \in (0, 0.5)$;
$\beta \in (0, 1)$;
$t \leftarrow 1$;
while $f(x + t\Delta x > f(x) + \alpha t \nabla f(x)^T \Delta x$
    $t \leftarrow \beta t$;
```

0.11.2 Search Direction - 1st Order Methods

Steepest Descent

$$\underset{\Delta x}{\operatorname{argmin}}\, f(x) + \nabla f(x)^T \Delta x_{sd}$$
$$\underset{\Delta x}{\operatorname{argmin}}\, \nabla f(x)^T \Delta x_{sd}$$

Normalized Steepest Descent

$$\Delta x_{nsd} = \underset{v}{\operatorname{argmin}}\{\nabla f(x)^T v : \|v\| \leq 1\}$$
$$\Delta x_{nsd} = \frac{\Delta x_{sd}}{\|\nabla f(x)\|_*}$$
$$\|\nabla f(x)\|_* = \sup_y \{\nabla f(x)^T y : \|y\| \leq 1\}\ (dualnorm)$$
$$\nabla f(x)^T \Delta x_{sd} = \nabla f(x)^T \Delta x_{nsd} \|\nabla f(x)\|_* = \|\nabla f(x)\|_*^2$$

Steepest Descent for L2-Norm

$$\Delta x = -\nabla f(x)$$
$$\Delta x_{nsd} = \frac{-\nabla f(x)}{\|\nabla f(x)\|_*}$$

Steepest Descent for Quadratic Norm

$$\|z\|_P = (z^T P z)^{\frac{1}{2}} = \|P^{\frac{1}{2}} z\|_2, P \in S^n_{++}$$
$$\Delta x_{nsd} = -(\nabla f(x)^T P^{-1} \nabla f(x))^{-\frac{1}{2}} P^{-1} \nabla f(x)$$
$$\Delta x_{sd} = -P^{-1} \nabla f(x)$$

Normalized Steepest Descent for L1-norm

$$\Delta x_{nsd} = \underset{v}{\operatorname{argmin}}\{\nabla f(x)^T v : \|v\|_1 \leq 1\}$$
$$\Delta x_{nsd} = -sign(\frac{\partial f(x)}{\partial x_i})\ e_i, i : \|\nabla f(x)\|_\infty = |(\nabla f(x))_i|$$

Search Direction - 2nd Order Methods, Unconstrained

Newton's Method

$$\begin{aligned}&\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)\\&\nabla^2 f(x)^{-1} \succ 0 \implies\\&\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0\end{aligned}$$

Affine invariance of Newton's method: auto-scaling of level curves to enable better descent direction

2nd order Taylor approximation of function is minimized with Δx_{nt}, so quadratic model gives good approximation near minimizer.

Newton Decrement: used for convergence/stopping criterion

Norm of Newton Step:

$$\lambda(x) = (\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt})^{\frac{1}{2}} = \|(\nabla^2 f(x) \Delta x_{nt})^{\frac{1}{2}}\|_2$$

Bounding difference of lower bound of 2nd order model and $f(x)$:

$$\begin{aligned}f(x) - \inf\{f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v\\ : A(x+v) = b\} = \frac{1}{2}\lambda(x)^2\end{aligned}$$

$$\begin{aligned}&\frac{1}{2}\lambda(x)^2 \approx f(x) - p^*\\&\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\\&\nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2 = \frac{d}{dt} f(x + \Delta x_{nt} t)|_{t=0}\end{aligned}$$

Algorithm 3: Newton Method Descent

```
init $x_0 \in dom f$;
do
    $\Delta x_{nt} \leftarrow -\nabla^2 f(x)^{-1} \nabla f(x)$;
    $\lambda(x)^2 \leftarrow \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$;
    $t \leftarrow$ Compute Step Size (line search);
    $x_{k+1} \leftarrow x_k + t\Delta x_{nt}$
while $\frac{1}{2}\lambda(x)^2 > \epsilon$;
```

Assumptions of above algorithm: KKT matrix invertible. Descent method is decreasing wrt. f.

Convergence: split into damped phase and quadratically convergent phase. Fast convergence once step size is 1.

BFGS: a Quasi-newton method
Direction:

$$p_k = -B_k^{-1} \nabla f_k$$

Properties for B:
$B \succ 0$
$B = B^T$
Satisfaction of Secant equation:

$$\begin{aligned}&B_{k+1}(x_{k+1} - x_k) = \nabla f_{k+1} - \nabla f_K\\&B_{k+1} s_k = y_k\end{aligned}$$

Curvature condition:

$$\begin{aligned}&B_{k+1} s_k = y_k\\&s_k^T B_{k+1} s_k = s_k^T y_k\\&B_{k+1} \succ 0 \implies s_k^T y_k > 0\end{aligned}$$

This need to be enforced if doing optimization on a non-convex function

General problem of find B

$$\begin{aligned}&\min_B \|B - B_k\|\\&s.t.\ B = B^T, B s_k = y_k\end{aligned}$$

BFGS Update: $B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k} - \frac{B_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} B_k^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}} B_k \mathbf{s}_k}$

Algorithm 4: BFGS Descent

```
init $x_0 \in domf$
init $B_0$ (eg: $I$)
while $\|y_k\| > \epsilon$
    solve for $p_k$ in $B_k p_k = -\nabla f(x_k)$
    $t \leftarrow \text{argmin}_s f(x_k + s p_k)$(or backtrack search)
    $s_k \leftarrow t p_k$
    $x_{k+1} \leftarrow x_k + s_k$
    $y_k = f(x_{k+1}) - f(x_k)$
    $B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k} - \frac{B_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} B_k^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}} B_k \mathbf{s}_k}$
```

Inversion of B via Sherman-Morrison:
$B_{k+1}^{-1} = \left(I - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) B_k^{-1} \left(I - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) + \frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}.$

Sherman-Morrison:
A^{-1} exists: $(A + uv^T)^{-1}$ exists $\iff 1 + v^T A^{-1} u \neq 0$
$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}$

0.12 Equality Constrained Optimization

Approaches:

- elimination of equality constraint, variable substitution into objective, unconstrained opt.
- Newton's Method with equality constraint, and assumed feasible start, unconstrained opt.
- Newton's Method with equality constraint, infeasible start,
  equivalence using primal-dual residual method

KKT optimality, with x assumed to be feasible:

$$\Delta x = \underset{v}{\operatorname{argmin}}\, f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$
$$s.t.\ A(x+v) = b$$
$$optimality\ condition:$$
$$L(v,w) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$
$$+ w^T(A(x+v) - b)$$
$$\frac{\partial L(v,w)}{\partial v} = \nabla f(x) + \nabla^2 f(x) v + A^T w = 0$$
$$A(x+v) = b$$
$$Ax = b \implies Av = 0$$
$$KKT\ system:$$
$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = K \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$
$$\begin{bmatrix} v \\ w \end{bmatrix} = K^{-1} \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}, K^{-1}\ exists$$
$$\Delta x = v$$

Alternatively Solve via elimination

$$\nabla f(x) + \nabla^2 f(x) v + A^T w = 0$$
$$v + \nabla^2 f(x)^{-1} A^T w + \nabla^2 f(x)^{-1} \nabla f(x) = 0$$
$$Av + A\nabla^2 f(x)^{-1} A^T w + A\nabla^2 f(x)^{-1} \nabla f(x) = 0$$
$$A\nabla^2 f(x)^{-1} A^T w + A\nabla^2 f(x)^{-1} \nabla f(x) = 0$$
$$w = -(A\nabla^2 f(x)^{-1} A^T)^{-1} A\nabla^2 f(x)^{-1} \nabla f(x)$$
$$v = \nabla^2 f(x)^{-1}(-\nabla f(x) - A^T w)$$
$$\Delta x = v$$

$\nabla^2 f(x)$ not invertible, augment KKT system s.t. $(\exists Q)\nabla^2 f(x) + AQA > 0$

$$\begin{bmatrix} \nabla^2 f(x) + AQA & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$
$$Av = 0$$
$$AQAv = 0 \implies solution\ same\ to\ original\ problem$$

Algorithm 5: Newton Method w/ Equality Constraint

```
init $x_0 \in dom f, Ax_0 = b$;
do
    $\Delta x_{nt} \leftarrow$ Solve KKT System / Elimination;
    $\lambda(x)^2 \leftarrow \Delta x_{nt}^T \nabla^2 f(x)^{-1} \Delta x_{nt}$;
    $t \leftarrow$ Compute Step Size (eg: backtrack);
    $x_{k+1} \leftarrow x_k + t\Delta x_{nt}$
while $\frac{1}{2}\lambda(x)^2 > \epsilon$;
```

0.12.1 Infeasible Start Newton Method

Modify KKT system for residual: $Av = b - Ax$

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = K \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ b - Ax \end{bmatrix}$$

Once step length becomes 1, all following iterates are feasible.

