SEEM 5380: Optimization Methods for High-Dimensional Statistics

Term 2, 2017-18

Instructor: Prof. Anthony Man-Cho So
Scribe: Zihao FU

2019.1.7

Natural of this course

Less structured

like a seminar

Recent advance

方式：

阅读论文 → 讲座

What's the course about

Integration of opt. and stat. techniques

Case study: Statistical estimation problems

- Samples: $z_1, z_2, \ldots z_n$

  n个样本，

- parameters: $\theta_1, \theta_2, \cdots, \theta_d$

  d个参数。

e.g.: linear regression

一般都先假设有一个generative model

$$y = X\theta^* + \varepsilon \qquad z = (x, y) \text{ samples}$$

$y \in \mathbb{R}^n$, $\theta^* \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^n$ noise

θ^* : star means ground truth

Classical Setting:

$n \gg d$ (Overdetermined) 通常假设 n远大于d

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2 \quad \text{Least-square estimator}$$

θ^* 是 ground truth (we don't know)

$\hat{\theta}$ 是 optimal solution to the opt. problem called estimator.

Note: $\hat{\theta} \neq \theta^*$ in general

find the error bound between $\hat{\theta}$ and θ^*

From Stat. we know:

If $\varepsilon \sim N(0, \sigma^2)$, Then the maximum likelihood estimator (MLE) is given by $\hat{\theta}$ (LSE那个)

Issues/Observations:

① Optimization side:

how to solve for $\hat{\theta}$

θ is convex $\therefore$ is polynomial-time solvable

KKT 是充要.

KKT: $\underbrace{X^T X}_{d \times d} \theta = X^T y$

nearly full rank.

是否存在更加 lightweight method?

e.g. Gradient Descent.

Gradient 是上升方向.

$$\theta^{k+1} \leftarrow \theta^k - \alpha_k X^T (X\theta^k - y)$$

α_k: step size > 0

那 performance 如何?

要做 convergence analysis

1) Does $\{\theta^k\}$ converge? to where?

convex 情况下必然收敛到最优

study to where is important especially non-convex

2) Convergence Rate?

$$F(\theta^k) - F(\hat{\theta}) \le h(k)$$

current value. optimal value ← bound, 希望 $h(k) \to 0$

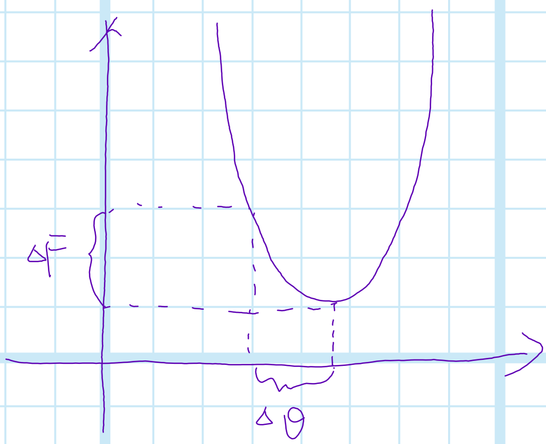
在上面的问题中 $F(\theta) = \|y - X\theta\|_2^2$

实际上 $F(\theta)$ 的值并不重要, 重要的是找到 $\hat{\theta}$

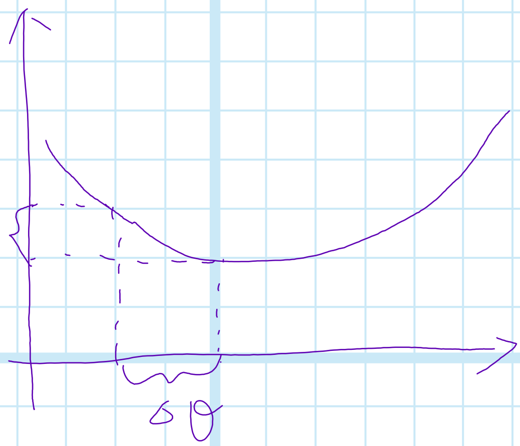
我们是否能直接找一个

$$\|\theta^k - \hat{\theta}\| \le h_2(k) \to 0$$

因为即使 $f(\theta)$ 收敛很快 $\|\theta^k - \hat{\theta}\|$ 也不一定收敛快. 如下图:



High curvature



Low curvature

在 low curvature 中 F 快收敛了, θ 却还差很多. 对于高维找是 Hessian 矩阵.

② Statistical error:

$$\|\hat{\theta} - \theta^*\| \leq ?$$

虽然我们解出了最好的 $\hat{\theta}$，但是它和 ground truth 差多少?

optimization 方面只关注了 θ_k 和 $\hat{\theta}$ 间的差异，并未涉及到 θ^*

Extentions to other settings

- in many applications:

$$y = X\theta^* + \varepsilon, \quad d >> n \text{ (underdetermined)}$$

$$n\,[\,y\,] = \underbrace{\overset{X}{[\qquad]}}_{n\times d} \cdot [\,\theta^*\,] + [\,\varepsilon\,]$$

例如 d 是基因个数，n 是病人样本数.

have too many parameter and few data can always fit it perfectly.

60~70年代，high dimentional 指的是样本很多.
现在指的是参数比样本多.

Need make more assumptions.

实际上就是假设数据具有一些结构信息，从而来降维

embed low-dimensional structure in θ^*

假设 θ^* 存在于某个特殊结构上.

e.g. sparisity：虽然 θ 维度很高，但是很多维度是不起作用的。

如何加 sparisity 的信息：

Idealy: 转成 regularized problem.

① $\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \|y - X\theta\|_2^2 + \lambda \|\theta\|_0$

$\|\theta\|_0 \triangleq |\{i : \theta_i \neq 0\}|$

$\|\theta\|_0$：θ 中非零元的个数

λ：>0，regularization parameter.

同样两个问题：

① optimization aspect

② stats.

求解方法：

① $\|\theta\|_0$ 是非凸的.（都不是连续的）

0-norm 不好求，直接用 1-norm 来近似.

$\hat{\theta}' \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \{ \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \}$ (LASSO)

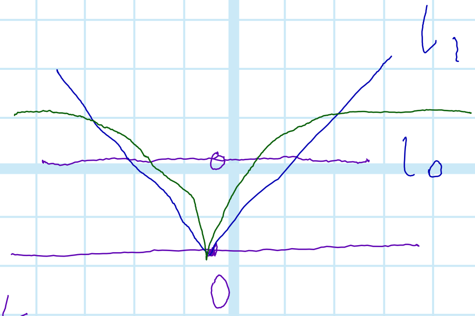
LASSO：least absolute shrinkage and

Selection Operator.

