Convex Optimization

1. Mathematical optimization = Mathematical programming : 수학적 계획법

어떠한 기준 아래서 최고의 값을 찾는 것 ⇒ conceptual idea를 mathematical compute로 찾는 것

가. 왜 convex optimization 인가?

local minimum을 생각하지 않아도 되기 때문 : convex optimization에서는 local minima가 global minima

나. Standard form

$$\min_{x \in D} f(x) \text{ subject to } \begin{array}{l} g_i(x) \le 0, i = 1, \cdots, m \\ h_j(x) = 0, j = 1, \cdots, r \end{array}$$

where f and g_i are all convex, and h_j are affine

$f(x)$: 목적함수 objective function (= cos, loss, utility, fitness function)

$g_i(x) \le 0$ and $h_i(x) = 0$: 제약함수 constraint function

제약식에 만족하는 집합 : feasible set

x^* : solution (= minimizer, optimal)

다. 예

① 회귀분석 LSE

C : 오차 제곱 최소화가 제일 좋다 ⇒ M : $\min_\beta \sum (y_i - x_i\beta)^2$

여기에 β가 크기 t보다는 작다는 idea를 추가 : $\min_\beta \sum (y_i - x_i\beta)^2$ subject to $\| \beta \|_2 \leqq t$

⇒ regularization

② PCA $z = \delta^T X$, 정보가 많다는 $Var(z)$가 크다 ⇒ $\max_\delta Var(z)$ subject to $\| \delta \|_2 = 1$

2. convex

가. convex set

$C \subseteq \boldsymbol{R}^n$ such that $x, y \in C \Rightarrow tx + (1-t)y \in C$ for all $0 \le t \le 1$

$tx + (1-t)y \in C$: 선분 즉 선분의 모든 점들이 집합 안에 들어가 있는 것

나. convex function

$f : \boldsymbol{R}^n \to R^1$ such that $dom(f) \subseteq \boldsymbol{R}^n$ convex,

and $f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$ for all $0 \le t \le 1$

and all $x, y \in dom(f)$

1) Concave function

opposite inequality above, so that f concave ⇔ - f convex

convex set가 주로 사용함 : 아닌 경우 concave set 이라기 보다는 non-concave set이라고 부름

2) strongly convex ⇒ strictly convex ⇒ convex

strictly convex : $f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$ for $x \ne y$ and $0 \le t \le 1$

즉 f가 convex이면서 greater curvature than a linear function

stronly convex : with parameter $m > 0$, $f - \frac{m}{2} \| x \|_2^2$ is convex

즉 f가 convex as a quadratic function

3) convex functions의 예 : Unvariate functions

① exponential function e^{ax} for any a

② power function x^a for $a \ge 1$ or $a \le 0$ 또는 $0 \le a \le 1$

③ logarithmic function $\log x$

4) convex functions의 다른 예

① affine function $a^T x + b$ is both convex and cocave

② quadratic function is $\frac{1}{2} x^T Q x + b^T x + c$ is convex, provided that $Q \ge 0$

즉 positive semidefinite라는 조건하에 convex

③ least squres loss $\| y - Ax \|_2^2$ is always convex

$A^T A$가 항상 positive semidefinite이기 때문

④ norm $\| x \|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p \ge 1$ is convex

norm은 $\| \quad \| : A \to R^1$인 함수로 $x \in A$의 크기를 의미

cf. $\| x \|_\infty = \max_{i=1,\cdots,n} |x_i|$, $\| x \|_0 =$ (0이 아닌 개수)

⑤ indicator funciton $I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$ is convex provided that C is convex stats와 다르게 쓰는 이유 $\min f(x)$ s.t $x \in c$를 $\min(f(x)+I_c(x))$로 사용 가능하기 때문

⑥ max function $\max\{x_1, \cdots, x_n\}$ is convex

다. convex combination of $x_1, \cdots, x_k \in R^n$

linear combination $\theta_1 x_1 + \cdots + \theta_k x_k$ with $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$

라. convex hull

$conv(C)$: convex hull of a set C is all convex combinations of elements

즉 convex hull 은 convex set이 아닌 것을 (필요한 부분을 포함하여) convex set으로 바꾸는 것

마. Cone

$C \subseteq R^n$ such that $x \in C \Rightarrow tx \in C$ for all $0 \leq t$

convex cone : cone that is also convex

conic combination : liniear combination $\theta_1 x_1 + \cdots + \theta_k x_k$

conic hull : collects all conic combinations

바. Key properties of convex

① separating hyperplane theorem

two disjoint convex sets have a separaing between hyperplane them

If nonempty convex sets with $C \cap D = \varnothing \Rightarrow \exists a, b$ such that $C \subseteq \{x : a^T x \leq b\}$

$D \subseteq \{x : a^T x \geq b\}$

② supporting hyperplane theorem

a boundary point of a convex set has a supporting hyperplane passing through it

SVM의 이론적 기반

③ operations preserving convexity

intersection of convex sets is convex

convex를 affine에 넣으면 convex가 된다.