Equivalence with Primal-Dual residual update:

$$\min_x f(x), s.t.\ Ax = b$$

Dual:

$$L(x,v) = f(x) + v^T(Ax - b)$$
$$s.t.\ Ax - b = 0$$
$$optimality:$$
$$\frac{\partial L}{\partial x} = \nabla f(x) + A^T v = 0$$

residual of primal and dual:

$$r = \begin{bmatrix} r_{dual} \\ r_{pri} \end{bmatrix} = \begin{bmatrix} \nabla f(x) + A^T v \\ Ax - b \end{bmatrix}$$

$$y = \begin{bmatrix} x \\ v \end{bmatrix}$$
$$\Delta y = \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix}$$
$$r(y + \Delta y) \approx r(y) + Dr(y)\Delta y$$
$$Dr(y) := Gradient\ of\ r(y)$$
$$Dr(y) = \begin{bmatrix} \nabla_x r_{dual}^T & \nabla_v r_{dual}^T \\ \nabla_x r_{pri}^T & \nabla_v r_{pri}^T \end{bmatrix}$$
$$Dr(y) = \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}$$

goal: $r(y + \Delta y) \to 0$

$$
\begin{aligned}
&r(y) + Dr(y)\Delta y = 0 \\
&\begin{bmatrix} \nabla f(x) + A^T v \\ Ax - b \end{bmatrix} + \begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix} = 0 \\
&\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix} = \begin{bmatrix} -\nabla f(x) - A^T v \\ b - Ax \end{bmatrix}
\end{aligned}
$$

comparison with defintion of infeasible start Newton Method:

$$
\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ b - Ax \end{bmatrix}
$$

Equivalence if for infeasible start, we select:

$$
\begin{aligned}
&v = \Delta x \\
&A^T(\Delta v + v) = A^T w = 0 \\
&w = v + \Delta v
\end{aligned}
$$

Solving using primal-dual formulation, obtain $\Delta v, \Delta x$
Backtrack line search using $\|r\|$ instead of f

Algorithm 6: Newton Method w/ Infeasible Start

```
init $x_0 \in dom f$
$\beta \in (0,1)$
$\alpha \in (0,0.5)$
do
    $\Delta v, \Delta x \leftarrow$ Solve primal-dual KKT system
    Backtrack Step Size Search with $\|r\|$:
    $t \leftarrow 1$
    while $\|r(x + t\Delta x, v + t\Delta v)\| > (1 - \alpha t)\|r(x,v)\|$
        $t \leftarrow \beta t$
    $x_{k+1} \leftarrow x_k + t\Delta x$
    $v_{k+1} \leftarrow v_k + t\Delta v$
while $\|r(x,v)\| > \epsilon \vee Ax \neq b$;
```

0.13 Inequality Constrained Optimization

Ideas:

- gradient projection: update x with descent direction, then project back onto feasibility set. Assume feasibility set is convex.

  $$\begin{aligned}
  &x \leftarrow descent\ update\\
  &\tilde{s} = \underset{s}{\operatorname{argmin}} \|x - s\|, s.t.\ s \in X = feasibility\ set\\
  &let\ [x]^+ = \underset{s}{\operatorname{argmin}} \|x - s\|, s.t.\ s \in X\\
  &x_{k+1} \leftarrow [x_k + t\Delta x]^+, eg : \Delta x = \nabla f(x)
  \end{aligned}$$

  Issues: projection hard in general, line search becomes harder. Ok for simple projection (eg: box constraints).

- Adapt Newton's Method with projection.

  $$\begin{aligned}
  \min_v \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v\\
  s.t.\ x_{k+1} = x_k + v \in X
  \end{aligned}$$

  Issues: Hard to solve.

- elimination inequality constraint and augment objective (eg: Interior Point)

Interior Point with Inequality & Equality Constraints

$$\begin{aligned}
&\min_x f_0(x)\\
&s.t.\ f_i(x) \leq 0, \forall i \in \{1, .., m\}\\
&Ax = b
\end{aligned}$$

Assumptions:

- solution exists
- strict feasibility (Slater's cond. hold), so strong duality
- objective and inequality functions differentiable and convex

Barrier Method:

$$\begin{aligned}
&\min_x f_0(x) + \sum_{i=1}^{m} I(f_i(x))\\
&s.t.\ Ax = b\\
&I(u) = \begin{cases} 0 & , u \leq 0\\ +\infty & , o/w \end{cases}
\end{aligned}$$

$I(u)$ convex, non-decreasing, but non-differentiable.

Use log barrier function to approximate $I(u)$

$$\begin{aligned}
\hat{I}(u) &= -\frac{1}{t} log(-u)\\
\hat{I}(u)|_{t \to +\infty} &\to I(u)
\end{aligned}$$

Log barrier:

$$\Phi(x) = -\sum_i log(-f(x))$$

$$\begin{aligned}
&\min_x f_0(x) - \frac{1}{t} \sum_{i=1}^{m} log(-f_i(x))\\
&\min_x f_0(x) + \frac{1}{t} \Phi(x)\\
&\min_x\ t f_0(x) + \Phi(x)\\
&s.t.\ Ax = b
\end{aligned}$$

Approach: fix t, optimize problem to obtain $x^*(t)$. Repeat for t increasing in value with previously solved intermediate solution so that $x^*(t)|_{t \to +\infty} = x^*$.

Find a starting strictly feasible point x. Set $t_0 > 0$. Inner optimization problem, solve $x^*(t)$(via iterative algo such as Newton):

Algorithm 7: Log Barrier Method

```
do
    $x^*(t) = \min_x\ t f_0(x) + \Phi(x)$, s.t. $Ax = b$
    update:
    $x = x^*(t)$
    $t \leftarrow \mu t, \mu > 1$ (eg: 10)
while $\frac{m}{t} < \epsilon$ not met;
```

Central Path:
Successive solving of inner optimization problem with varying t traces out a path, leading closer to the optimal final solution.

Central path lies in the interior of the feasibility region, thus potential optimal solution on the boundary of the feasibility set is never exactly reached.

Inner optimization usually uses 2nd order methods for auto-scaling of level set of optimization objective when t is large.

Search for Initial Feasible Point (Phase 1):
Formulate problem as:

$$\begin{aligned}&\min_{x,s} s\\ &s.t.\ f_i(x) \leq s, \forall i\\ &Ax = b\end{aligned}$$

If $s^* < 0$, x^* is strictly feasible. Else infeasibility.

Initialization of Phase 1 problem: set $s > \max_i f_i(x)$ so that starting point is interior of feasible set. Then proceed to solve using Interior Point method.

If $s^* < 0$, then original problem is solved using x^* as the strictly feasible starting point.