注意到上面用的 $\hat{\Theta}' \in$，有一个绝全都满足。是使用哪个要用stats的方法，

② 先画出 $|x|_0$ 的图像：

$$|x|_0 = \begin{cases} 1 & x \neq 0 \\ 0 & x = 0 \end{cases}$$



l_1 是近似 l_0，可以更近一步，如绿线一样继续逼近 $|x|_0$

$$\hat{\Theta}^R \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \{ \|y - X\theta\|_2^2 + \lambda \underbrace{R(\theta)}_{\text{regularizer}} \}$$

好处是更逼近 l_0 了，
坏处是问题又变成了非凸的问题(原因)

③ 上面两种方法都是为了逼近 l_0，原因是 ① 是 intractable 的，之所以是 intractable 因为问题是 NP 的。能不能直接解？

Mixed-integer optimization

问题是 practical tractable，即不在最坏情况下，还是 tractable 的。所以并不能用到 deep learning 上。

2019.1.8

Linear Regression

$$\underset{n}{y} = \underset{n\times d}{X}\,\underset{d}{\theta^*} + \underset{n}{w}$$ 就是上节课的 ε

θ^*: ground-truth

w: noise

X: design matrix

Warm-up: classic setting $n \gg d$

Least-squares estimate:

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^d}{\arg\min}\ \frac{1}{2}\|y - X\theta\|_2^2$$

Statistics Error: $\|\hat{\theta} - \theta^*\|_2$

① $\frac{1}{2}\|y - X\hat{\theta}\|_2^2 \le \frac{1}{2}\|y - X\theta^*\|_2^2$

optimality of $\hat{\theta}$ $\because \hat{\theta}$ 是最优解

② Define $\hat{\Delta} = \hat{\theta} - \theta^*$

展开

$$\frac{1}{2}\left(y^Ty - 2y^TX\hat{\theta} + \hat{\theta}^TX^TX\hat{\theta}\right) \le \frac{1}{2}\left(y^Ty - 2y^TX\theta^* + \theta^{*T}X^TX\theta^*\right)$$

利用 $\hat{\Delta}$ 的定义，以及 $X\theta^* = y - w$

展开后：$\hat{\theta}^T X^T X \hat{\theta} - \theta^{*T} X^T X \theta^* \leq 2 y^T X \hat{\Delta}$

代入 $y = X\theta^* + w$：

$\hat{\theta}^T X^T X \hat{\theta} + \theta^{*T} X^T X \theta^* - 2\theta^{*T} X^T X \hat{\theta} \leq 2 w^T X \hat{\Delta}$

$$\|X\hat{\Delta}\|_2^2 \leq 2\hat{\Delta}^T X^T w$$

③ Assume X has full column rank (因为 $n \gg d$)

$\Rightarrow$ $X^T X$ is invertible $\lambda_{min}(X^T X) > 0$

By the Courant - Fischer Theorm

$$\left(\lambda_{min}(A) = \min_{\text{s.t. } \|x\|_2 = 1} x^T A x \qquad A \in S^n \right)$$

$$\|X\hat{\Delta}\|_2^2 \geq \lambda_{min}(X^T X) \cdot \|\hat{\Delta}\|_2^2$$

④ $\lambda_{min}(X^T X) \cdot \|\hat{\Delta}\|_2^2 \leq \|X\hat{\Delta}\|_2^2 \leq 2\hat{\Delta}^T X^T w$

$\leq 2\|\hat{\Delta}\|_2 \cdot \|X^T w\|_2$ (Cauchy - Swartz)

$$\Rightarrow \|\hat{\Delta}\|_2 \leq \frac{2}{\lambda_{min}(X^T X)} \|X^T w\|_2 \qquad (*)$$

推出了一个 bound

Rmk:

ⓐ The bound $(*)$ is a deterministic bound with various statistical modess for X w, then the bound can be probabilistic

这里X确定,就是一个deterministic bound
如果假设了X的概率分布就有一个high probabilistic bound.

e.g.

① $X_{ij} \sim N(0,1)$ w bounded

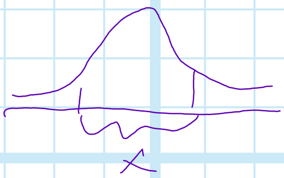
$X^T w = \sum_i w_i x_i$ x_i: ith col of X^T

高斯变量的和仍是高斯变量.

② X has bounded col, $w \in N(0, \sigma^2 I)$

$\Rightarrow$ concentration of measure to get high-prob bounds on $\lambda_{min}(X^T X)$ or $\|X^T w\|_2$

measure 就是 mass, concentration 就是说



很窄的 x 集中了很多的 mass.

ⓑ We use Cauchy-Schwarz at the last step. 也可以用 Hölder-inequality (C-S不等式的推广). 因此:

$2\hat{\Delta}^T X^T w \le 2\|\hat{\Delta}\|_2 \|X^T w\|_2$ 变成了:

$$\le 2\|\hat{\Delta}\|_p \|X^T w\|_q \quad \frac{1}{p}+\frac{1}{q}=1$$

例如 $p=1, q=\infty$, 在sparsity场景中更好.

$\|u\|_2 \le \|u\|_1 \le \sqrt{n}\|u\|_2$

$\because \|u\|_1 = \sum_i |u_i| \cdot 1 \le \left(\sum |u_i|^2\right)^{\frac{1}{2}} \left(\sum_i 1^2\right)^{\frac{1}{2}}$

如果 u 是 sparse vector, $u \begin{bmatrix} \vdots \\ 0 \end{bmatrix} \begin{matrix} \}s \\ \}n \end{matrix}$

把 2-norm 换成 1-norm 不会差太多 ($\sqrt{s}$)

Recall X^TX is invertible $X^TX > 0$ PD

Incidentally, X^TX is Hessian of the loss function: $\mathcal{L}(\theta) = \frac{1}{2}\|y - X\theta\|_2^2$

$\Rightarrow \mathcal{L}$ is strongly convex ($\because$ Hessian 是 PD)

Definition/Propisition We say $f: \mathbb{R}^d \to \mathbb{R}$ is strongly convex with modulus $c > 0$ (aka: c-strongly convex) if any of the following equivalent conditions holds:

① $\forall x, y \in \mathbb{R}^d$, $\alpha \in [0,1]$

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y) \underbrace{- \frac{1}{2} c\alpha(1-\alpha)\|y-x\|_2^2}_{\text{减掉一些仍然 convex}}$$

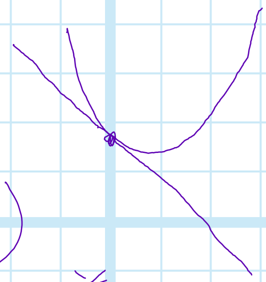
② The function $f(x) - \frac{1}{2}c\|x\|_2^2$ is convex

(weakly convex 指是加一些东西后 convex)

③ In the presence of differentiability

$\forall x, y \in \mathbb{R}^d$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

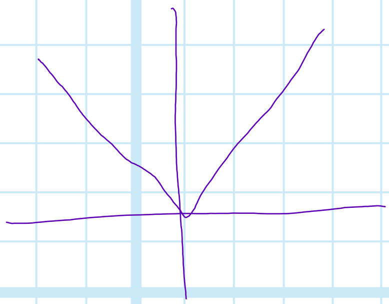


← 普通 convex

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} c \|y-x\|_2^2$$

← strongly convex

lower bound 从线性函数变成一个二次函数.



