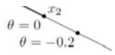
1. 概念

正定矩阵 P 对称，且 x^TPx>0 特征值大于 0

仿射和凸组合

- **affine combination**

$$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \quad \text{with} \quad \theta_1 + \theta_2 + \cdots + \theta_k = 1$$

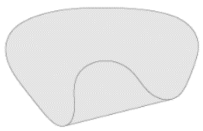


- **convex combination**

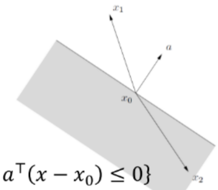
$$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \quad \text{with} \quad \theta_1 + \theta_2 + \cdots + \theta_k = 1 \quad \text{and} \quad \theta_i \ge 0, \forall i$$

凸集：集合内的点的凸组合也属于集合　　内点（领域也在集合内）**空集和全集即开又闭**

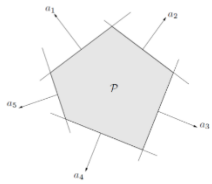
凸包（最小凸集）　　**超平面，半空间（a 是法向量）**　　多面体由超平面和半空间围成



$b\}$



$\{x: a^\top (x - x_0) \le 0\}$



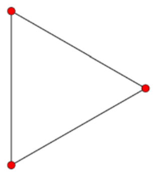
$$\{x: a_i^\top x \le b_i, c_j^\top x = d_j\}$$

仿射独立（一个点无法由其他点的仿射变换表示，**2D 最多 3 个**）　　锥：**过原点**的射线

单纯形　（一系列仿射独立的点的凸包）　　凸锥：锥变换在集合内

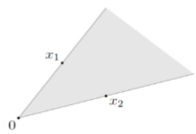
两点一线段，三点一图形

$$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \quad \text{with} \quad \theta_i \ge 0$$



椭球 **P 正定**

$$\{x: (x - x_c)^\top P^{-1} (x - x_c) \le 1\}$$



Proper Cone

Generalized Inequality

- K is a proper cone:
  - closed
  - solid: non-empty interior
  - pointed: contains no line

perspective function $P: \mathbf{R}^{n+1} \to \mathbf{R}^n$

$$P(x, t) = x/t$$

保凸运算　1.交集　2.仿射变换（凸集的和，放大）3.透视降维

4.线性分式函数　仿射+透视

$$f(x) = \frac{Ax + b}{c^\top x + d}$$

dom $f = \{x | c^\top x + d > 0\}$

5. 非负加权和

$$f = w_1 f_1 + \cdots + w_m f_m, w_m \ge 0$$

6.逐点上确界　$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex

透视（KL 散度）　g(x,t)

- $f(x) = -\log x$ is convex and its perspective function is

$$g(x,t) = tf(x/t) = -t\log\frac{x}{t} = t\log t - t\log x$$

- Kullback-Leibler divergence

$$D_{kl}(u, v) = \sum_{i}^{n} (u_i \log(u_i/v_i) - u_i + v_i)$$

7 最小值和舒尔补

- if $f(x, y)$ is convex w.r.t. (x, y) and C is a convex set, then the minimization $g(x) = \inf_{y \in C} f(x, y)$ is convex
- distance to a convex set S

$$\mathrm{dist}(x, S) \triangleq g(x) = \inf_{y \in S} \|x - y\|$$

difference to
$f(x) = \sup_{y \in C} \|x - y\|$

- Shur complement
  - $f(x,y) = x^\top Ax + 2x^\top By + y^\top Cy$ with $\begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \ge 0, C > 0$

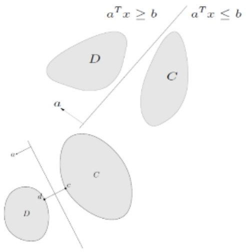
分离超平面（必须凸集，不可逆，**严格的话不可取等号**）　　支撑超平面（必须凸集）