Interpretation of Interior Point Method with KKT:
Primal of Original:

$$\begin{aligned}&\min_{x} f_0(x)\\ &s.t.\ f_i(x) \leq 0, \forall i \in \{1, .., m\}\\ &Ax = b\end{aligned}$$

Lagrangian of Primal:
$f_0(x) + \sum_i \lambda_i f_i(x) + w^T(Ax - b)$
KKT optimality conditions:

$$\begin{aligned}&\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) + A^T w = 0\\ &\lambda_i \geq 0\\ &Ax - b = 0\\ &\lambda_i f_i(x) = 0\\ &f_i(x) \leq 0\end{aligned}$$

Primal of Log-Barrier:

$$\begin{aligned}&\min_{x} f_0(x) - \frac{1}{t}\sum_i log(-f_i(x))\\ &s.t.\ Ax = b\end{aligned}$$

KKT optimality conditions for Log-Barrier problem:

$$\begin{aligned}&\nabla f_0(x) - \frac{1}{t}\sum_i \frac{\nabla f(x)}{f_i(x)} + A^T w = 0\\ &Ax - b = 0\\ &f_i(x) \leq 0\ (domain\ of\ log)\\ &let\ \lambda_i = -\frac{1}{t f_i(x)}\\ &\nabla f_0(x) + \sum_i \lambda_i \nabla f(x) + A^T w = 0\\ &Ax - b = 0\end{aligned}$$

$$t > 0, f_i(x) \leq 0 \implies \lambda_i = -\frac{1}{t f_i(x)} \geq 0$$

Overall (Modified) KKT conditions for Log-Barrier:

$$\begin{aligned}&\nabla f_0(x) + \sum_i \lambda_i \nabla f(x) + A^T w = 0\\ &f_i(x) \leq 0\\ &Ax - b = 0\\ &\lambda_i \geq 0\\ &\lambda_i f_i(x) = -\frac{1}{t}\end{aligned}$$

This approaches true KKT conditions as $t \to +\infty$.

Obtaining a bound on duality gap of the modified problem from original:
Dual of original:

$$\begin{aligned}
&g(\lambda) = \min_{Ax=b} f_0(x) + \sum_i \lambda_i f_i(x)\\
&g(\lambda) \leq f_0(x) + \sum_i \lambda_i f_i(x), Ax = b\\
&g(\lambda) \leq f_0(x^*) + \sum_i \lambda_i f_i(x^*), Ax^* = b\\
&(\forall i) f_i(x^*) \leq 0, \lambda_i \geq 0 \implies\\
&g(\lambda) \leq f_0(x^*) + \sum_i \lambda_i f_i(x^*) \leq f_0(x^*)\\
&g(\lambda) \leq f_0(x^*), \lambda \geq 0
\end{aligned}$$

Log-Barrier Problem:

$$\begin{aligned}
&\min_x f_0(x) - \frac{1}{t} \sum_i log(-f_i(x))\\
&s.t.\ Ax = b
\end{aligned}$$

KKT optimality conditions:

$$\begin{aligned}
&\nabla f_0(x) + \sum_i -\frac{1}{tf_i(x)} \nabla f(x) + A^T w = 0\\
&f_i(x) \leq 0\\
&Ax - b = 0\\
&from\ dual:\\
&g(\lambda) \leq f_0(x) + \sum_i \lambda_i f_i(x), Ax = b\\
&\lambda_i = -\frac{1}{tf_i(x)}, f_i(x) \leq 0 \implies same\ KKT\ conditions\\
&x^*(t)\ minimizes\ for\ particular\ \lambda_i^*(t) = -\frac{1}{tf_i(x)}\\
&g(\lambda^*(t)) = f_0(x^*(t)) + \sum_i \lambda_i^*(t) f_i(x^*(t)), Ax^*(t) = b\\
&g(\lambda^*(t)) = f_0(x^*(t)) + \sum_i -\frac{1}{tf_i(x^*(t))} f_i(x^*(t))\\
&g(\lambda^*(t)) = f_0(x^*(t)) + \sum_{i=1}^m -\frac{1}{t}, Ax^*(t) = b\\
&g(\lambda^*(t)) = f_0(x^*(t)) - \frac{m}{t} \leq f_0(x^*)\\
&f_0(x^*(t)) \leq f_0(x^*) + \frac{m}{t}\\
&f_0(x^*) \leq f_0(x^*(t)) \implies gives\ bound\ of\ gap: \frac{m}{t}
\end{aligned}$$

$x^*(t)$ gives solution that is bounded by $\frac{m}{t}$ away from optimal wrt. objective.
$\frac{m}{t}$ usable as stopping criterion.

Primal-Dual Interior Point Method (Alternative to Barrier Method)

Overall (Modified) KKT conditions for Log-Barrier:

$$\begin{aligned}
&\nabla f_0(x) + \sum_i \lambda_i \nabla f(x) + A^T v = 0\\
&f_i(x) \leq 0\\
&Ax - b = 0\\
&\lambda_i \geq 0\\
&-\lambda_i f_i(x) - \frac{1}{t} = 0
\end{aligned}$$

$$y = \begin{bmatrix} x \\ \lambda \\ v \end{bmatrix}$$

$$r(y) = \begin{bmatrix} \nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) + A^T v \\ -\lambda_1 f_1(x) - \frac{1}{t} = 0 \\ .. \\ -\lambda_m f_m(x) - \frac{1}{t} = 0 \\ Ax - b \end{bmatrix}$$

Goal: drive $r(y)$ to 0

1st order approx of $r(y + \Delta y)$:

$$r(y + \Delta y) \approx r(y) + Dr(y)\Delta y = 0$$

$$Dr(y) = \begin{bmatrix} \nabla^2 f_0(x) + \sum_i \lambda_i \nabla^2 f(x) & f_1'(x) & .. & f_m'(x) & A^T \\ -\lambda_1 f_1'(x) & -f_1(x) & 0 & .. & \\ .. & & .. & & \\ -\lambda_m f_m'(x) & & 0 & -f_m(x) & 0 \\ A & 0 & .. & & \end{bmatrix}$$

$$Dr(y)\Delta y = -r(y)$$

Overall algorithm:

Algorithm 8: IP Primal-Dual Method

```
init strictly feasible
  $x, \lambda > 0, \eta = -\sum_i \lambda_i f_i(x), \mu = 10$
$\eta = -\sum_i \lambda_i f_i(x)$
do
    $t \leftarrow \frac{\mu m}{\eta}$
    Solve for $\Delta y$ in $Dr(y)\Delta y = -r(y)$
    Line search for step size, $s$, using norm of
      residual $\|r\|$, s.t. $\lambda > 0, f(x) < 0$
    $y \leftarrow y + s\Delta y$
    $\eta = -\sum_i \lambda_i f_i(x)$ (proxy for duality gap,
      the smaller the better)
while $\|r_{prim}\| > \epsilon \vee \|r_{dual}\| > \epsilon \vee \eta > \epsilon_2$;
```

Note: one loop only, but stopping in middle of algorithm does not guarantee a feasible solution.