是 convex, 但不是 strongly convex, ∵ 找不到一个二次函数的LB

④ (In the presence of 2^{nd} order diff)

$\forall x \in \mathbb{R}^d$, $v^T \nabla^2 f(x) v \geq c \cdot \|v\|_2^2 \quad \forall v \in \mathbb{R}^d$

普通 convex 的 Hessian 是 PSD, 而 strongly convex 的 Hessian 是 PD

下面看 strongly convex 如何帮助 optimization.

Optimization Aspect

Def: Let $f: \mathbb{R}^d \to \mathbb{R}$ be continuously diff., We say f has L-Lipschitz continuous gradient if $\forall x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x-y\|_2$$

原始的 Lipschitz 描述 针对 $f(x)-f(y)$，现在扩展到梯度上.

e.g. $\mathcal{L}(\theta) = \frac{1}{2}\|y - X\theta\|_2^2$, $\nabla\mathcal{L}(\theta) = X^T(X\theta - y)$

$$\|\nabla\mathcal{L}(\theta_1) - \nabla\mathcal{L}(\theta_2)\|_2 = \|X^T X(\theta_1 - \theta_2)\|_2$$

$$\leq \|X^T X\| \cdot \|\theta_1 - \theta_2\|_2$$

针对那个 L

∴ 我们之前那个 $\mathcal{L}(\theta)$ 是 gradient-Lipschitz 连续的.

注意 $L(\theta)$ 不是 Lipchitz 连续.

Prop: Let f be c-strongly convex and have L-Lipschitz gradient. Then, $\forall x, y \in \mathbb{R}^d$

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{cL}{c+L}\|x-y\|_2^2 + \frac{1}{c+L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Assume this prop, we study the convergence behavior of the gradient method for solving:

(P) $\min_{\theta \in \mathbb{R}^d} f(\theta)$ f is as in the prop (strongly convex, Lipschitz-gradient)

Gradient method: $\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla f(\theta^k)$

Thm: Let f be as in the prop, Suppose $\alpha_k \equiv \alpha$, $\alpha \in \left(0, \frac{2}{c+L}\right)$, Then,

$$\|\theta^k - \hat{\theta}\|_2^2 \le \left(1 - \frac{2\alpha c L}{c+L}\right)^k \|\theta^0 - \hat{\theta}\|_2^2 \quad (\triangle)$$

$\hat{\theta}$ is opt soln to (P)

就是说 θ^k 收敛到 $\hat{\theta}$

observation:

$\left(1 - \frac{2\alpha c L}{c+L}\right)^k < 1$ ∴ 每一步都在减小.

Pf: $\|\theta^{k+1} - \hat{\theta}\|_2^2 = \|\theta^k - \alpha \nabla f(\theta^k) - \hat{\theta}\|_2^2$

$= \|\theta^k - \hat{\theta}\|_2^2 - 2\alpha \nabla f(\theta^k)^T(\theta^k - \hat{\theta}) + \alpha^2 \|\nabla f(\theta^k)\|_2^2$

又观察到 gradient Lipschitz 有两项梯度, 而这里只有一项, 注意到 $\nabla f(\hat{\theta}) = 0$ ($\hat{\theta}$ is minimizer)

∴ 有 $\nabla f(\theta^k)^T(\theta^k - \hat{\theta}) = (\nabla f(\theta^k) - \underbrace{\nabla f(\hat{\theta})}_{=0})(\theta^k - \hat{\theta})$

$\ge \frac{cL}{c+L}\|\theta^k - \hat{\theta}\|_2^2 + \frac{1}{c+L}\|\nabla f(\theta^k)\|_2^2$ 代入得:

← SCVX prop 的结论.

$\le \left(1 - \frac{2\alpha c L}{c+L}\right)\|\theta^k - \hat{\theta}\|_2^2 + \underbrace{\alpha\left(\alpha - \frac{2}{c+L}\right)\|\nabla f(\theta^k)\|_2^2}_{\le 0}$ 我们选的 $\alpha \in \left[0, \frac{2}{c+L}\right]$

$$\le \left(1-\frac{2\alpha cL}{c+L}\right)\|\theta^k-\hat{\theta}\|_2^2$$

Q.E.D

把(△)取对数: $\ln(\|\theta^k-\hat{\theta}\|_2) \le k\cdot \overset{<0}{c_0} + c_1$

Gradient ~~method~~ converges linearly for strongly convex

线性收敛.

是指取对数后是linear

When moving to more general setting:

$$\gamma\|\hat{\Delta}\|_2^2 \le \|X\hat{\Delta}\|_2^2 \le 2\hat{\Delta}^T X^T w$$

如果 $n \ll d$, 这个 γ 就不一定有了.

但能否让 γ 对当前 $\hat{\Delta}$ 存在?

Restricted eigenvalues (RE)

$\Longleftrightarrow$ restricted strong convexity (RSC)

Loss function need not ~~to~~ be curve everywhere, but in small region.

2019.1.14

Recall:

standard linear model

$$\underset{n}{y} = \underset{n\times d}{X}\,\underset{d}{\theta^*} + \underset{n}{w}$$

consider $n >> d$

Least square estimator:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2$$

Interested: $\|\hat{\theta} - \theta^*\|_2$

Crucial step:

Eigen value condition:

$$\|X\Delta\|_2^2 \geq \lambda_{min}(X^TX)\|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{R}^d$$

在之前的假设中，有 $n>>d$ ∴ $\lambda_{min}(X^TX) > 0$，但如果 $n<<d$，$\lambda_{min}=0$，就不能用上一节的方法，需要假设一些结构.

If X has full col rank, then $\lambda_{min}(X^TX) > 0$

Take $\Delta = \hat{\theta} - \theta^*$ we can bound the error

本章研究更 general 的 case.

Today $n << d$ underdetermined

→ need additional constraints on the model.

→ typically, it is assumed θ^* has some low-dim structure.

e.g. sparsity, low-rankness, etc.

Regularization:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta, \{z\}_{i=1}^n) + \lambda R(\theta) \right\}$$

* $z_1, z_2, \cdots, z_n$: samples from space Z

e.g. in the linear model,

$$z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$$

* $\mathcal{L} : \mathbb{R}^d \times Z^n \to \mathbb{R}$: smooth. convex loss function

* $R : \mathbb{R}^d \to \mathbb{R}_+$: norm regularizer

就是线性里面那个 norm.