사. Key properties of convex function

① A function is convex ⇔ its restriction to any line is convex.

② Epigraph characterization

a function f is convex ⇔ its epigraph $epi(f) = \{f(x,t) \in dom(f) \times R : f(x) \leq t\}$ is convex set

② Convex sublevel sets

f is convex ⇒ its sublevel sets $\{x \in dom(f) : f(x) \leq t\}$ are convex

③ First-order characterization ★★★

if f is differentiable,

f is convex ⇔ $dom(f)$ is convex, and $f(y) \geq f(x) + \nabla f(x)^T(y-x)$ for all $x,y \in dom(f)$

즉 differentiable convex function $\nabla f(x) = 0$ ⇔ x minimizes f

④ Second-order characterization

if f is twice differentiable,

f is convex ⇔ $dom(f)$ is convex, and $\nabla^2 f(x) \geq 0$ for all $x \in dom(f)$

⑤ Jensen's inequality

if f is convex, and X is a random variable supported on $dom(f)$, then $f(E[X]) \leq E[f(X)]$

⑥ Operations preserving convexity

Nonnegative linear combination : $f_1, \cdots, f_m$ convex implies $a_f f_1 + \cdots + a_m f_m$ convex for any $a_i \geq 0$

Pointwise maximization : if f_s is convex for any $s \in S$, then $f(x) = \max_{s \in S} f_s(x)$ is convex

Partial minimization : if $g(x,y)$ is convex in x,y, and C is convex ⇒ $f(x) = \min_{y \in C} g(x,y)$ is convex

즉 하나씩 minimization 할 수 있다는 의미

3. Optimization Basics

① First-order condition for optimality ★★★

For a convex problem $\min_x f(x)$ subject to $x \in C$ and differentiable f,

a feasible point x is optimal ⇔ $\nabla f(x)^T(y-x) \geq 0$ for all $y \in C$

② Partial optimization

If we decompose $x=(x_1,x_2) \in R^{n_1+n_2}$,

$\min_{x_1,x_2} f(x_1,x_2) \quad subject\ to\ g_1(x_1 \leqq 0), g_2(x_2 \leqq 0)$ ⇔ $\min_{x_1} \tilde{f}(x_1) \quad subject\ to\ g_1(x_1 \leqq 0)$

where $\tilde{f}(x_1) = \min\{f(x_1,x_2) : g_2(x_2) \leq 0\}$.

즉 변수 하나씩 먼저 해도된다는 의미

③ Transformations and change of variables

If $h : R \rightarrow R$ is a monotone increasing transformation,

$\min_x f(x) \quad subject\ to\ x \in C \Leftrightarrow \min_x h(f(x)) \quad subject\ to\ x \in C$

④ Eliminating equality constraints and Introducing slack variables

If we can express any feasible point as $x = My + x_0$, where $Ax_0 = b$ and $col(M) = null(A)$

$$\min_x f(x) \quad subject\ to\ \begin{array}{l} g_i(x) \leqq 0 \ \ \mathrm{for}\, i=1,\cdots,m \\ Ax=b \end{array}$$

⇔ $\min_y f(My+x_0) \quad subject\ to\ g_i(My+x_0) \leqq 0 \ \ \mathrm{for}\, i=1,\cdots,m$: eliminating equality contraints

⇔ $\min_x f(x) \quad subject\ to\ \begin{array}{l} s_i \geqq 0 \ \ \mathrm{for}\, i=1,\cdots,m \\ g_i(x)+s_i=0 \ \ \mathrm{for}\, i=1,\cdots,m \\ Ax=b \end{array}$: introducing slack variables

4. Canonical Problem Forms

Linear Programs ⊂ Quadratic Programs ⊂ SemiDefinite Programs ⊂ Conic Programs

가. Linear Programs : $\min_{x} c^T x \quad subject\ to\ \ Dx \leqq d \text{ and } Ax = b$

standard form of LP : $\min_{x} c^T x \quad subject\ to\ \ Ax = b \text{ and } x \geqq 0$

1) example : basis pursuit

변수의 수가 칼럼수보다 많은경우 일종의 변수 선택과정

$\min_{\beta} \ \| \beta \|_0 \quad subject\ to\ \ X\beta = y$를 해야하나 approximation 하여 $\min_{\beta} \ \| \beta \|_1 \quad subject\ to\ \ X\beta = y$

이를 linear form인 $\min_{\beta, z} 1^T z \quad subject\ to\ \ z \geqq \beta,\ z \geqq -\beta,\ X\beta = y$로 표현가능