0.14 Ellipsoid Method

f is convex, C^1 only

$(\forall x, x_0) f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0)$

$\nabla f(x_0)^T(x - x_0) \geq 0 \implies f(x) \geq x(x_0)$

Idea: use of halfspace to perform elimination of search space

Selection of potential search points: use centroid of feasible polyhedron to maximize elimination space

Volumne reduction: $(X_{max} - X_{min})/2^k$

Algorithm 9: Halfspace Elimination

```
do
    update select $x^{(k)}$ using center of $C^{(k)}$ where $x^* \in C^{(k-1)}$
    $(\forall i) C^{(k-1)} = \{x : \nabla f(x^(i))^T(x - x^{(i)}) \leq 0\}$
    bisection:
    $f'(x^{(k)}(x - x^{(k)})$ eliminates 1 side of halfspace
    $C^{(k)} = C^{(k-1)} \cap \{x : \nabla f(x^{(k)})^T(x - x^{(k)}) \leq 0\}$
while Not Satisfy stopping criterion;
```

Pros:

- cnetroid selection easy
- bisection evaluation easy
- exponential rate

Implementation issues: centroid search, intersection in high dimension

Simplify: ellipsoid instead of polyhedral for localization

Algorithm 10: Ellipsoid Method

```
do
    set $x^{(k+1)}$ as center of $\varepsilon^{(k)}$
    eval $\nabla f(x^{(k+1)})$
    find min. volume ellipsoid covering:
    $S = \{x : \nabla f(x^{(k+1)})(x - x^{(k+1)}) \leq 0\}$
    $\varepsilon^{(k+1)} = \varepsilon \cap S$
while Not Satisfy stopping criterion;
```

Intersection search:

$g = \nabla f(x^{(k+!)})$

$\tilde{g} = \frac{g}{\sqrt{g^T P g}}$

$\varepsilon^{(k)} = \{x : (x - x^{(k+1)})^T P^{-1}(x - x^{(k+1)}) \leq 1\}$

$\varepsilon^{(k+1)} = \{x : (x - x^+)^T (P^+)^{-1}(x - x^+) \leq 1\}$

$x^+ = x^{(k-1)} - \frac{1}{n+1} P\tilde{g}, n = dim\ of\ X$

$P^+ = \frac{n^2}{n^2-1}(P - \frac{2}{n+1} P\tilde{g}\tilde{g}^T P)$

$Vol(\varepsilon^{(k+1)}) \leq e^{-\frac{1}{2^n}} Vol(\varepsilon^{(k)})$

0.15 Subgradient Method

useful for f convex, but not C^1, C^2, and when subgradient of certain form is known

Idea: need a g such that:

$(\forall x) f(x) \geq f(x_0) + g^T(x - x_0)$

subdifferential:

$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$

properties:

closed and convex for all f (nonconvex as well)

can be empty for nonconvex f

f is atleast C^1 $at\ x \iff \partial f(x) = \{\nabla f(x)\}$

$f\ convex \implies f(x^*) = min_x f(x) \iff 0 \in \partial f(x^*)$

subgradient calculus:

$$(\forall a > 0)\partial(af) = a\partial f$$

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2$$

$$g(x) = f(Ax + b) \implies \partial g(x) = A^T \partial f(Ax + b)$$

$$f(x) = max_i f_i(x) \implies$$

$$\partial f(x) = conv\Bigg(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x)\Bigg)$$

subgradient descent:

eg: $\sum_k^\infty t_k^2 < \infty, \sum_k^\infty t_k = \infty, x^{k+1} = x^k - t_k g^k$

keep track of the best x

0.15.1 Application

Decomposition of LP into decoupled systems:

Original LP:

$$\min_{u,v} c^T u + d^T v$$

$$s.t. Au \leq b$$

$$Pv \leq q$$

$$Fu + Gv \leq h$$

Augment objective w/ constraint, partial Lagrandian

$$L(u, v, \lambda) = c^T u + d^t v + \lambda^T(Fu + Gv - h)$$

$$g(\lambda) = \min_{u,v} c^T u + d^t v + \lambda^T(Fu + Gv - h)$$

$$s.t. Au \leq b$$

$$Pv \leq q$$

For fixed λ, we can solve 2 separate LPs involving only u and v

$$\begin{aligned}&\min_u c^T u + \lambda^T Fu\\ &s.t. Au \leq b\end{aligned}$$

$$\begin{aligned}&\min_v d^T v + \lambda^T Gv\\ &s.t. Pv \leq q\end{aligned}$$

max . $g(\lambda)$ s.t. $\lambda \geq 0$ gives optimal λ^*

$g(\lambda)$ concave, but not differentiable
Use subgradient, s, s.t. $g(\tilde{\lambda}) \leq g(\lambda) + s^T(\tilde{\lambda} - \lambda), \forall \tilde{\lambda}$

(u^*, v^*) for a fixed λ:
$S = Fu^* + Gv^* - h$ is a subgradient of g at λ

$g(\tilde{\lambda}) = \min_{u,v} c^T u + d^T v + \tilde{\lambda}^T(Fu + Gv - h) \leq c^T u^* + d^T v^* + \tilde{\lambda}^T(Fu^8 + Gv^* - h)$

For $\forall u, v$ feasible:

$$\begin{aligned}g(\tilde{\lambda}) = &\min_{u,v} c^T u + d^T v + \tilde{\lambda}^T(Fu + Gv - h)\\ &s.t. Au \leq b\\ &\quad Pv \leq q\end{aligned}$$

$$\begin{aligned}g(\tilde{\lambda}) &\leq c^T u^* + d^T v^* + \tilde{\lambda}^T(Fu^* + Gv^* - h)\\ &+ \lambda^T(Fu^* + Gv^* - h)\\ &- \lambda^T(Fu^* + Gv^* - h)\\ g(\lambda) &= c^T u^* + d^T v^* + \lambda^T(Fu^* + Gv^* - h)\\ g(\tilde{\lambda}) &= g(\lambda) + (\tilde{\lambda} - \lambda)^T(Fu^* + Gv^* - h)\\ &= g(\lambda) + s^T(\tilde{\lambda} - \lambda)\end{aligned}$$

$s = Fu^* + Gv^* - h$ is subgradient of g at point λ

$\max_{\lambda \geq 0} g(\lambda)$ for expression of subgradient

Update rule for λ:
$\lambda^{(k+1)} = [\lambda^{(k)} + \alpha_k s^{(k)}]^+$

Interpretation of subgradient $s = Fu^* + Gv^* - h$

Original problem $Fu + Gv - h \leq 0$

λ, pricing variable associated w/ constraint

Penalize objective w/ $\lambda^T(Fu + Gv - h)$

$(\exists u^8, v^*) Fu^* + Gv^* \leq h \implies \lambda$ decreases
$(\exists u^8, v^*) Fu^* + Gv^* \geq h \implies \lambda$ increases

Summary for decoupling LP:

$$\begin{aligned}&\min_{u,v} c^t u + d^T v\\ &s.t. Au \leq b\\ &\quad Pv \leq q\\ &\quad Fu + Gv \leq h\end{aligned}$$

Subproblems:

$$\begin{aligned}&\min_u c^t u + \lambda^T Fu\\ &s.t. Au \leq b\end{aligned}$$

$$\begin{aligned}&\min_v d^T v + \lambda^T Gv\\ &s.t. Pv \leq q\end{aligned}$$

Algorithms:

Algorithm 11: Decoupling LP with Subgradient

```
set $\lambda^{(0)} > 0$
do
    solve 2 LP problems for fixed $\lambda^{(k)} \to (u^*, v^*)$
    calculate $s^(k) = Fu^* + Gv^* - h$
    update $\lambda^(k+1) = [\lambda^{(k)} + \alpha_k s^{(k)}]^+$
while *not meeting* $\|f(u,v) - g(\lambda)\| < \epsilon$;
```

Choosing α_k, derivation ultimately comes to:
$\sum_{k=1}^{\infty} a_k^2 < \infty, \sum_{k=1}^{\infty} a_k \to \infty \implies$
$\lambda^{(k)} \to \lambda^*, k \to \infty$
eg: $\alpha_k = \frac{1}{k}$

Issues with subgradient method:

- can be slow
- step size selection difficult

0.16 Practical Optimization Strategies

Ellipsoid Method:

- slower
- easier to code

Interior Point Method:

- faster
- harder to code

Tradeoff:
program vs. run time leverage existing software: use problem conversion (eg: LP to SDP)
LP $\iff$ SDP:

$$\begin{aligned} &\min_x c^T x \\ &s.t. Ax \le b \end{aligned}$$

$$\begin{aligned} &\min_x c^T x \\ s.t.diag(Ax-b) \preceq 0 \iff& \\ &\sum_i x_i F_i \preceq 0 \end{aligned}$$

SOCP $\iff$ SDP:

$$\begin{aligned} &\min_x f^T x \\ s.t. \|Ax+b\|_2 \le& c^T x + d \end{aligned}$$

$$\begin{aligned} &\min_x f^T x \\ s.t. &\begin{bmatrix} (c^T x + d)I & Ax+b \\ (Ax+b)^T & c^T x + d \end{bmatrix} \succeq 0 \end{aligned}$$

use Schur Complement:

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

$$X \succeq 0 \iff A \succeq 0 \wedge C - B^T A^{-1} C \succeq 0$$

$$X \succeq 0 \iff C \succeq 0 \wedge A - BC^{-1} B^T \succeq 0$$

$$let\ X = \begin{bmatrix} (c^T x + d)I & Ax+b \\ (Ax+b)^T & c^T x + d \end{bmatrix} \succeq 0 \iff$$

$$\begin{cases} c^T x + d \succeq 0 \\ c^T x + d - \frac{(Ax+b)^T I (Ax+b)}{c^T x + d} \succeq 0 \end{cases}$$

$$(c^T x + d)^2 \ge \|Ax+b\|_2^2$$

$$c^T x + d \ge \|Ax+b\|_2$$

0.16.1 Coordinate Desecent

Unconstrained minimization problem

$$\begin{aligned} &\min_{x_1,x_2} f(x_1,x_2) \\ &assume\ x_1^*, x_2^*\ exist \end{aligned}$$

Algorithm 12: Coordnate Descent

```
do
    fix $x_1$, min. over $x_2$
    fix $x_2$, min. over $x_1$
while not until convergence;
```

Convergence: yes
Sequence of objective value is non-increasing
Optimal value exists and bounded below
CD algo. guaranteed to converge to a (local) minima whether or not objective function is convex

$$\text{global optimal: } \begin{cases} \text{No,} & \text{nonconvex obj} \\ \text{Yes,} & \text{convex, differentiable obj} \end{cases}$$

Assume CD converges to (x_1^*, x_2^*), ∇f exists.