* $\lambda > 0$: regularization parameter

在线性回归中，$\mathcal{L}(\theta, \{z_i\}) = \|y - X\theta\|_2^2$,

$R(\theta) = \|\theta\|_1$ 这个 $R(\theta)$ 是 convex 的.

在 $n \gg d$ 时，之前那个不等式 $\|X\Delta\|_2^2 \geq |X^TX| \|\Delta\|_2^2$ 成立，条件是 X 列满秩，$\lambda_{min}(X^TX) > 0$.

当 $n \ll d$ 时，能否有一个弱一点的结论，例如：

$$\|X\Delta\|_2^2 \geq c \|\Delta\|_2^2 \quad \forall \Delta \in \mathcal{C}$$

即在某个子域里面成立.

之前是要求对任意Δ成立，现在放松要求，能否对部分Δ成立加regularizer来实现.

下面看为什么加了regularizer就能构造出来这个新不等式.

Decomposability of R

* understand effect of R in inducing the desired structure.

* let $\mathcal{M} \subseteq \overline{\mathcal{M}} \subseteq \mathbb{R}^d$ be supspaces of $\mathbb{R}^d$

  $\overline{\mathcal{M}}$ 不是闭包，只是一个随意的子空间 (θ的空间)

  - $\mathcal{M}$: model subspace. intends to capture the low-dim structure in θ^*

  - $\overline{\mathcal{M}}^\perp$: $\overline{\mathcal{M}}^\perp = \{v \in \mathbb{R}^d : u^\top v = 0 \ \forall u \in \overline{\mathcal{M}}\}$

    $\overline{\mathcal{M}}^\perp$ 叫 perturbation space, capture deviation from the model subspace

    如果参数落到 $\overline{\mathcal{M}}^\perp$ 中，则需要被惩罚

  - for simplicity: take $\mathcal{M} = \overline{\mathcal{M}}$

下面定义 regularizer 的 composability

Def: R is decomposable wrt $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$ if

$$R(\theta+r) = R(\theta)+R(r)$$

$$\forall \theta \in \mathcal{M},\ r \in \bar{\mathcal{M}}^{\perp}$$

注：$R(\theta+r) \le R(\theta)+R(r)$ 是恒成立的，∵R是norm，r是我们不想要的，所以想惩罚到最大，那么取等号的时候，惩罚最大。

r是perturbation
$\bar{\mathcal{M}}$是perturbation space

Rmk: The pair $(\mathcal{M}, \bar{\mathcal{M}})$ can be chosen.

e.g. $\mathcal{M} = \mathbb{R}^d$, $\bar{\mathcal{M}}^{\perp} = \{0\}$

此时 $r \equiv 0$, $R(\theta+r) = R(\theta)+R(r)$ 显然成立

但这个pair用处不大，

what would be a useful choice?

有用的 $(\mathcal{M}, \bar{\mathcal{M}})$ 需满足两点：

(1) $\theta^* \in \mathcal{M}$ / $\Pi_{\mathcal{M}}(\theta^*) \simeq \theta^*$ θ^*在$\mathcal{M}$上的投影很接近自己

(2) $\mathcal{M}$ is "small" 不满足要求的都会被惩罚

e.g. sparse linear model

$$y = X\theta^* + w \qquad \theta^*: s\text{-sparse (has } s \text{ non-0 entries)}$$

$$R(\theta) = \|\theta\|_1$$

For a subset $S \subseteq \{1, \cdots, d\}$ of cardinality s,

$$\mathcal{M}(S) \triangleq \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \notin S\}$$ 只要不在S中的维度都为0

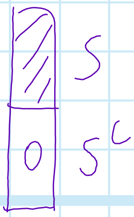
if θ^* is supported on S, $\Pi_{\mathcal{M}(S)}(\theta^*) = \theta^*$

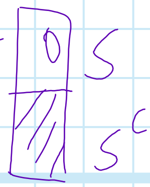
Take $\bar{\mathcal{M}} = \mathcal{M}(S)$, Then: $\bar{\mathcal{M}}^\perp$是在S中的维度为0

$$\bar{\mathcal{M}}^\perp = \{\theta \in \mathbb{R}^d : \theta_j = 0, \forall j \in S\}$$

下面看 decomposibility

(1) $R(\theta) = \|\theta\|_1$

$\theta \in \mathcal{M}$: 

$\gamma \in \bar{\mathcal{M}}^\perp$ 

$R(\theta + \gamma) = R(\theta) + R(\gamma)$ 成立

(2) 另一个例子，X变成了一个矩阵. R用 nuclear norm

$y_i = \langle X_i, \theta^* \rangle + w_i$ 即: $y_i = tr(\underset{m\times d}{X_i^T} \underset{m\times d}{\theta^*}) + w_i$

$$R(\theta) = \|\theta\|_* = \sum_{i=1}^{\min(m,d)} \sigma_i(\theta)$$ rank $r \le \min(m,d)$

$$\sigma_i(\theta) = \sqrt{\lambda_i(\theta^T\theta)}$$

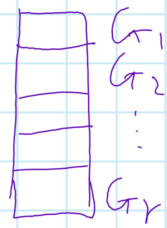
vector 中的 non-zero entry 类似于矩阵 rank

(3) group sparsity

把θ^*切成若干个非相交的组，希望大多数

组是 sparse 的，而组内是否 sparse 无所谓

$$R(\theta)=\sum_{i=1}^{r}\|\theta_{G_i}\|_2$$



1-norm 作用在 group 上，group 内的 norm 无所谓

下面看如何利用 decomposability 来构造 $\|X\Delta\|_2^2 \geq c\|\Delta\|_2^2$ $\forall \Delta \in C_e$ 的形式.

Consequence of Decomposability

Prop1: Suppose $\mathcal{L}$ is smooth convex.

$$\lambda \geq 2R^*\left(\nabla\mathcal{L}(\theta^*;\{z_i\}_{i=1}^{n})\right)$$ （λ 要足够大才有意义）

$R^*(v)$ = dual norm of R 即: $\sup_{R(u)\leq 1} u^T v$

例如: $R(v)=\|\cdot\|_2$，则 $R^*(v)=\|\cdot\|_2$　　$\|\cdot\|_1 \Leftrightarrow \|\cdot\|_\infty$

dual is to minimize a linear function over a unit ball of origional norm.

Then for any $(\mathcal{M},\bar{\mathcal{M}}^\perp)$ over which R is decomposable

$$\hat{\Delta} \triangleq \hat{\theta}-\theta^* \in C_e \triangleq \left\{\Delta\in\mathbb{R}^d : R(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\mathcal{M}^\perp})\right\}$$

where $\Delta_{\bar{\mathcal{M}}}=\Pi_{\bar{\mathcal{M}}}(\Delta)$ and so on.