Separating Hyperplane Theorem

- if C and D are disjoint convex sets, i.e., $C \cap D = \emptyset$, then there exists at least one **separating hyperplane** such that

$$a^\top x \le b, \forall x \in C$$
$$a^\top x \ge b, \forall x \in D$$

- Proof sketch for the case, the distance between C and D are positive
  - find the closest points
  - define the hyperplane by the middle point and the orthogonal direction



Supporting Hyperplane Theorem

For a set $C \subseteq \mathbf{R}^n$ and a point in its boundary x_0, if $a \ne 0$ satisifes $a^\top x \le a^\top x_0, \forall x \in C$, then we call the corresponding hyperplane a **supporting hyperplane** to C at x_0

- **Supporting Hyperplane Theorem**: for a convex set, there exists at least one supporting hyperplane at every boundary point.

~~how about non-convex sets?~~
can we have more?

- a convex set could be represented by (maybe infinite) linear inequalities
- (partial converse): if a set is close, has nonempty interior, and has a supporting hyperplane at every point in its boundary, then it is convex.

PANDA 2020/5/28, 16..
vertex

凸集可以由线性不等式的解集表示（通过分离超平面定理）

2. 凸函数（定义域，可不光滑，是否严格） 一阶和二阶

Definition

PANDA 2020/5/28, 17:00:02
定义域必须是凸集

- $f: R^n \to R$ is convex, if **dom** f is a convex set and

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in \textbf{dom}\, f, \theta \in [0,1]$$

- **"line is above the curve":**

- **strictly convex** if strict inequality holds for $x \ne y$ and $\theta \in (0,1)$

- **First-order condition**

$$f(y) \ge f(x) + \nabla f(x)^{\mathsf{T}}(y-x) \quad \forall x, y \in \textbf{dom}\, f$$

"tangent plane is below the surface"

$\nabla^2 f(x) \ge 0, \forall x \in \textbf{dom}\, f \iff f$ is convex

$\nabla^2 f(x) > 0, \forall x \in \textbf{dom}\, f \implies f$ is strictly convex

水平集（**针对定义域的**，保凸）

sublevel set of $f: \mathrm{R}^n \to \mathrm{R}$ $\quad C_\alpha = \{x \in \textbf{dom}\, f : f(x) \le \alpha\}$

sublevel sets of convex functions are convex

上镜图（升维，保凸）

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

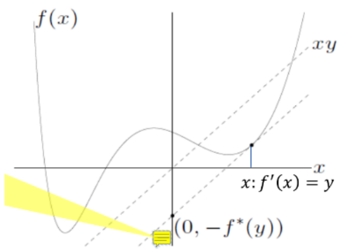
$$\textbf{epi}\, f = \{(x,t) \in \mathbf{R}^{n+1}\ x \in \textbf{dom}\, f : f(x) \le t\}$$

- f is convex iff **epi** f is a convex set

共轭函数

切线和函数差值最大值，y 为斜率

在 f(x)'=y 的点找 永远是凸函数



for a given y, $f^*(y)$ is the largest gap between the affine function xy and the function $f(x)$

when f is differentiable, the optimality condition tells

$$\nabla(y^{\mathsf{T}}x - f(x)) = y - f'(x) = 0$$

3. 凸优化问题 在凸集上找凸函数的最小值

可行解属于可行集（凸函数和约束的定义域交集）

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

全局最优解 f0(x)最小

局部最优解：距离 x 范围 d 内，f0(x+d)>=f0(x)

线性规划（relaxation 法）：例如整数约束变成非整数

线性分式（规划换变量法）

二次规划（梯度法）向量优化

反正拉格朗日对偶最好用

Linear-fractional Programming

$d = f = 0$

$$\begin{aligned} \min \quad & \frac{c^{\mathsf{T}}x + d}{e^{\mathsf{T}}x + f} \\ \text{s.t.} \quad & Gx \preccurlyeq h \\ & Ax = b \\ & e^{\mathsf{T}}x + f > 0 \end{aligned} \qquad y = \frac{x}{e^{\mathsf{T}}x + f}, \quad z = \frac{1}{e^{\mathsf{T}}x + f}$$