Piecing together:

$$\nabla_{x_1} f|_{x_1^*,x_2^*} = 0$$

$$\nabla_{x_2} f|_{x_1^*,x_2^*} = 0$$

$$\nabla f = \begin{bmatrix} \nabla_{x_1} f \\ \nabla_{x_2} f \end{bmatrix}\Bigg|_{x_1^*,x_2^*} = 0 \implies optimal$$

Non-differentiable objective function, iterate may get stuck.

Constrained problem

CD works if constraints are separable
eg:

$$\begin{aligned} &\min_{x_1,x_2} f(x_1,x_2) \\ s.t. &\begin{bmatrix} A & 0 \\ 0 & F \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \le \begin{bmatrix} b \\ g \end{bmatrix} \end{aligned}$$

x_1, x_2 decoupled, separable
Solve separate problems, concatenate results

At optimality:

$$
\begin{aligned}
&\nabla_{x_1} f(x_1^*, x_2^*) + A^T \lambda_1 = 0 \\
&\nabla_{x_2} f(x_1^*, x_2^*) + F^T \lambda_2 = 0 \\
&\implies \\
&\nabla_x f(x_1^*, x_2^*) + \begin{bmatrix} A^t & 0 \\ 0 & F^T \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \\
&= \nabla_x f(x_1^*, x_2^*) + \tilde{A}^T \lambda = 0
\end{aligned}
$$

CD works for separable constraints due to:

- convergence guarantee
- at convergence, KKT condition of orig. problem satisfied

0.16.2 Sequential Quadratic Programming (SQP)

$\min_x f_0(x)$, f_0 nonconvex, differentiable

Idea: per iteration, approximate $f_0(x)$ by a convex function (eg: quadratic)
Solves for a local optimum

At point x:

$$
\begin{aligned}
f_0(x+v) \approx f_0(x) + \nabla f_0(x)^T v + \frac{1}{2} v^T \nabla^2 f_0(x) v \\
\min_v \nabla f_0(x)^T v + \frac{1}{2} v^T \nabla f_0^2(x) v \\
\Delta x = v^* \implies x = x + \alpha \Delta x, \alpha\ via\ line\ search
\end{aligned}
$$

Constrained problem:
Idea: form Lagrandian $L(x, \lambda, v) = f_0(x) + \sum_i \lambda_i f_i(x) + \sum_i v_i h_i(x)$

$$
\begin{aligned}
\min_x L(x, \lambda, v) \\
s.t. f_i(x) \leq 0 \\
h_i(x) = 0
\end{aligned}
$$

At λ, v optimal dual dual variable

$$
\begin{aligned}
\min_v (\nabla_x L)^T v + \frac{1}{2} v^T (\nabla_x^2 L) v \\
s.t. \nabla f_i(x)^T v + f_i(x) \leq 0 \\
\nabla h_i(x)^T v + h_i(x) \leq 0
\end{aligned}
$$

Updating λ, v?
Select $\Delta\lambda, \Delta v$ in direction of subgradient

For fixed λ, v: $x^{(k+1)} \leftarrow x + \alpha \Delta x, \Delta x = v$

Algo:

Algorithm 13: Sequential Quadratic Programming

```
do
    solve QP to obtain $(\Delta x, \Delta\lambda, \Delta v)$
    line search
      $(x, \lambda, v) = (x, \lambda, v) + \alpha(\Delta x, \Delta\lambda, \Delta v)$
    reapproximate by QP
while not until convergence;
```

0.17 Augmented Lagrangian

Given:

$$\min_x f(x)$$
$$s.t.\ Ax = b$$

Add quadratic term and form the Lagrangian:

$$\begin{aligned}\max_y \min_x L(x,y) =& \max_y \min_x f(x) + y^T(Ax-b)\\ &+ \frac{\rho}{2}\|Ax-b\|_2^2\end{aligned}$$

Optimality conditions for primal and dual:

$$Ax^* - b = 0$$
$$\partial f(x^*) + A^T y^* = 0$$

step size parameter for dual variable:

$$\nabla_x L(x^{k+1}, y^k) = 0\ [by\ definition]$$
$$\partial f(x^{k+1}) + A^T y^k + \rho A^T(Ax^{k+1} - b) = 0$$
$$\partial f(x^{k+1}) + A^T(y^k + \rho(Ax^{k+1} - b)) = 0$$
$$y^{k+1} = y^k + \rho(Ax^{k+1} - b)$$
$$\partial f(x^{k+1}) + A^T y^{k+1} = 0$$

Thus, the step size ρ for y dual variable works to make (x^{k+1}, y^{k+1}) dual feasible.