可以观察到 $\mathcal{C}$ 现在不是整個的了，下面看每一项：

$4R(\theta^*_{M^\perp})$: misspecification of model

∵ 正常情况下 θ^* 在 perturbation space $M^\perp$ 的投影应该为0.

下面用 linear model 举例说明：

取 $d=3$, $\Delta=(\Delta_1, \Delta_2, \Delta_3)$

$R(\Delta)=\|\Delta\|_1$, $S=\{3\}$ 假设参数只在第3维.

$$M(S)=\{\Delta : \Delta_1=\Delta_2=0\}$$

$$M^\perp(S)=\{\Delta : \Delta_3=0\}$$

① 假设 $\theta_1^*=\theta_2^*=0$ (model is exact)

θ^* 就在 M 中，那个 $\mathcal{C}$ 变成：

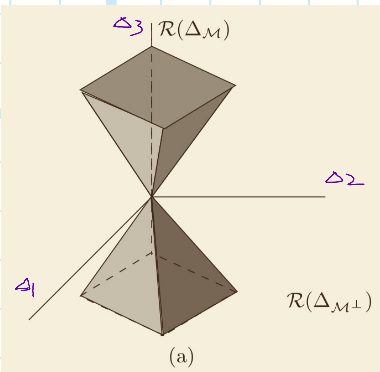
$$\mathcal{C}=\{\Delta : \underbrace{|\Delta_1|+|\Delta_2|}_{R(\Delta_{M^\perp})} \leq \underbrace{3|\Delta_3|}\}$$

$R(\Delta_{M^\perp})$ 把 Δ 在 $M^\perp(S)=\{\Delta : \Delta_3=0\}$ 上投影

把 Δ 在 $M(S)=\{\Delta : \Delta_1=\Delta_2=0\}$ 上投影

画出 $\mathcal{C}$ 的图形：

参数空间变成了一个 cone.

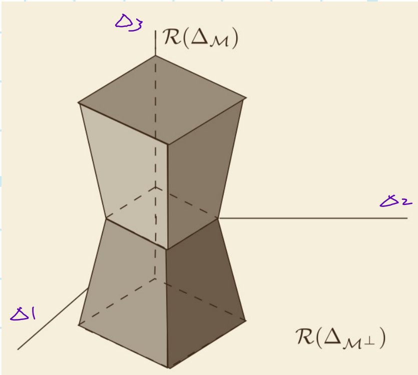


(a)

② 假设 model is not exact ($R(\theta^*_{m^\perp}) \neq 0$)

真实的参数 θ^* 在 perturbation 的空间 $m^\perp$ 上仍有分量. ∴ 当 $|\Delta_3|$ 取 0 时, 右边仍 >0

$$\mathcal{C} = \{\Delta : |\Delta_1| + |\Delta_2| \le 3|\Delta_3| + 4R(\theta^*_{m^\perp})\}$$



(b)

总结:

↙ 0-norm?

nuclear norm 就是 向量的 1-norm 向矩阵 norm的推广, ∵矩阵 1-norm 表征非0元素, 而矩阵的非0 则用 rank 来表示. 可以这么理解: 向量非0元指的是在这个向量上投影, 能剩下多少维, 而对应到矩阵, 则是其 rank 的大小, rank 大, 矩阵变换后剩的多。

Decomposable regularizer 的思路是: 对于 $R(\theta)$, 首先希望 $R(\theta^*) \approx 0$, 而如果在 θ^* 上加一个扰动 γ, 则希望 R 的反应非常剧烈. 即原来的 $R(\theta)$ 加扰动 γ 变成 $R(\theta+\gamma)$ 由三角不等式 $R(\theta+\gamma) - R(\theta) \le R(\gamma)$ 即这个 R 的变化 $R(\theta+\gamma) - R(\theta)$ 有一个上界 $R(\gamma)$, 即然希望越剧烈越好, 就直接取 = 放到最大.

R 是 Decomposable 的一个直接好处是 把 Δ 限定在一个子空间中, 而非全空间. 即是一个 low-dim structure.

2019.1.15

Recamp:

θ^* : ground truth

$z_1, \cdots, z_n$: Samples

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \{ \mathcal{L}(\theta; \{z_i\}_{i=1}^n) + \lambda R(\theta) \}$$

Goal: Bound stat. errors : $\hat{\Delta} = \hat{\theta} - \theta^*$

R: decomposable wrt $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$

decomposability 是关于一对空间的.

$$R(\theta + \gamma) = R(\theta) + R(\gamma) \quad \forall \theta \in \mathcal{M}, \gamma \in \bar{\mathcal{M}}^\perp$$

$\mathcal{M}$: model subspace

$\bar{\mathcal{M}}^\perp$: perturbation subspace

就是说如果Regulizer合适的话, 会被限制到一个子域中,而不是整个空间

prop: Suppose $\mathcal{L}$ is smooth, convex,

$\lambda \geq 2R^*(\nabla \mathcal{L}(\theta^*))$ $\quad R^*$: dual norm

R is decomposable wrt $(\mathcal{M}, \mathcal{M}^\perp)$

Then,

$$\hat{\Delta} \in \mathcal{C}_0 \triangleq \{ \Delta \in \mathbb{R}^d \; R(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\mathcal{M}^\perp}) \}$$

注: 这个结论是 deterministic 的, 即与 noise 无关, 也不是什么分布.

下面证明那个奇怪的定理 C_6

Pf: Define

$$D(\Delta)=\mathcal{L}(\theta^*+\Delta)-\mathcal{L}(\theta^*)+\lambda(R(\theta^*+\Delta)-R(\theta^*))$$

(在 θ^* 上加的一个扰动 Δ)

Obs: $D(\hat{\Delta})\le 0$ ($\because \hat{\theta}$ 是最优解, $\therefore \mathcal{L}(\hat{\theta})+R(\hat{\theta})$ 最小)

先给两个 Claim, 看整体的证明框架.

Claim 1:

$$R(\theta^*+\Delta)-R(\theta^*)\geqslant R(\Delta_{\bar{m}^\perp})-R(\Delta_{\bar{m}})-2R(\theta^*_{m^\perp})$$

Claim 2: under the assumptions of Prop:

$$\mathcal{L}(\theta^*+\Delta)-\mathcal{L}(\theta^*)\geqslant -\frac{\lambda}{2}\left[R(\Delta_{\bar{m}})+R(\Delta_{\bar{m}^\perp})\right]$$

代入 claim 1, claim 2 得:

把最值的不等式拆成两块分别 bound

$$0\geqslant D(\hat{\Delta})$$

$$\geqslant\lambda\left[R(\hat{\Delta}_{\bar{m}^\perp})-R(\hat{\Delta}_{\bar{m}})-2R(\theta^*_{m^\perp})\right]$$

$$-\frac{\lambda}{2}\left[R(\hat{\Delta}_{\bar{m}})+R(\hat{\Delta}_{\bar{m}^\perp})\right]$$

$$=\frac{\lambda}{2}\left[R(\hat{\Delta}_{\bar{m}^\perp})-3R(\hat{\Delta}_{\bar{m}})-4R(\theta^*_{m^\perp})\right]$$

下面来证那两个 claim.