- **Linear-fractional Programming** is equivalent to the following LP

$x, y \in \mathbf{R}^n$, $z \in \mathbf{R}$

$$\begin{aligned} \min \quad & c^{\mathsf{T}}y + dz \\ \text{s.t.} \quad & Gy \preccurlyeq hz \\ & Ay = bz \\ & e^{\mathsf{T}}y + fz = 1 \\ & z \ge 0 \end{aligned} \qquad x = y/z$$

4. 常用模型

监督学习

线性拟合用最小二乘法 缺点：偏离值影响太大

解决方法：**huber loss**（连续和光滑问题）

Huber loss

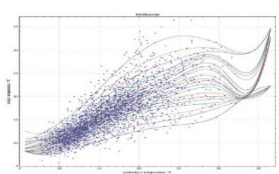
$$L_{\text{huber}}(r) = \begin{cases} r^2 & \text{if } |r| < c \\ |r| & \text{if } |r| \ge c \end{cases}$$

分位拟合：拟合曲线上下点成固定比例

esult is the median value

regression with asymmetric loss

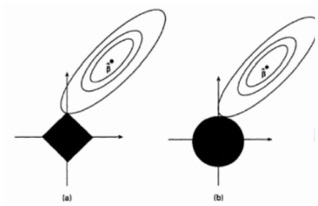
$$L_\tau(r) = \begin{cases} -(1-\tau)r & \text{if } r < 0 \\ \tau r & \text{if } r \ge 0 \end{cases}$$



正则化项： 额外 min 所有参数

Lasso: l1norm，使系数稀疏

Ridge: l2 norm

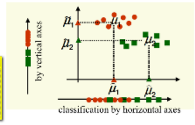


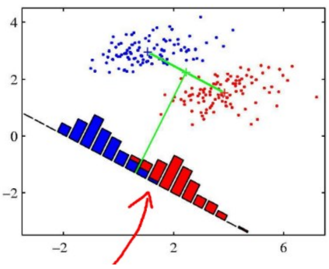
LDA: 在直线上投影均值最大

方差越小越好

- Fisher Discriminant

$$\max_{x} \frac{(x^\top \bar{a}_+ - x^\top \bar{a}_-)^2}{S_+^2 + S_-^2}$$





CCA 找到两条线,使得数据上面的投影相关性最大

数据 A,B 描述同个 object

$$\max_{x,y} \frac{x^\top \mathrm{cov}(A,B) y}{\sqrt{x^\top \mathrm{cov}(A,A) x}\sqrt{y^\top \mathrm{cov}(A,A) y}}$$

SVM 最大边距 硬边距

$$\min_{x,z} \ \|x\|_2^2 \quad \text{s.t.} \quad b_i(x^\top b_i + z) \geq 1$$

软边距

$$\min_{x,z} \ \frac{1}{2}\|x\|_2^2 + C\sum s_i \quad \text{s.t.} \quad b_i(x^\top b_i + z) \geq 1 - s_i, \quad s_i \geq 0$$

hinge loss（和分类结果<=1 数据有关）

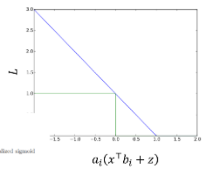
- loss function + regularization term

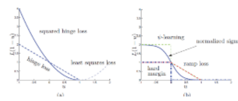
$$\min_{x,z} \ \frac{1}{2}\|x\|_2^2 + C\sum_i L(b_i(x^\top a_i + z))$$

- hinge loss

$$L(u) = \max\{0, 1-u\}$$

- the convex approximation for misclassification loss





非监督学习

降维分析

主成分分析 PCA（线性分有用）

A 是数据，x 和 x^T 是主成分和无效成分，x 就是 C 的特征向量

- find a direction that maximizes the data's variance

$$\max_{x^T x = 1} \ x^T C x$$

$$C = \frac{1}{n-1}\boxed{A^T A}$$

it should be zero-mean, otherwise, the covariance matrix will be ?

