Convex sets and functions

- Convex sets : hyperplane, half-space, convex cone, $\|\cdot\|$-ball, ellipsoid, polyhedron,
- Intersection, affine, perspective, linear-fractional functions preserve set convexity.
- f is α-strongly convex iff $f-\alpha\|\cdot\|^2$ is convex, ie. if $\forall(x,y) \in \mathrm{dom}(f), \forall\lambda \in [0,1], f(\lambda x+(1-\lambda)y) < \lambda f(x)+(1-\lambda)f(y)-\alpha\lambda(1-\lambda)\|x-y\|^2$. $\alpha = 0$ is usual convexity,
- *Warning :* set $\mathrm{dom}(f)$ has to be convex for f to be convex !
- Convexity is *strict* if $\forall\lambda \in (0,1), f(\lambda x+(1-\lambda)y) < \lambda f(x)+(1-\lambda)f(y)$,
- Ex : *affine* functions $f(x) = ax+b$, $\exp(ax)$, $-\log$, *log-sum-exp* $\log\left(\sum_{i=1}^n \exp g_i(x)\right)$, all norms, x^p $(p > 0, x > 0)$, $|x|^p$ $(p > 1)$, negative entropy $x\log x$ are all convex,
- Non-negative weighted sum, comp. with affine f., point-wise max or sup, composition, minimization, perspective all *preserve convexity*,
- On matrices : $f(X) = \mathrm{Tr}(A^T X)+b$ (affine), spectral norm $\|X\|_2 = \sigma_{\max}(X) = \left(\lambda_{\max}(X^T X)\right)^{1/2}$ (biggest spectral value) are convex,
- Condition for cxty : $f(y) \geqslant f(x)+\nabla f(x)^T(y-x)$ (SNC, if f diff., *supporting hyperplane*), $\nabla^2 f(x) = Hf(x) \succeq 0$ (SNC, if f twice diff.), $Hf(x) \succ 0 \Leftrightarrow$ str.cxty,
- 3 conditions. Convex : all local minima are global. Strictly cvx. : if local minimum, then unique and global. Strongly cvx : $\exists!\, x^\star$ local (and global) minimum.

Optimization problems, convex problems

- Generic form : $\min_{x\in H} F(x)$ for $F : \mathbb{R}^n \to \mathbb{R}$ on a domain $H \subset \mathbb{R}^n$,
- Solution $x^\star \in \mathrm{argmin}_{x\in H} F(x), \in \mathbb{R}$ can be not unique, and maybe no solution. *Infeasible* if no x satisfies the constraints $(x^\star = +\infty)$. Unbounded below if $x^\star = -\infty$,
- *Equality* constraints : $H_j(x) = 0$ $(j = 1\dots p)$, and *inequality* constraints : $F_i(x) \leqslant 0$ $(i = 1\dots m)$ ($\leqslant$ can be $\preceq_{\mathcal{K}}$ a generalized inequality on a cone),
- Problem is *convex* if F_0, H_j, F_i are **all** convex (it is then *tractable*). Usual form : $H_j(x) = a_j^T x + b_j$ (affine equality constraints).

Optimality conditions (1)

- *Fermat/Euler's condition* : if F is differentiable, $x^\star \in \mathrm{argmin}_x F(x) \implies \nabla F(x^\star) = 0$ (stationary point). If F is convex, it's a $\Leftrightarrow$,
- 2^{nd}*-order condition* : if F is twice differentiable, $x^\star \in \mathrm{argmin}_x F(x) \implies \nabla F(x^\star) = 0, \nabla^2 F(x^\star) \succeq 0$. Converse : $\nabla^2 F(x^\star) \succ 0$ is required for $\Leftarrow$.