Updates:

$$\begin{aligned}x^{k+1} =& \operatorname*{argmin}_x f(x) + {y^k}^T(Ax-b)\\ &+ \frac{\rho}{2}\|Ax-b\|_2^2\\ y^{k+1} =& y^k + \rho(Ax^{k+1} - b)\end{aligned}$$

0.18 ADMM

Form augmented Lagrangian, minimize wrt. primal variables, maximize dual problem, use coordinate descent for update. Example, for constrained minimization:

$$\begin{aligned}&\min_x g(x) + g(x)\\ &s.t. Ax + Bx = c\\ &L_p = f(x) + g(z) + y^T(Ax + Bz - c)\\ &+ \frac{\rho}{2}||Ax + Bz - c||_2^2\end{aligned}$$

If multiple rounds of update of primal variable are done before dual variable update, then it is equivalent to method of multipliers where joint optimization is done:

$$x^{k+1}, z^{k+1} = \operatorname*{argmin}_{x,z} L_p(x, z, y^k)\ [descent]$$
$$y^{k+1} = \operatorname*{argmax}_y L_p(x^{k+1}, z^{k+1}, y)\ [ascent]$$

ADMM update:

$$\begin{aligned}x^{k+1} =& \operatorname*{argmin}_x L_p(x, z^k, y^k)\\ z^{k+1} =& \operatorname*{argmin}_z L_p(x^{k+1}, z^k, y^k)\\ y^{k+1} =& \operatorname*{argmax}_y L_p(x^{k+1}, z^{k+1}, y)\\ =& y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)\end{aligned}$$

alternative formulation with scaled form of ADMM, with the same constrained optimization above:

$$\begin{aligned}L_p =& f(x) + g(z) + y^T(Ax + Bz - c)\\ &+ \frac{\rho}{2}||Ax + Bz - c||_2^2\\ r =& Ax + Bz - c\\ u =& \frac{y}{\rho}\\ L_p =& f(x) + g(z) + y^T r + \frac{\rho}{2}||r||_2^2\end{aligned}$$

$$\begin{aligned}y^T r + \frac{\rho}{2}||r||_2^2 =& \frac{\rho}{2}(\frac{2}{\rho}y^T r + r^T r)\\ =& \frac{\rho}{2}(r^T r + \frac{2y^T r}{\rho} + \frac{y^T y}{\rho^2}) - \frac{y^T y}{2\rho}\\ =& \frac{\rho}{2}||r + \frac{y}{\rho}||_2^2 - \frac{1}{2\rho}||y||_2^2\\ =& \frac{\rho}{2}||r + u||_2^2 - \frac{\rho}{2}||u||_2^2\\ =& \frac{\rho}{2}(||r + u||_2^2 - ||u||_2^2)\end{aligned}$$

$$\begin{aligned}
L_p =& f(x) + g(z) + \frac{\rho}{2}(||r+u||_2^2 - ||u||_2^2)\\
x^{k+1} =& \operatorname*{argmin}_x L_p = \operatorname*{argmin}_x f(x)\\
&+ \frac{\rho}{2}(||r+u||)_2^2 - ||u||_2^2)\\
\frac{\partial u}{\partial x} =& 0 \implies\\
x^{k+1} =& \operatorname*{argmin}_x f(x) + \frac{\rho}{2}||r+u||_2^2\\
=& \operatorname*{argmin}_x f(x) + \frac{\rho}{2}||Ax + Bz^k - c + u^k||_2^2\\
z^{k+1} =& \operatorname*{argmin}_z L_p = \operatorname*{argmin}_z g(z)\\
&+ \frac{\rho}{2}(||r+u||_2^2 - ||u||_2^2)\\
\frac{\partial u}{\partial z} =& 0 \implies\\
z^{k+1} =& \operatorname*{argmin}_z g(z) + \frac{\rho}{2}||r+u||_2^2\\
=& \operatorname*{argmin}_z g(z) + \frac{\rho}{2}||Ax^{k+1} + Bz - c + u^k||_2^2\\
y^{k+1} =& y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)\\
\rho u^{k+1} =& \rho u^k + \rho(Ax^{k+1} + Bz^{k+1} - c)\\
u^{k+1} =& u^k + Ax^{k+1} + Bz^{k+1} - c
\end{aligned}$$

thus, ADMM scaled form update becomes:

$$\begin{aligned}
x^{k+1} =& \operatorname*{argmin}_x f(x) + \frac{\rho}{2}||Ax + Bz^k - c + u^k||_2^2\\
z^{k+1} =& \operatorname*{argmin}_z g(z) + \frac{\rho}{2}||Ax^{k+1} + Bz - c + u^k||_2^2\\
u^{k+1} =& u^k + Ax^{k+1} + Bz^{k+1} - c
\end{aligned}$$

residual at iteration k:

$$r^k = Ax^k + Bz^k - c$$

sum of residuals:

$$\begin{aligned}
u^k &= u^0 + \sum_{j=1}^{k} r^j\\
u^{k+1} &= u^k + r^{k+1}
\end{aligned}$$

Example for constrained convex set minimization:

$$\begin{aligned}
&\min_x f(x)\\
&s.t.\ x \in C
\end{aligned}$$

augmented Lagrangian and scaled form ADMM

$$\begin{aligned}
&\min_x f(x) + g(z)\\
&g(z) = I_C(z)\\
&s.t.\ x - z = 0
\end{aligned}$$

$$\begin{aligned}
r =& x - z\\
L_p(x,z,y) =& f(x) + g(z) + y^T(x-z) + \frac{\rho}{2}\|x-z\|_2^2\\
=& f(x) + g(z) + \frac{\rho}{2}(r^T r + \frac{2y^T r}{\rho} + \frac{y^T y}{\rho^2})\\
&- \frac{y^T y}{2\rho}\\
=& f(x) + g(z) + \frac{\rho}{2}\|r + \frac{y}{\rho}\|_2^2 - \frac{\|y\|_2^2}{2\rho}\\
u =& \frac{y}{\rho}\\
L_p(x,z,y) =& f(x) + g(z) + \frac{\rho}{2}(\|r+u\|_2^2 - \|u\|_2^2)\\
x^{k+1} =& \operatorname*{argmin}_x f(x) + \frac{\rho}{2}\|r^k + u^k\|_2^2\\
x^{k+1} =& \operatorname*{argmin}_x f(x) + \frac{\rho}{2}\|x - z^k + u^k\|_2^2\\
z^{k+1} =& \operatorname*{argmin}_z g(z) + \frac{\rho}{2}\|x^{k+1} - z + u^k\|_2^2\\
z^{k+1} =& \operatorname*{argmin}_z I_C(z) + \frac{\rho}{2}\|x^{k+1} - z + u^k\|_2^2\\
z^{k+1} =& \Pi_C(x^{k+1} + u^k)\ [projection\ to\ C]\\
\rho u^{k+1} =& \rho u^k + \rho(x^{k+1} - z^{k+1})\\
u^{k+1} =& u^k + x^{k+1} - z^{k+1}
\end{aligned}$$

0.18.1 Convergence

1. f and g convex, closed, proper
2. unaugmented Lagrandgian has a saddle point

$\implies$

1. residual convergence as $k \to \infty$
2. objective convergence as$k \to \infty$
3. dual variable convergence as$k \to \infty$

0.18.2 Optimality and Stopping Conditions

use primal and dual residual as proxy for stopping

$$\begin{aligned}
&f(x^k) + g(z^k) - p^* \leq -{y^k}^T r^k + (x^k - x^*)^T s^k \\
&\|x^k - x^*\|_2 \leq d \\
&f(x^k) + g(z^k) - p^* \leq -{y^k}^T r^k + d\|s^k\|_2 \\
&f(x^k) + g(z^k) - p^* \leq \|y^k\|_2 \|r^k\|_2 + d\|s^k\|_2
\end{aligned}$$

make residuals small:

$$\begin{aligned}
&\|f^k\|_2 \leq \epsilon^{pri} \\
&\|s^k\|_2 \leq \epsilon^{dual}
\end{aligned}$$

Eg:

$$\begin{aligned}
\epsilon^{pri} &= \sqrt{p}\epsilon^{abs} + \epsilon^{rel} max(\|Ax^k\|_2, \|Bz^k\|_2, \|c\|_2) \\
\epsilon^{dual} &= \sqrt{n}\epsilon^{abs} + \epsilon^{rel}\|A^T y^k\|_2
\end{aligned}$$

0.18.3 L1 problems

least absolute deviation

$$\min_x \|Ax - b\|_1$$

$$\begin{aligned}
&z = Ax - b \\
&\min_{x,z} f(x) + g(z) \\
&s.t.\ Ax - b - z = 0 \\
&g(z) = \|z\|_1 \\
&f(x) = 0 \\
&L(x, z, y) = \|z\|_1 + y^T(Ax - b - z) \\
&\ + \|Ax - b - z\|_2^2 \\
&x^{k+1} = \operatorname*{argmin}_x L(x, z^k, y^k) \\
&\frac{\partial L}{\partial x} = A^T y + \rho A^T(Ax - b - z) = 0 \\
&x^{k+1} = (A^T A)^{-1} A^T(b + z^k - u^k) \\
&z^{k+1} = \operatorname*{argmin}_z L(x^{k+1}, z, y^k) \\