나. Quadratic Programs : $\min_{x} c^T x + \frac{1}{2} x^T Q x \quad subject\ to\ \ Dx \leqq d \text{ and } Ax = b$ where $Q \geq 0$

$Q \geq 0$는 결국 positive semidefinite를 의미

standard form of QP : $\min_{x} c^T x + \frac{1}{2} x^T Q x \quad subject\ to\ \ Ax = b \text{ and } x \geqq 0$

1) example : lasso

$\min_{\beta} \ \| y - X\beta \|_2^2 \quad subject\ to\ \ \| \beta \|_1 \leqq s$

이를 다시 Lagrange form인 $\min_{\beta} \frac{1}{2} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1$로 표현가능

5. Gradient Descent

가. gradient descent method

To solve convex optimization $\min_x f(x)$ we use gradient descent :

f is convex and differentiable with $dom(f) = R^n$

→ choose initial point $x^{(0)} \in R^n$

repeat $x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)})$ for $k = 1,2,3,\cdots$

1) backtracking line search : t_k를 찾는 기술적 방법

① fix parameters $0 < \beta < 1$ and $0 < \alpha \leqq \frac{1}{2}$

② start with $t = t_{init}$,

③ at each iteration, shrink $t = \beta t$ if $f(x - t\nabla f(x)) > f(x) - \alpha t \| \nabla f(x) \|_2^2$

else perform gradient descent update $x^+ = x - t\nabla f(x)$

2) Exact line search $t = \underset{s \geqq 0}{argmin} f(x - s\nabla f(x))$: 가능은 하나 효율적이지 않음

가) convergence analysis

f is convex and differentiable with $dom(f) = R^n$, ∇f is lipschitz continuous with constant $L > 0$

⇒ gradient descent has convergence rate $O(1/k)$

나) ∇f is lipschitz continuous with constant $L > 0$

$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2$ for any x, y (or when twice differentiable : $\nabla^2 f(x) \leq LI$)

① Theorem

Gradient descent with fixed step size $t \leq 1/L$ satisfies $f(x^{(k)}) - f^* \leq \frac{\| x^{(0)} - x^* \|_2^2}{2tk}$

and same result holds for backtracking, with t replaced by β/L

② Theorem

Gradient descent with fixed step size $t \leq 2/(m+L)$ or with backtracking line search satisfites

$f(x^{(k)}) - f^* \leq \gamma^k \frac{L}{2} \| x^{(0)} - x^* \|_2^2$ where $o < \gamma < 1$

3) 코드를 통한 $x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)})$구현 순서

def ① cost function 목적함수

② gradient function 경사 함수

③ backtracking line search : $f(x - t\nabla f(x)) > f(x) - \alpha t \| \nabla f(x) \|_2^2$

④ early stopping : if $\| \nabla f(x) \|_2^2 \leq \epsilon$ than break (ϵ은 작은 수)

나. Subgradients

subgradient of a convex function f at x : any $g \in R^n$ such that $f(y) \geq f(x) + g^T(y-x)$ for all y

- always exists on the relative interior of $dom(f)$
- if f differentiable at x, then $g = \nabla f(x)$ uniquely

subdifferential : set of all subgradients of convex f

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

- nonempty (only for convex f)
- $\partial f(x)$ is closed and covex (even for nonconvex f)
- if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- if $\partial f(x) = \{g\}$, then f is differentialble at x and $\nabla f(x) = g$

1) subgradient calculus

scaling : $\partial(af) = a \cdot \partial f$

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

affine composition : if $g(x) = f(Ax+b)$, then $\partial g(x) = A^T \partial f(Ax+b)$

finite pointwise maximum : if $f(x) = \max_{i=1,\cdots,m} f_i(x)$, then $\partial f(x) = \mathrm{conv}(\bigcup_{i : f_i(x) = f(x)} \partial \mathrm{f_i(x)})$

2) Subgradient optimality conditon

For any f (covex or not),

$f(x^*) = \min_x f(x) \Leftrightarrow 0 \in \partial f(x^*)$

3) Subgradient Method

f is convex with $dom(f) = R^n$

→ choose initial point $x^{(0)} \in R^n$

repeat $x^{(k)} = x^{(k-1)} - t_k \nabla g^{(k-1)}$ for $k = 1,2,3,\cdots$ where $g^{(k-1)} \in \partial f(x^{(k-1)})$

4) subgradient method이 사용하지 않는 이유(단점)

not necessarily descent method이므로 $f(x_{best}^{(k)}) = \min_{i=0,\cdots,k} f(x^{(i)})$를 해야 함

step size t_k가 작아져야 함 : $\sum_{k=1}^{\infty} t_k^2 < \infty$ and $\sum_{k=1}^{\infty} t_k = \infty$ (square summable, but not summable)

gradient와 차이는 여기서는 pre-specified로 adaptively computed되지 않음

convergence rate가 $O(I/\epsilon^2)$로 느리다. (gradient method의 경우 $O(I/\epsilon)$)