Lagrangian and dual problem

- KKT theorem justifies to introduce λ, ν the *Lagrange multipliers* ($\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$,
- *Lagrangian* of a constrained problem : $\mathcal{L}(x,\lambda,\nu) = F_0(x) + \sum_{i=1}^m \lambda_i F_i(x) + \sum_{j=1}^p \nu_j H_j(x)$ $(\lambda \succeq 0)$. Dual function : $g^\star(\lambda,\nu) = \inf\limits_{x\in\mathbb{R}^n} L(x,\lambda,\nu)$ (always convex),
- *Dual problem* is $\max_{\lambda,\nu} g^\star(\lambda,\nu)$ with $\lambda \succeq 0$. Dual solution is $d^\star \in \mathbb{R}$ (optimal dual value),
- How to compute the dual : write $L(x,\lambda,\nu)$, regroup terms with and without x, minimize wrt x to compute $g^\star(\lambda,\nu)$, and (try to) solve the dual problem,

— *Weak duality* : $d^\star \leqslant p^\star$ (always : max min $\leqslant$ min max), *duality gap* is $p^\star - d^\star$ (unknown !),

— *Slater's condition* : $\exists$ one strictly feasible point $\implies$ strong duality : $d^\star = p^\star$,

— *Complementary slackness* :
$\lambda_i^\star \cdot F_i(x^\star) = 0$: $\lambda_i^\star > 0 \implies F_i(x^\star) = 0$ and $F_i(x^\star) < 0 \implies \lambda_i^\star = 0$.

Conjugate functions

— Definition : $f^\star(y) \overset{\text{def}}{=} \sup_{x \in \mathbf{dom} f} \left(y^T x - f(x)\right)$. $f^\star$ is *always* convex,

— Examples : $f(x) = -\log x \implies f^\star(y) = -1 - \log y$ if $y < 0, +\infty$ otherwise. $f(x) = x^T Q x, Q \in \mathbf{S}^n_{++}, \implies f^\star(y) = y^T Q^{-1} y$.

KKT optimality conditions (2)

— *KKT conditions* : (1) Primal feasibility ($F_i(x) \leqslant 0, H_j(x) = 0$) + (2) dual feasibility ($\lambda \succeq 0$) + (3) complementary slackness + (4) $\nabla_x \mathcal{L}(x, \lambda, \mu) = 0$,

— If strong duality : x, λ, μ optimal have to satisfy KKT conditions,

— For a convex problem with no = cstr and $\preceq$ cstr, $x = x^\star \Leftrightarrow \exists \lambda$ s.t. (x, λ) is a *saddle point* (*point selle*) of $\mathcal{L}(x, \lambda)$, ie. $F(x) \preceq 0, p \succeq 0, F(u) \cdot p = 0, \nabla F_0(x) + \sum_{i=1}^m p_i \nabla F_i(x) = 0$,

— Still apply for linear = constraints : $A_i x = b_i \Leftrightarrow A_i x \preceq b_i$ and $A_i x \preceq b_i$,

— Note : for (completely) convex problem, KKT conditions are SNC ($\Leftrightarrow$).

Classical convex problems

— Least-Squares : $\min \|Ax - b\|_2^2$. $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ are parameters, $x \in \mathbb{R}^n$ is variable. Normal equations solution : $x^\star = (A^T A)^{-1} A^T b$.

— Linear Programs : $\min c^T x$ with $a_i^T x \leqslant b_i, i = 1 \dots m$. $c, a_i \in \mathbb{R}^n, b_i \in \mathbb{R}$ are parameters, $x \in \mathbb{R}^n$ is variable. With *affine* equality constraints : $\min c^T x$ with $Gx \preceq h, Ax = b$.

— Quadratic Programs : $\min \|Fx - g\|_2^2$ with $a_i^T x \leqslant b_i, i = 1 \dots m$. $F \in \mathbb{R}^{m \times n}, g \in \mathbb{R}^m, a_i \in \mathbb{R}^n, b_i \in \mathbb{R}$ are parameters, $x \in \mathbb{R}^n$ is variable.

— Symmetric Cone ($\mathcal{K}$) Programs : $\min c^T x$ with $Ax - b \in \mathcal{K}$.

— Quadratically constrained Quadratic Program : QP + QC.

Gradient descent methods

For *unconstrained* problems, $\min_x F(x)$, F differentiable. (order-1 Taylor apprx.)

— Algorithm from x_0, $x_{(n+1)} \leftarrow x_n - \alpha_n d_n$, direction $d_n = \nabla_x F(x)$, step size $\alpha_n > 0$,

— Variants : *fixed* step ($\alpha_n = \alpha > 0$), *optimal* step ($\alpha_n = \text{argmin}_{\alpha \in \mathbb{R}} F(x_n - \alpha \nabla_x F(x))$ optimal, with exact or backtracking line search), *conjugate* step (d_n depends on history of $d_i, i < n$),

— With = constraints : write $H = \{x : F_i(x) \leqslant 0, i = 1 \dots m\}$, $\texttt{proj}_H$ its projection operator, then $x_{(n+1)} \leftarrow \texttt{proj}_H (x_n - \alpha_n d_n)$. Also : Uzawa algorithm (not covered).