Pf(Claim 1)

$$R(\theta^*+\Delta)=R(\theta^*_m+\theta^*_{m^\perp}+\Delta_{\bar{m}}+\Delta_{\bar{m}^\perp})$$

注意 θ 投影在 m 和 $m^\perp$，而 Δ 投影在 $\bar{m}$ 和 $\bar{m}^\perp$ 上.

$$\geq R(\theta^*_m+\Delta_{\bar{m}^\perp})-R(\theta^*_{m^\perp}+\Delta_{\bar{m}})$$ 三角不等式

$$\geq R(\theta^*_m+\Delta_{\bar{m}^\perp})-R(\theta^*_{m^\perp})-R(\Delta_{\bar{m}})$$

$$=R(\theta^*_m)+R(\Delta_{\bar{m}^\perp})-R(\theta^*_{m^\perp})-R(\Delta_{\bar{m}}) \quad ①$$

注意到：$R(\theta^*)\leq R(\theta^*_m)+R(\theta^*_{m^\perp})$ ②

① − ② ⇒

$$R(\theta^*+\Delta)-R(\theta^*)\geq R(\Delta_{\bar{m}^\perp})-R(\Delta_{\bar{m}})-2R(\theta^*_{m^\perp})$$

Q.E.D.

Pf(Claim 2)

观察到左边是 $\mathcal{L}(\theta^*+\Delta)-\mathcal{L}(\theta^*)$ 我们需要对一个差来找一个下界，自然想到 gradient inequality

By convexity + smoothness,

$$\mathcal{L}(\theta^*+\Delta)-\mathcal{L}(\theta^*) \geq \nabla\mathcal{L}(\theta^*)^T\Delta \geq -|\nabla\mathcal{L}(\theta^*)^T\Delta|$$

$$|\nabla\mathcal{L}(\theta^*)^T\Delta| \leq R^*(\nabla\mathcal{L}(\theta^*))\cdot R(\Delta)$$

(generalized Cauchy-Schwarz)

当 R 是 2-norm 时，就是普通的 C-S

当 R 是 1-norm，R^* 是 ∞-norm，得出 Hölder

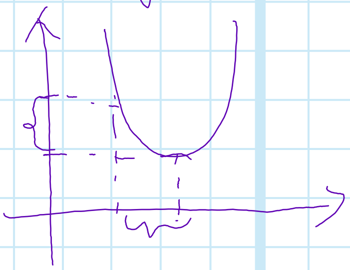
$$\leq \frac{\lambda}{2} R(\Delta)$$

$$\leq \frac{\lambda}{2}\left[R(\Delta_{\overline{\mathcal{M}}})+R(\Delta_{\overline{\mathcal{M}}^\perp})\right] \quad (\text{三角不等式})$$

Q.E.D.

下面看，对于 Δ 的空间 $\Delta \in \mathcal{C}$，如果这个空间有 strong convexity，我们就能得到 error bound

strong convexity 直观理解：enough curved.



如果是 strong convex，我们一定能找到一个兜底的二次曲线，

Definition: $\mathcal{L}$ satisfies the restricted strong convex (RSC) property if $\exists \kappa > 0$, a function $\tau(\cdot)$ s.t.:

↑就是上周那个 c

$$\mathcal{L}(\theta^*+\Delta) \geq \mathcal{L}(\theta^*) + \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \underbrace{\tau^2(\theta^*)}_{\text{tolerance}}$$

$$\forall \Delta \in \mathcal{C}$$

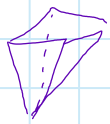
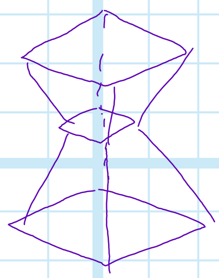
上周讲的 strong convexity △是属于全空间，而且 $\tau^2(\theta^*)=0$，因此这里的这个式子要求弱些

同时，我们这里的 θ^* 是固定的，只能动 Δ

对比正常的 strong convexity：

$$\mathcal{L}(v) \geq \mathcal{L}(u) + \nabla\mathcal{L}(u)^T(v-u) + \kappa\|v-u\|_2^2$$

$$\forall u, v \in \mathbb{R}^d$$

$\tau^2(\theta^*)$ 是 tolerance，把  变成 

直观上讲，我们只需要 $\mathcal{C}$ curve enough，减掉一点不影响。

Thm: Under the assumption of Prop + RSC of L

$$\|\hat{\theta}-\theta^*\|_2^2 \leq \frac{9\lambda^2}{4\kappa^2}\psi^2(\bar{m}) + \frac{1}{\kappa}\left[2\tau^2(\theta^*) + 4\lambda R(\theta^*_{\bar{m}^\perp})\right]$$

where $\psi(m) = \sup_{u\in m\setminus\{0\}} \frac{R(u)}{\|u\|_2}$ 　ψ 是关于空间 m 的一个 const.

ψ is the "Lipschitz const" of R restricted on m

注：这个 bounds 是一个 bounds 族，m 扩大，ψ 扩大.

Rmk:

① Tradoff: Error has 2 parts:

(i) $\frac{9\lambda^2}{4K^2}\psi^2(m)$ 和 (ii) $\frac{4\lambda}{K}R(\theta^*_{m^c})$ (assuming $\tau=0$)

estimation error

模型对 θ^* 估计所产生的误差，

approx error.

由对模型加入各种 specification 所引入的 error，例如假设是 sparse，但实际不是 sparse 的.

"bigger" the m; lower approx error

higher estimation error.

如果对模型的假设少，那么模型空间就很大，estimation error 就大，(没有引入更多外部的信息) 此时 approx error 就小. 没做任何假设.

对 m 限制小，

K 越大，收敛越好.

② For concreate applications, need to compute/bound the parameters

Quick example: Sparse linear regression

θ^*: s-sparse.