局部线性嵌入 LLE（流形学习）

1 用一个点周围的点的加权和表示它（用最小二乘法得到权重系数）

2 通过低维的线性变换矩阵 b 尽量保留所有点和周围点的加权关系

$$M = (I - W)^T (I - W)$$

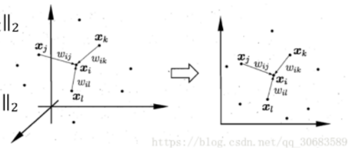
W 是权重系数矩阵，特征向量为低维特征向量

- find the relationship $a_0 = \sum x_i a_i$

$$\min_x \ \|a_0 - \sum x_i a_i\|_2$$

- keep the relationship in reduction

$$\min_b \ \|b_0 - \sum x_i b_i\|_2$$

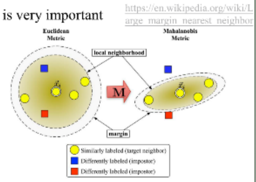


聚类 K-means 质心迭代

度量学习

Metric Learning

- distance/similarity/dissimilarity/metric is very important
  - non-negative
  - identity
  - symmetry
  - subadditivity (triangle inequality)
- Mahalanobis distance



$$d_M(a,b) = \sqrt{(La - Lb)^\top (La - Lb)} = \sqrt{(a-b)^\top L^\top L (a-b)} = \sqrt{(a-b)^\top M (a-b)}$$

$$\min_M \sum_{i,j \in N_i} d(a_i, a_j) + \lambda \sum_{i,j,l} \xi_{ijl} \quad \text{s.t.} \begin{cases} d(a_i,a_j) + 1 \leq d(a_i,a_l) + \xi_{ijl}, \forall j \in N_i, l \notin N_i, \forall i \\ \xi_{ijl} \geq 0 \\ M \succcurlyeq 0 \end{cases}$$

自编码器（高维数据编码至低维神经元再解码）

目标：**神经网络输出和输入一样**

5．无约束问题解法

1.OLS 直接求梯度为 0

Pseudo inverse 计算量大

$$\min_x \sum_{i=1}^m (a_i^\top x - y_i)^2$$

$$\min_x \|Ax - Y\|_2^2 = \min_x \ (Ax - Y)^\top (Ax - Y)$$

$$\nabla \|Ax - Y\|_2^2 = \nabla (Ax - Y)^\top (Ax - Y) = 2A^\top (Ax - Y) = 0$$

$$x^* = (A^\top A)^{-1} A^\top Y$$

2 梯度下降 梯度为 0 需要 strong convexity 复杂度和 m/M 成线性关系

General descent method.

given a starting point $x \in \mathbf{dom}\, f$.

repeat

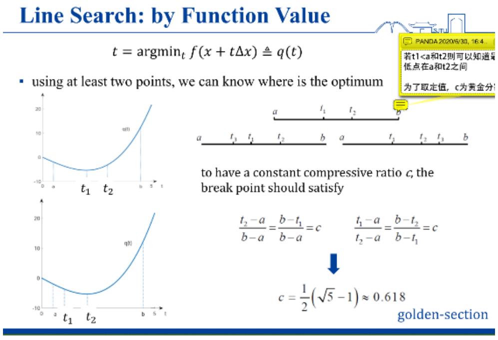
1. Determine a descent direction Δx.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

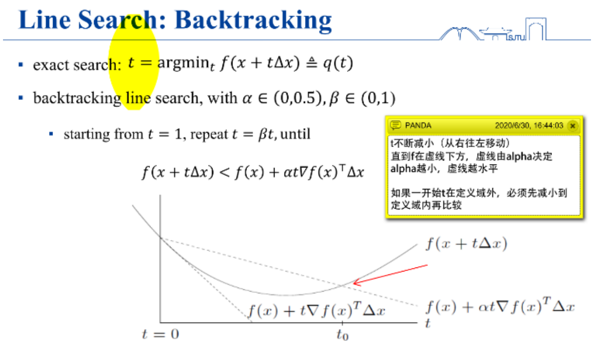
until stopping criterion is satisfied.