Newton's method

For *unconstrained* problems, $\min_x F(x)$, F twice differentiable. (order-2 Taylor apprx.)

— Algorithm : from x^0 close "enough" to $x^\star$, update $x_{n+1} \leftarrow x_n - \alpha_n (HF(x))^{-1} \nabla F(x)$,

— 2 : slow (*damped*, require backtracking for α_n) then quick (*quadratically*, all $\alpha_n = 1$),

— Newton's decrement : $\lambda(x) = \nabla F(x)^T \nabla^2 F(x)^{-1} \nabla F(x)$, stopping criterion $\lambda^2/2 \leqslant \varepsilon$,

— Concretely : works really well, always with < 150 steps,

— Can be used with $=$ constraints (need a valid starting point and new update formula).

Other methods

— *Sub-gradient* for f at x = vector g s.t. $f(y) \geqslant f(x) + g^T(y - x)$ (supporting hyperplane). Method = descent with direction d_n a sub-gradient (not unique).

— Coordinate descent algorithm : minimize in 1-D successively on *one* coordinate (cycle $x_1, \dots, x_n$), if domain $H = H_1 \times \cdots \times H_k$ product of simpler sets ($n \neq k$ is possible),

— How-to find an *strictly feasible* initial point x_0 ? Phase 1 : $(x, s)^\star = \operatorname{argmin}_{x,s} s$ with $f_i(x) \leqslant 0, Ax = b$. If $s^\star < 0$ then $x_0 = x^\star$ OK, else no x_0.

— Central path / barrier method : start x_0, $t_0 > 0, \mu > 1$, then repeat : 1. Centering : $x_{n+1} = x^\star(t) = \operatorname{argmin}(t_n f_0 + \phi)$ with $Ax = b$, 2. Increase $t_{n+1} \leftarrow \mu t_n$ (stops if $m/t < \varepsilon$).

Stochastic optimization : example

Stochastic Linear Program : $\min c^T x$ with $\mathbb{P}(a_i^T x \leqslant b_i) \geqslant \eta, i = 1 \dots m$ (*required reliability* $0 \leqslant \eta \leqslant 1$). $c \in \mathbb{R}^n, b_i \in \mathbb{R}$ are parameters, a_i follows $\mathcal{N}(\bar{a}_i, \Sigma_i)$, $x \in \mathbb{R}^n$ is variable. SLP is convex $\Leftrightarrow \eta \geqslant 1/2$.

Schur's complements

If $M = [[A; B]; [C; D]]$, its *Schur's complements* are : bottom $M/D \overset{\text{def}}{=} A - BD^{-1}C$, top $M/A \overset{\text{def}}{=} D - CA^{-1}B$. Formula : $M \succ 0 \Leftrightarrow A \succ 0, M/A \succ 0$ (resp. with D).

Examples of gradients

— $\nabla_x \left(a^T x + b\right) = a$ (vector affine function),

— $\nabla_x \left(\frac{1}{2} x^T A x\right) = \frac{1}{2}(A^T + A)x$ (vector), $= A^T x$ if A sym.,

— $\nabla_X \left(\operatorname{Tr}(A^T X) + b\right) = A$ (matrix affine function),

— $\nabla_X (\det(X)) = \bar{X}$ with $\bar{X} = (\det X)(X^{-1})^T$ the *comatrix* of X),

— $\nabla_X (\log(\det X)) = X^{-1}$ (matrix, proof with LU decomposition, and $A + H = A^{1/2}(I + A^{-1/2} H A^{-1/2}) A^{1/2}$),

— $f(X) = X^{-1} \implies (\nabla_X f)(H) = -X^{-1} H X^{-1}$,

Sublevel set (bonus) The α-sublevel set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $C_\alpha(f) \overset{\text{def}}{=} \{x \in \operatorname{dom}(f) : f(x) \leqslant \alpha\}$ (convex for f cvx).