$S \subseteq \{1, \dots, d\}$ of cardinality s,

$m(S) = \{\theta : \theta_j = 0 \ \forall j \notin S\}$

$R(\theta) = \|\theta\|_1$

$$\psi(m(S)) = \sup_{u \in m(S)\setminus\{0\}} \frac{\|u\|_1}{\|u\|_2} \overset{\text{C-S}}{=} \sqrt{s}$$

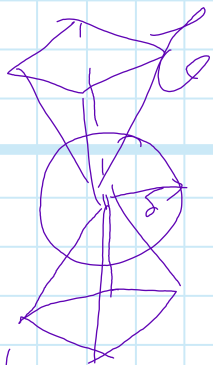
下面证明这个 Theorm

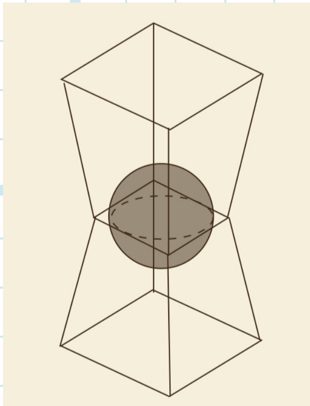
Pf:

$$D(\Delta) = \mathcal{L}(\theta^*+\Delta) - \mathcal{L}(\theta^*) + \lambda[R(\theta^*+\Delta) - R(\theta^*)]$$

$$D(\hat{\Delta}) \le 0$$

Let $\delta > 0$, define: $K(\delta) = \mathcal{C} \cap \{\Delta : \|\Delta\|_2 = \delta\}$





(b)

先给一个claim看证明框架.

Claim: if $D(\Delta) > 0 \ \forall \Delta \in K(\delta)$, then $\|\hat{\Delta}\|_2 \le \delta$

就是说一个 δ 球面上的点都不是 optimal 的，那么 optimal 的点会在 δ 球中.

下面来看claim中的那个条件 $D(\Delta) > 0$

Goal: Show that $D(\Delta) > 0 \ \forall \Delta \in K(\delta)$ hold for some choice of δ

For $\Delta \in K(\delta)$

$$D(\Delta) \ge \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) + \lambda[R(\theta^*+\Delta) - R(\theta^*)] \quad \text{(RSC)}$$

(claim 1)

$$\ge \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) +$$

$$\lambda[R(\Delta_{\bar{m}^\perp})-R(\Delta_{\bar{m}})-2R(\theta^*_{m^\perp})]$$

$$\overset{G\text{-}C\text{-}S}{\geqslant} K\|\Delta\|_2^2-\tau^2(\theta^*)+\lambda[R(\Delta_{\bar{m}^\perp})-R(\Delta_{\bar{m}})-2R(\theta^*_{m^\perp})]$$
$$-R^*(\nabla\mathcal{L}(\theta^*))\cdot R(\Delta)$$

注意到：$R^*(\nabla\mathcal{L}(\theta^*))\leq\frac{\lambda}{2}$

代入：$R(\Delta)\leq R(\Delta_{\bar{m}})+R(\Delta_{\bar{m}^\perp})$

$$\geqslant K\|\Delta\|_2^2-\tau^2(\theta^*)+\frac{\lambda}{2}[\underbrace{R(\Delta_{\bar{m}^\perp})}_{\geqslant 0,\text{扔掉}}-3R(\Delta_{\bar{m}})-4R(\theta^*_{m^\perp})]$$

$$\geqslant K\|\Delta\|_2^2-\tau^2(\theta^*)-\frac{\lambda}{2}[3R(\Delta_{\bar{m}})+4R(\theta^*_{m^\perp})]$$

因为 $R(\Delta_{\bar{m}})\leq\psi(\bar{m})\cdot\|\Delta_{\bar{m}}\|_2$

$$=\psi(\bar{m})\cdot\|\Pi_{\bar{m}}(\Delta)-\Pi_{\bar{m}}(0)\|_2$$
$$\leq\psi(\bar{m})\|\Delta-0\|_2 \quad (\text{non-expansiveness of projection})$$
$$=\psi(\bar{m})\|\Delta\|_2$$

综上有：

$$D(\Delta)\geqslant K\|\Delta\|_2^2-\tau^2(\theta^*)-\frac{\lambda}{2}[3\psi(\bar{m})\cdot\|\Delta\|_2+4R(\theta^*_{m^\perp})]$$

现在要来凑那个claim中的条件 $D(\Delta)>0$

观察到右边是 Δ 的二次函数 $K>0$ ∴ 二次函数开口向上，只要最小点 >0 即可。$-\frac{b}{2a}$ 代入

大和计算能得出那个Thm.

2019.1.21

Recamp:

Regularized loss minimization

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{\mathcal{L}(\theta) + \lambda R(\theta)\}$$

Thm: Suppose $\mathcal{L}$ is smooth and convex, $\lambda \geq 2R^*(\nabla \mathcal{L}(\theta^*))$ R is decomposable wrt $(m, \bar{m}^{\perp})$. Then,

$$\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C} \{\Delta : R(\Delta_{\bar{m}^{\perp}}) \leq 3R(\Delta_{\bar{m}}) + 4R(\theta^*_{m^{\perp}})\}$$

Note: if $\mathcal{L}$ is RSC, i.e.

$$\mathcal{L}(\theta^* + \Delta) \geq \mathcal{L}(\theta^*) + \nabla \mathcal{L}(\theta^*)^T \Delta + K\|\Delta\|_2^2 - \underbrace{\tau^2(\theta^*)}_{\text{tolerance.}} \quad \forall \Delta \in \mathcal{C}$$

(只用对于θ^*成立，不必对于所有θ成立)

Then

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{9\lambda^2}{4K^2} \psi^2(\bar{m}) + \frac{2}{K}\left[\tau^2(\theta^*) + 2\lambda R(\theta^*_{m^{\perp}})\right]$$

where $\psi(\bar{m}) = \sup_{u \in \bar{m} \setminus \{0\}} \frac{R(u)}{\|u\|_2}$

Example 1: LASSO w/ exactly sparse models

$$y = X\theta^* + w$$

assume: θ^* is s-sparse

LASSO: $$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \Big\{ \underbrace{\frac{1}{2n}\|y - X\theta\|_2^2}_{\mathcal{L}(\theta)} + \lambda \underbrace{\|\theta\|_1}_{R(\theta)} \Big\}$$

is $\hat{\theta}$ a good estimator for θ^*?

Take $S = \operatorname{Supp}(\theta^*)$. Define the model subspace as:

$$\mathcal{M} = \mathcal{M}(S) = \bar{\mathcal{M}}(S) = \{\theta : \theta_j = 0 \ \forall j \notin S\}$$

$$\Rightarrow \theta^* \in \mathcal{M} \Rightarrow \theta^*_{\mathcal{M}^\perp} = 0 \qquad \bar{\mathcal{M}}^\perp = \{\theta : \theta_j = 0, \forall j \in S\}$$

$$\mathcal{C} = \{\Delta : \underset{R(\Delta_{\bar{\mathcal{M}}^\perp})}{\underset{\|}{\|\Delta_{S^c}\|_1}} \le 3\|\Delta_S\|_1\}$$

下面验证 RSC:

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \nabla\mathcal{L}(\theta^*)^T\Delta = \frac{1}{2n}\|X\Delta\|_2^2$$

这是把 $\mathcal{L}(\theta^*)$ 在 θ^* 处展开，减掉了常数和一阶项，剩下了二阶项.