- staring point $x^{(0)}$
- descent direction $\Delta x^{(k)}$
- step size /step length/ learning rate $t^{(k)}$

$$mI \leq \nabla^2 f(x) \leq MI,$$

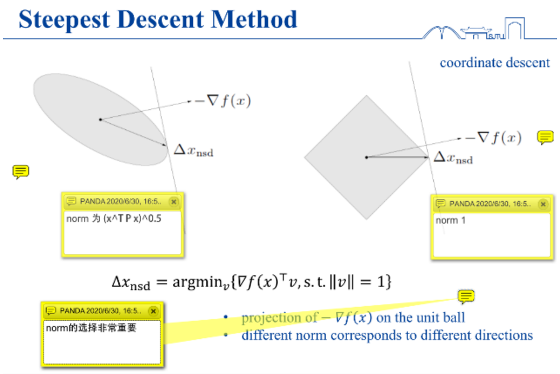
line search　exact line search:找一条直线上 f(x)的最小值,计算量大





最速下降法 先假设模长和梯度，找负梯度上投影最长点（不同 norm 结果不同）

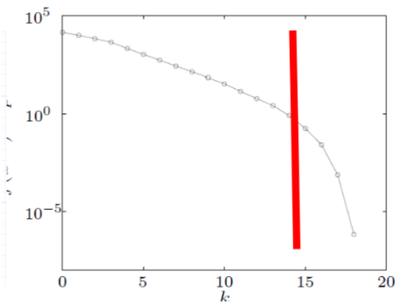
牛顿法 仿射不变 最优解的仿射变换是仿射变换后最优解，t^k 用 line search



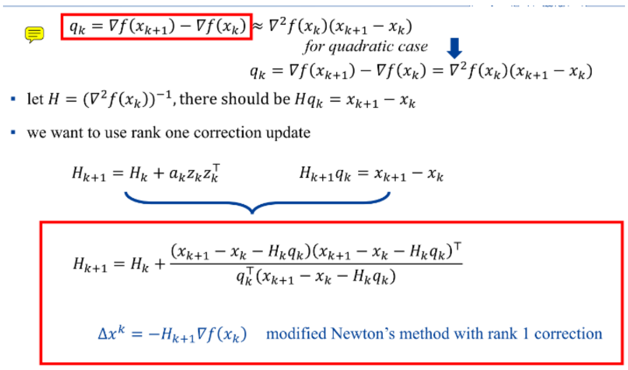
$$x^{k+1} = x^k + t^k d^k \qquad d^k = \nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

Damp phase 下降速度为 c 线性

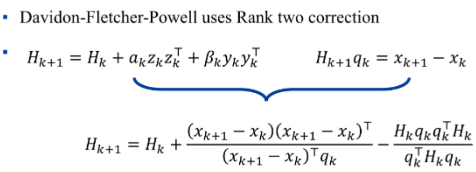
Quadratic phase 速度很快，6 步到位



拟牛顿法



DFP 和 BFGS(都是求二阶黑塞矩阵)



BFGS　　**L-BFGS 直接算** $B_{k+1}\nabla f(x_{k+1})$

$$B_{k+1} = B_k - \frac{B_k(x_{k+1}-x_k)^T B_k}{(x_{k+1}-x_k)^T B_k (x_{k+1}-x_k)} + \frac{q_k q_k^T}{q_k^T(x_{k+1}-x_k)}$$

动量法：Polyaks, Nestrov

- Polyak's momentum algorithm

$$x_{k+1} = x_k - t_k\nabla f(x_k) + \delta_k(x_k - x_{k-1}) \qquad x_{k+1} \approx x_k - t_k \nabla f(x_k + \delta_k(x_k - x_{k-1})) + \delta_k(x^k - x^{k-1})$$

实用的 SGD 实用少数样本更新梯度 计算量小 能收敛 需要 normalization

shuffling: pure randomness to avoid biasness

curriculum: supplying the training examples in a meaningful order may actually lead to improved performance