&0 \in \partial g(z) - y^k - \rho(Ax^{k+1} - b - z) \\
&0 \in \frac{1}{\rho}\partial g(z) - \frac{y^k}{\rho} - (Ax^{k+1} - b - z) \\
&0 \in \frac{1}{\rho}\partial g(z) - (Ax + u - b - z) \\
&z^{k+1} = \operatorname*{argmin}_z g(z) + \frac{\rho}{2}\|Ax^{k+1} + u^k - b - z\|_2^2 \\
&z^{k+1} = prox_{||*||_1, \frac{1}{\rho}}(Ax^{k+1} + u^k - b) \\
&y^{k+1} = y^k + Ax^{k+1} - b - z^{k+1}
\end{aligned}$$

Huber fitting

$$\begin{aligned}
&\min f_{huber}(Ax - b) \\
&f_{huber}(v) = \begin{cases} \frac{v^2}{2} & , v \leq 1 \\ \|v\|_1 - \frac{1}{2} & , v \geq 1 \end{cases}
\end{aligned}$$

use proximal operator for Huber function instead of L1 when updating z:

$$\begin{aligned}
z^{k+1} =& \frac{\rho}{1+\rho}(Ax^{k+1} - b + u^k) \\
&+ \frac{1}{1+\rho} ST_{1+\frac{1}{\rho}}(Ax^{k+1} - b + u^k) \\
ST :=& \text{ SoftThresholding}
\end{aligned}$$

L1 minimization with equality constraint (basis pursuit)

$$\begin{aligned}
&\min_x \|x\|_1\\
&s.t.\ Ax = b
\end{aligned}$$

ADMM:

$$\begin{aligned}
&\min_{x,z} I_{Ax-b=0}(x) + \|z\|_1\\
&s.t.\ x - z = 0\\
&L(x,z,y) = I_{Ax-b=0}(x) + \|z\|_1 + y^T(x-z)\\
&+ \frac{\rho}{2}\|x-z\|_2^2\\
&x^{k+1} = \Pi_{Ax-b=0}(z^k - u^k)\\
&z^{k+1} = \text{SoftThresholding}_{\frac{1}{\rho}}(x^{k+1} + u^k)\\
&u^{k+1} = u^k + x^{k+1} - z^{k+1}
\end{aligned}$$

general L1 regularized loss

$$\min_x l(x) + \lambda\|x\|_1,\ l\ convex$$

Lagrangian and ADMM:

$$\begin{aligned}
&g(x) = \lambda\|x\|_1\\
&min_{x,z} l(x) + g(z)\\
&s.t.\ x - z = 0\\
&L(x,z,y) = l(x) + g(z) + y^T(x-z) + \frac{\rho}{2}\|x-z\|_2^2\\
&x^{k+1} = \operatorname*{argmin}_x L(x, z^k, y^k)\\
&0 \in \partial l(x) + y + \rho(x-z)\\
&u = \frac{y}{\rho}\\
&x^{k+1} = \operatorname*{argmin}_x l(x) + \frac{\rho}{2}\|x - z^k - u^k\|_2^2\\
&z^{k+1} = \operatorname*{argmin}_z L(x^{k+1}, z, y^k)\\
&0 \in= \partial g(z) - y^T + \rho(x^{k+1} - z)(-1)\\
&z^{k+1} = \operatorname*{argmin}_z \lambda\|z\|_1 + \frac{\rho}{2}\|x^{k+1} - z + u\|_2^2\\
&z^{k+1} = \operatorname*{argmin}_z \|z\|_1 + \frac{\rho}{2\lambda}\|x^{k+1} + u - z\|_2^2\\
&z^{k+1} = prox_{\|*\|_1, \frac{\lambda}{\rho}}(x^{k+1} + u^k)\\
&z^{k+1} = \text{SoftThresholding}_{\frac{\lambda}{\rho}}(x^{k+1} + u^k)\\
&u^{k+1} = u^k + x^{k+1} - z^{k+1}
\end{aligned}$$

Thus, this problem is reduced to solving a series of L2 regularized loss.

If $l(x)$ is smooth, various 1st and 2nd order methods can be used: L-BFGS, Newton, Quasi-Newton, Conjugate Gradient.

L1 regularized linear regression (Lasso)

$$\begin{aligned}
&\min_x f(x) + g(x)\\
&f(x) = \frac{1}{2}\|Ax - b\|_2^2\\
&g(x) = \lambda\|x\|_1
\end{aligned}$$

ADMM:

$$\begin{aligned}
&\min_x f(x) + g(z)\\
&s.t.\ x - z = 0\\
&L(x, z, y) = f(x) + g(z) + y^T(x - z) + \frac{\rho}{2}\|x - z\|_2^2\\
&x^{k+1} = \operatorname*{argmin}_x L(x, z^k, y^k)\\
&A^T(Ax - b) + y^k + \rho(x - z^k) = 0\\
&x = (A^TA + \rho I)^{-1}(A^Tb + \rho(z^k - u^k))\\
&z^{k+1} = \operatorname*{argmin}_z L(x^{k+1}, z, y^k)\\
&0 \in \partial g(z) + \rho(x - z + u)(-1)\\
&z^{k+1} = \operatorname*{argmin}_z \lambda\|z\|_1 + \frac{\rho}{2}\|x^{k+1} + u^k - z\|_2^2\\
&z^{k+1} = prox_{\|*\|_1, \lambda/\rho}(x^{k+1} + u^k)\\
&z^{k+1} = \text{SoftThresholding}_{\lambda/\rho}(x^{k+1} + u^k)\\
&u^{k+1} = u^k + x^{k+1} - z^{k+1}
\end{aligned}$$

generalized lasso

$$\begin{aligned}
&\min_x f(x) + g(x)\\
&f(x) = \frac{1}{2}\|Ax - b\|_2^2\\
&g(x) = \lambda\|Fx\|_1
\end{aligned}$$

ADMM:

$$\begin{aligned}
&\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|Fx\|_1\\
&z = Fx\\
&\min_{x,z} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|z\|_1\\
&s.t.\ Fx - z = 0\\
&L(x, z, y) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|z\|_1 + y^T(Fx - z)\\
&+ \frac{\rho}{2}\|Fx - z\|_2^2\\
&\frac{\partial L}{\partial x} = A^T(Ax - b) + F^Ty + \rho F^T(Fx - z) = 0\\
&A^TAx + \rho F^TFx = A^Tb - F^Ty + \rho F^Tz\\
&x^{k+1} = (A^TA + \rho F^TF)^{-1}(A^Tb + \rho F^T(z^k - u^k))\\
&\frac{\partial L}{\partial z} = \partial(\lambda\|z\|_1) + \rho(Fz - z + u)(-1)\\
&z^{k+1} = \operatorname*{argmin}_z \lambda\|z\|_1 + \frac{\rho}{2}\|Fx - z + u\|_2^2\\
&z^{k+1} = prox_{\|*\|_1, \lambda/\rho}(Fx^{k+1} + u^k)\\
&u^{k+1} = u^k + Fx^{k+1} - z^{k+1}
\end{aligned}$$

Special case for total variation denoising:

$$\begin{aligned}
&A = I\\
&F = \text{1st order difference matrix}\\
&F_{ij} = \begin{cases} 1 & , i + 1 = j\\ -1 & , i = j\\ 0 & , o/w \end{cases}
\end{aligned}$$

0.18.4 consensus

$$\min_x \sum_i f_i(x)$$

ADMM:

$\min\limits_x \sum\limits_i f_i(x_i)$

$s.t.\ x_i - z = 0$

$L(x,z,y) = \sum\limits_i f_i(x_i) + y_i^T(x_i - z) + \frac{\rho}{2}\|x_i - z\|_2^2$

$x_i^{k+1} = \underset{x_i}{\operatorname{argmin}} f_i(x_i) + {y_i^k}^T(x_i - z^k) + \frac{\rho}{2}\|x_i - z^k\|_2^2$

$\frac{\partial L}{\partial z} = \sum\limits_i -y_i + \rho(x_i - z)(-1) = 0$

$z^{k+1} = \sum\limits_i \frac{y_i}{\rho} + x_i^{k+1}$

$y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1})$

simplification and averaging:

$z^{k+1} = \frac{\bar{y_i}}{\rho} + \bar{x}^{k+1}$

$\bar{y}^{k+1} = \bar{y}^k + \rho(\bar{x}^{k+1} - z^{k+1})$

$substitution \implies$

$\bar{y}^{k+1} = 0$

$z^{k+1} = \bar{x}^{k+1}$

$update\ equations:$

$x_i^{k+1} = \underset{x_i}{\operatorname{argmin}} f_i(x_i) + {y_i^k}^T(x_i - \bar{x}^k) + \frac{\rho}{2}\|x_i - \bar{x}^k\|_2^2$

$y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - \bar{x}^{k+1})$

can also add another regularization term to objective, in which case the z update is retained

0.18.5 computation shortcuts

caching factorization

matrix inversion lemma

block separability

component separability

warm start

0.18.6 distributed techniques

splitting across samples

TODO

splitting across features

TODO

Reference: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers [Boyde et al.]