다. Proximal Gradient Descent

$f(x) = g(x) + h(x)$ where g is convex and differentiable with $dom(f) = R^n$, h is convex

→ choose initial point $x^{(0)} \in R^n$

repeat $x^{(k)} = \mathrm{prox}_{h,t_k}(x^{(k-1)} - t_k \nabla g(x^{(k-1)}))$ for $k = 1,2,3,\cdots$

where $\mathrm{prox}_{h,t}(x) = \underset{z}{\mathrm{argmin}} \ \frac{1}{2t} \| x - z \|_2^2 + h(z)$

만약 $G_t(x) = \frac{x - \mathrm{prox_{h,t}(x - t\nabla g(x))}}{t}$라 쓴다면 $x^{(k)} = x^{(k-1)} - t_k G_{t_k}(x^{(k-1)})$

1) backtracking line search : t_k를 찾는 기술적 방법

① fix parameters $0 < \beta < 1$

② start with $t = t_{init}$,

③ at each iteration, shrink $t = \beta t$ if $g(x - G_t(x)) > g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2} \| G_t(x) \|_2^2$

else perform gradient descent update $x^+ = x - tG_t(x)$

라. Accelerated proximal gradient method

$f(x) = g(x) + h(x)$ where g is convex and differentiable with $dom(f) = R^n$, h is convex

→ choose initial point $x^{(0)} \in R^n$

repeat $v = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$ and $x^{(k)} = \mathrm{prox}_{t_k}(v - t_k \nabla g(v))$ for $k = 1,2,\cdots$

마. Stochastic gradient descent method

repeart $x^{(k)} = x^{(k-1)} - t_k \nabla f_{ik}(x^{(k-1)})$ for $k = 1,2,3,\cdots$

where $i_k \in \{1,\cdots,m\}$ is some chosen index at iteration k

바. Mini-batches stochastic gradient descent method

repeart $x^{(k)} = x^{(k-1)} - t_k \frac{1}{b} \sum_{i \in I_k} \nabla f_i(x^{(k-1)})$ for $k = 1,2,3,\cdots$

where $I_k \subseteq \{1,\cdots,m\}$ and $|I_k| = b < m$

사. Gradient Descent convergence rate 비교

gradient descent : $O(1/\epsilon)$, (strong convexity인 경우 $O(\log(1/\epsilon))$)

subgradient descent : $O(1/\epsilon^2)$

proximal gradient descent : $O(1/\epsilon)$

accelation proximal gradient descent : $O(1/\sqrt{\epsilon})$

6. Duality

가. Duality for general form Linear Programs

Given $c \in R^n, A \in R^{m \times n}, b \in R^m, G \in R^{r \times n}, h \in R^r$,

Primal LP : $\min_x c^T x$ subject to $Ax = b$
$Gx \leq h$

⇔ Dual LP : $\max_{u,v} -b^T u - h^T v$ subject to $-A^T u - G^T v = c$
$v \geq 0$

나. Lagrangian

Lagrangian : $L(x,u,v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x)$ where $u \in R^m, v \in R^r, u \geq 0$

Lagrange dual function : $g(u,v) \doteq \min_x L(x,u,v) \leq \min_{x \in C} L(x,u,v) \leq f^*$

다. Lagrange dual problem

Primal problem : $\min_x f(x)$ subject to $h_i(x) \leq 0$ for $i = 1, \cdots, m$
$l_j(x) = 0$ for $i = 1, \cdots, r$

⇔ Lagrange dual problem : $\max_{u,v} g(u,v)$ subject to $u \geq 0$

Lagrange dual problem always hold weak duality, and always convex optimization

- weak duality : if dual optimal value is g^*, then $f^* \geq g^*$
- strong duality : $f^* = g^*$ (KKT condition일 경우 성립)
- duality gap : $f(x) - g(u,v)$

For a problem with strong duality,
x^*, u^*, v^* are primal and dual solutions ⇔ x^*, u^*, v^* satisfy the KKT conditions

라. KKT condition

Given general problem $\min_x f(x)$ subject to $h_i(x) \leq 0$ for $i = 1, \cdots, m$
$l_j(x) = 0$ for $i = 1, \cdots, r$

① stationarity : $0 \in \partial_x (f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x))$

② complementary slcakness : $u_i h_i(x) = 0$ for all i

③ primal feasibility : $h_i(x) \leq 0$ and $l_j(x) = 0$ for all i, j

④ daul feasibility : $u_i \geq 0$ for all i