Adagrad(学习率一直减小)　　RMSprop（梯度大学习率小）　　Adam

$$G_{ii} = \Sigma_{t=1}^{k}\left(\nabla f(x_t)(i)\right)^2$$

$$x_{k+1} = x_k - \frac{\eta}{\sqrt{G_{ii}+\varepsilon}}\nabla f(x_k)$$

$$z_{k+1} = \rho z_k + (1-\rho)\left(\nabla f(x_k)\cdot\nabla f(x_k)\right)$$

$$x_{k+1} = x_k - \frac{\eta}{\sqrt{z_{k+1}+\varepsilon}}\nabla f(x_k)$$

also for stoch

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0,1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t+1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1)\cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2)\cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t/(1-\beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t/(1-\beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha\cdot\widehat{m}_t/(\sqrt{\widehat{v}_t}+\epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

ISTA FISTA

$$\min_x \quad \lambda\|x\|_1 + \|Ax-B\|_2^2$$

$$x^{k+1} = S_\lambda\left(x^k - A^\top(Ax^k - B)\right)$$

$$S_\lambda = \underset{\mathbf{x}}{\operatorname{argmin}}\left\{g(\mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)\right\|^2\right\}$$

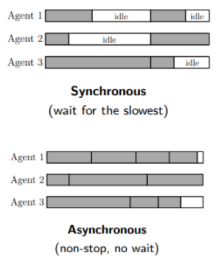
$$y^{k+1} = S_\lambda\left(x^k - A^\top(B - Ax^k)\right)$$

$$t^{k+1} = \frac{1+\sqrt{1+4t_k}}{2}$$

$$x^{k+1} = y^{k+1} + \left(\frac{t^k - 1}{t^{k+1}}\right)\left(y^{k+1} - y^k\right)$$

$$\begin{aligned} x^{k+1} &= x^k + \lambda_k A^\top(B - Ax^k) \\ &= (I - \lambda_k A^\top A)x^k + \lambda_k A^\top B \\ &\triangleq T_k x^k + b_k \end{aligned}$$

Parallel computing（同时收敛，非同时效率高）
x 没有高项就可以拆分 x 并行算



6. 拉格朗日对偶

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1,\dots,m \\ & h_i(x) = 0, i = 1,\dots,p \end{aligned}$$

$$g(\lambda,\nu) = \inf_{x\in D} L(x,\lambda,\nu)$$

$$= \inf_{x\in D}\left(f_0(x) + \sum_{i=1}^{m}\lambda_i f_i(x) + \sum_{i=1}^{p}\nu_i h_i(x)\right)$$

- **Primal**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, i = 1,\dots,m \\ & h_i(x) = 0, i = 1,\dots,p \end{aligned}$$

$f^* = f_0(x^*)$

Dual

$$\begin{aligned} \max \quad & g(\lambda,\nu) = \inf_{x\in D}\left(f_0(x) + \nu^\top f(x) + \nu^T h(x)\right) \\ \text{s.t.} \quad & \lambda \ge 0 \end{aligned}$$

$g^* = g(\lambda^*, \nu^*)$

Slater 条件

Slater's constraint qualification requires the problem is strictly feasible:

$$\exists x \in \operatorname{int} D, f_i(x) < 0, h_i(x) = 0$$

- if the problem is convex and the Slater's qualification satisfied, then there is

strong duality

KKT 条件（必要，slater 满足则充要）

primal feasible	$f_i(x^*) \le 0$
	$h_i(x^*) = 0$
dual feasible	$\lambda_i^* \ge 0$
complementary slackness	$\lambda_i^* f_i(x^*) = 0$