0.19 Subgradient Methods

todo

0.20 Appendix

0.20.1 Gradient of Log Det

$f(x) = log(detX)$

$$\begin{aligned}
f(X+\delta X) &= logdet(X+\delta X) \\
&= logdet((X^{\frac{1}{2}}(I+X^{-\frac{1}{2}})\delta X X^{-\frac{1}{2}})X^{\frac{1}{2}}) \\
&= logdet(X) + logdet(I+X^{-\frac{1}{2}}\delta X X^{-\frac{1}{2}}) \\
&\text{let } M = X^{-\frac{1}{2}}\delta X X^{-\frac{1}{2}} \\
&= logdet(X) + logdet(I+M)
\end{aligned}$$

claim eigenvalues of $I+M$: $1+\lambda_i$

$$\begin{aligned}
Mv_i &= \lambda_i v_i \\
(I+M)v_i &= (1+\lambda_i)v_i \\
det(M) &= \prod_i (1+\lambda_i) \\
f(X+\delta X) &= logdet(X) + log\prod_i (1+\lambda_i) \\
&= logdet(X) + \sum_i log(1+\lambda_i) \\
&\approx logdet(X) + \sum_i \lambda_i \text{ since } \delta X \text{ is small} \\
&\approx logdet(X) + trace(X^{-\frac{1}{2}}\delta X X^{\frac{1}{2}}) \\
&\approx logdet(X) + trace(X^{-1}\delta X) \\
trace(X^{-1}\delta X) &= (X^{-T})^T \delta X \\
f(X+\delta X) &= f(X) + (\nabla f(X))^T \delta X \implies \nabla f(X) = X^{-1} \\
logdet(X) &= log(X) \implies f'(X) = \frac{1}{X}
\end{aligned}$$

0.20.2 2nd order approximation of Log Det

$f(X+\delta X) = f(X) + < \nabla f(X), \delta X > + 1/2 < \delta X, \nabla^2 f(x) \delta X >$

first look at first order approximation: $g(X) = X^{-1}$

$$\begin{aligned}
g(X+\delta X) &= (X+\delta X)^{-1} = (X^{\frac{1}{2}}(I+X^{-\frac{1}{2}}\delta X X^{-\frac{1}{2}})X^{\frac{1}{2}})^{-1} \\
&= X^{-\frac{1}{2}}(I+X^{-\frac{1}{2}}\delta X X^{-\frac{1}{2}})^{-1}X^{-\frac{1}{2}} \\
&\text{for small A(small eigenvalues): } (I+A)^{-1} \approx I - A \\
&= X^{-\frac{1}{2}}(I-X^{-\frac{1}{2}}\delta X X^{-\frac{1}{2}})X^{-\frac{1}{2}} \\
&= X^{-1} - X^{-1}\delta X X^{-1}
\end{aligned}$$

$$logdet(X+\delta X) = logdet(X) + tr(X^{-1}\delta X) - \frac{1}{2}tr(\delta X X^{-1}\delta X X^{-1})$$

0.21 Miscellaneous Properties

$(a+x)^{-1} \approx 1-x$
$\lim_{t\to 0} \frac{f(x+\epsilon t)-f(x)}{t} = \frac{\partial f(x)}{\partial x}\epsilon$

0.21.1 Pseudo-inverse

Overconstrained case:
Cast as L2 norm approximation problem

$$\min_x \|Ax-b\|_2^2$$

$$\begin{aligned}
&(Ax-b)^T(Ax-b) = x^TA^TAx - 2x^TA^Tb + b^Tb\\
&\frac{\partial}{\partial x}(x^TA^TAx - 2x^TA^Tb + b^Tb) = 2A^TAx - 2A^Tb\\
&x = (A^TA)^{-1}A^Tb
\end{aligned}$$

Underconstrained case:
Cast as a least-norm problem w/ equality constraint

$$\begin{aligned}
&\min_x \|x\|_2^2\\
&s.t.:\ \ Ax = b
\end{aligned}$$

$$\begin{aligned}
&\min_x L(x,\lambda,v) = x^Tx - v^T(Ax-b)\\
&\frac{\partial(x^Tx - v^T(Ax-b))}{\partial x} = 2x - A^Tv = 0\\
&x = -\frac{1}{2}A^Tv\\
&g(\lambda,v) = [x^Tx - v^T(Ax-b)]_{x=-\frac{1}{2}A^Tv}\\
&g(\lambda,v) = -\frac{1}{4}v^TAA^Tv - v^Tb\\
&dual:\ \max_{\lambda,v} g(\lambda,v) = -\min_{\lambda,v} g(\lambda,v)\\
&\frac{\partial(\frac{1}{4}v^TAA^Tv + v^Tb)}{\partial v} = 0\\
&v = \frac{1}{2}AA^Tv + b = 0\\
&v = -2(AA)^{-1}b\\
&x = -\frac{1}{2}A^Tv|_{v=-2(AA)^{-1}b} = A^T(AA^T)^{-1}b
\end{aligned}$$

0.22 Problems

0.22.1 min. vol. covering ball

Find minimal volume norm ball B covering $B_j, \forall j$, where B_j is a norm ball. Consider 2-norm:

Let c_j, r_j be center, radius of norm ball B_j
Let c, r be center, radius of B

$$\begin{aligned}
&\min_{c,r}\ r \\
&s.t. \| |c - c_j| + r_j \|_2 \leq r, \forall j \\
&c^c + (-c_j + r_j)^T(-c_j + r_j) + 2c^T(-c_j + r_j) \leq r, \forall j \\
&c^c + (c_j + r_j)^T(c_j + r_j) + 2c^T(c_j + r_j) \leq r, \forall j
\end{aligned}$$

QCQP problem

Find minimal volume norm ball B covering $B_j, \forall j$, where B_j is a norm ball. Consider ∞-norm:

$$\begin{aligned}
&\min_{c,r}\ r \\
&s.t. \| |c - c_j| + r_j \|_\infty \leq r, \forall j \\
&c - c_j + r_j \leq r, \forall j \\
&-c + c_j + r_j \leq r, \forall j
\end{aligned}$$

LP problem

0.22.2 polyhedron intersection

$D = \{x \in \mathbb{R}^n : Cx \leq d\}$
$G = \{x \in \mathbb{R}^n : Cx \leq d\}$
Solve $D \cap G = \emptyset$? using LP:

$$\begin{aligned}
&\min_{a,b,x}\ 0 \\
&s.t. a^T x = b \\
&Cx - d \nleq 0 \\
&Hx - g \nleq 0
\end{aligned}$$

$$\begin{aligned}
&\min_{a,b,x,e}\ \sum_i e_i \\
&s.t. a^T x = b \\
&Cx - d + e \leq 0 \\
&Hx - g + e \leq 0 \\
&e \leq 0
\end{aligned}$$

f^* feasible and $f^* < 0 \implies D \cap G = \emptyset$
otherwise, $\implies D \cap G \neq \emptyset$

Solve $D \subseteq G$? using LP:

Algorithm 14: Descent Overview

```
$verts$ of D ← solve intersection of halfspace equations of D
for $\forall vert \in D$ do
    $min\ 0$
    s.t. $Hv \leq g$
    $f^* = \begin{cases} NaN \implies returnD \nsubseteq G \\ o/w \implies continue \end{cases}$
return $D \subseteq G$
```