$$\nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_i \nu_i^* h(x^*) = 0$$

7 大规模数据

$x^{k+1} = P_D(x^k - t_k \Delta x_{nt})$

约束（投影）梯度下降法

ADMM（轮流更新两种参数）

增广项影响并行运算

Alternating direction method of multipliers

- **Dual decomposition: Lagrangian**

$$L(x, z; \alpha) = \lambda \|x\|_1 + \|b - Az\|_2^2 + \alpha^T(x - z)$$

PANDA 2020/7/21, 16:5...
增广项，也是惩罚项

- **Method of multipliers: augmented Lagrangian**

$$L_\rho(x, z; \alpha) = \lambda \|x\|_1 + \|b - Az\|_2^2 + \alpha^T(x - z) + \frac{\rho}{2}\boxed{\|x - z\|_2^2}$$

strong duality

condition number adjustment

- **ADMM**

$$x^{k+1} = \text{argmin}_x \; L_\rho(x, v^k; \alpha^k)$$

$$z^{k+1} = \text{argmin}_z \; L_\rho(x^{k+1}, z; \alpha^k)$$

$$\alpha^{k+1} = \alpha^k + \boxed{\rho}(x^{k+1} - z^{k+1})$$

PANDA 2020/7/27, 16:3...
固定一个参数更新另一个

PANDA 2020/7/21, 16:5...
这个rou由KKT中的导数为0得到

Alternating direction method of multipliers

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c \end{aligned}$$

- **Method of multipliers: augmented Lagrangian**

$$L_\rho(x, z; \alpha) = \lambda f(x) + g(z) + \alpha^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

- **ADMM**

$$x^{k+1} = \text{argmin}_x \; L_\rho(x, z^k; \alpha^k)$$

$$z^{k+1} = \text{argmin}_z \; L_\rho(x^{k+1}, z; \alpha^k)$$

$$\alpha^{k+1} = \alpha^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

z^{k+1} minimize $L(x^{k+1}, z, y^k)$

$\Rightarrow 0 = \nabla_z L_\rho(x^{k+1}, z^{k+1}, y^k)$

$= \nabla g(z^{k+1}) + B^T y^k + \rho B^T(Ax^{k+1} + Bz^{k+1} - c)$

$= \nabla g(z^{k+1}) + B^T(y^k + \underbrace{\rho(Ax^{k+1} + Bz^{k+1} - c)})$

$(x^{k+1}, z^{k+1}, y^{k+1}) = \nabla g(z^{k+1}) + B^T y^{k+1}$

Satisfies 2nd dual feasibility.

Frank-Wolfe method

...mizing over all Previous Directions (with $s^{(0)} := x^{(0)}$)

... as Algorithm 2, except replacing line b) with

b') Update $x^{(k+1)} := \underset{x \in \text{conv}(s^{(0)}, \dots, s^{(k+1)})}{\arg\min} f(x)$

$$\min_x \; f(x) \;\; \text{s.t.} \; x \in D$$

$$z^k = \underset{z}{\text{argmin}} \;\; z^\top \nabla f(x^k) \;\; \text{s.t.} \; z \in D$$

there is no projection

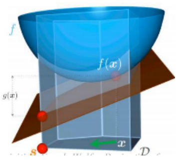
affine invariance

set the direction as $\Delta x^k = z^k - x^k$

set $t^k = \frac{2}{k+1}$ or do line search

$$x^{k+1} = x^k - t^k \Delta x^k$$

x^k is kept in the feasible domain



x-update

$$\begin{aligned} x^{k+1} &= \text{argmin}_x \; L_\rho(x, z^k; \alpha^k) \\ &= \text{argmin}_x \; \lambda \|x\|_1 + x^\top \alpha^k + \frac{\rho}{2}\|x - z^k\|_2^2 \\ &= S_{\lambda/\rho}(z^k + \alpha^k/\rho) \end{aligned}$$

- **z-update**

$$\begin{aligned} z^{k+1} &= \text{argmin}_z \; L_\rho(x^{k+1}, z; \alpha^k) \\ &= \text{argmin}_z \; \|b - Az\|_2^2 + z^\top \alpha^k + \frac{\rho}{2}\|x^{k+1} - z\|_2^2 \\ &= (A^\top A + \rho I)^{-1}(A^\top b + \rho x^{k+1} - \alpha^k) \end{aligned}$$