$\Rightarrow$ to establish RSC, suffices to show:

$$(*) \quad \frac{\|X\Delta\|_2^2}{2n} \ge \kappa\|\Delta\|_2^2 \quad \forall \Delta \in \mathcal{C}$$

（如果成立的话，这里取 $\tau(\cdot)\equiv 0$）

这里将RSC那个条件转化成了一个类似于特征值不等式的一个式子，只要(*)成立，就是RSC。

下面看(*)的应用。

prop: suppose

(i) X satisfies (*)

(ii) (normalization) $\frac{\|X_j\|_2}{\sqrt{n}} \le 1\ \forall j$, X_j: j^{th} col of X

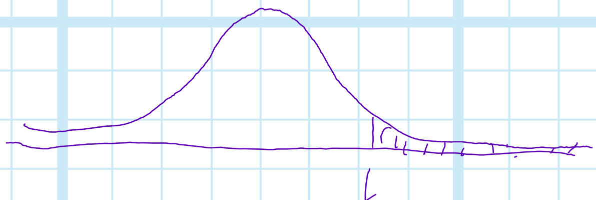
(iii) w is sub-Gaussian w/ parameter $\sigma > 0$

for a fixed $\|v\|=1$ w is mean zero and

$$P_r[|v^T w| \ge t] \le 2\exp\left(\frac{-t^2}{2\sigma^2}\right)\ \forall t.$$

sub-Gaussian:

意思是高斯分布的长尾以一个指数函数为上界。只要长尾以指数分布为上界的就是sub-Gaussian



e.g. $w \sim N(0, I)$ w是高斯，那么w就是sub-Gaussian

Then $g = v^T w = \sum_i v_i w_i \sim N(0, \|v\|_2^2)$

if $\|v\|_2 = 1$, g is standard normal

$$P_r[|g| \ge t] = 2\int_t^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \le 2\exp\left(-\frac{t^2}{2}\right)$$

Then setting $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$, Then

$$\|\theta - \theta^*\|_2^2 \leq \frac{36\sigma^2}{\lambda^2} \frac{s\log d}{n} \quad \text{w/ prob} \geq 1 - \frac{2}{d}$$

σ: power of noise, σ越小, 效果越好.

① $\psi^2(\bar{M}) = \sup_{u \in \bar{M} \setminus \{0\}} \frac{\|u\|_1^2}{\|u\|_2^2} = s$

② $R^*(v) = \|v\|_\infty$; $\nabla\mathcal{L}(\theta^*) = \frac{1}{n} X^T (X\theta^* - y)$

$$= -\frac{1}{n} X^T w$$

开始凑:

Need: $\lambda \geq \frac{2}{n}\|X^T w\|_\infty = \frac{2}{n}\max_i |(X^T w)_i| = \frac{2}{n}\max_i |X_i^T w|$

Note: $P_r\left[\frac{|X_i^T w|}{n} \geq t\right] \leq P_r\left[\left|\left(\frac{X_i}{\sqrt{n}}\right)^T w\right| \geq \sqrt{n}\, t\right]$

$$\leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

之前要求的是 $\|u\| = 1$, 这里 $v = \frac{X_i}{\sqrt{n}}$

$\|u\| \leq 1$ ∴ 也成立. 可以这么看:

$$P_r\left[\left|\frac{\left(\frac{X_i}{\sqrt{n}}\right)^T w}{a}\right| \geq \frac{\sqrt{n}\, t}{a}\right] \leq 2\exp\left(-\frac{nt^2}{a^2 2\sigma^2}\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

$(a \leq 1)$

$$\Rightarrow P_r\left[\frac{\|X^T w\|_\infty}{n} \geq t\right] \leq 2d \cdot \exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

每个维都有个bound, 全部加起来就是d倍的bound.

Set $t^2 = 4\sigma^2 \frac{\log d}{n}$

$$\Rightarrow RHS = 2d \exp(-2\log d)$$

$$= 2/d$$

w/ prob $\geq 1 - \frac{2}{d}$ $\quad \frac{\|X^T w\|_\infty}{n} \leq 2\sigma^2 \sqrt{\frac{\log d}{n}}$

代入 $\lambda \geq \frac{2}{n}\|X^T w\|_\infty$ 得出 $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$ 为什么这么取.

下面看如何证 (*)

Thm: Suppose the rows of X are iid $\sim N(0, \Sigma)$

$\Sigma > 0$, Then,

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}\Delta\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log d}{n}}\|\Delta\|_1$$

$\forall \Delta \in \mathbb{R}^d$ with high probability

where $\rho(\Sigma) = \max_{1\leq j\leq d} \Sigma_{jj}$

下面从这个 Thm 到 (*)

for $\Delta \in \mathcal{C}$ $\qquad$ $\mathcal{C}$ 的约束.

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{s}\|\Delta_S\|_2$$
$$\leq 4\sqrt{s}\|\Delta\|_2$$

For simplicity, take $\Sigma = I$, by Thm,

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Delta\|_2 - 9\sqrt{\frac{\log d}{n}}\left(4\sqrt{s}\|\Delta\|_2\right)$$

$$= \underbrace{\left(\frac{1}{4} - 36\sqrt{\frac{s\log d}{n}}\right)}_{K}\|\Delta\|_2$$

K，这也给出了 n 的下界.

到此，除了那个 Thm，example 1 基本证完

这个 Thm 没去证明，一些观察

Fix Δ:

$$E\left[\|X\Delta\|_2^2\right] = E\left[\Delta^T \underbrace{X^T X}_{\text{Random quadratic form.}} \Delta\right] \qquad x \sim N(0, I)$$

$$\Pr\left(\Delta^T X^T X \Delta - E\left[\Delta^T X^T X \Delta\right] \geq t\right) \leq \cdots$$

这个 concentration 不等式和 Thm 很像，但不能直接得出 Thm，$\because$ 不能直接对所有 $\Delta \in C$ 加起来，因为空间很大.

Example 2: LASSO w/ weakly sparse models.

$y = X\theta^* + w$ 并不是严格 sparse，而是很多接近于 0.

Assume: $\theta^* \in B_q(R_q) = \{\theta : \|\theta\|_q^q \leq R_q\}$ $q \in [0, 1]$

$$\|\theta\|_q^q \le \sum_i |\theta_i|^q$$

$$\sum_i |\theta_i|^q \le R_q$$ impose decay rate on θ

如果 θ_i 大，后面的就小。

下面套用 Thm 在 Example 上的问题。

$\mathcal{C}$ 表示：$\mathcal{C} = \{\Delta : \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1 + \underbrace{4\|\theta^*_{S^c}\|_1}_{\text{多了这项}}\}$

是一个 ball，

而 (*) $\frac{\|X\Delta\|_2^2}{2n} \ge \kappa\|\Delta\|_2^2$ 是齐次的，如果 Δ 在一个 ball 中成立，它在整个空间都成立，这显然是不行的。因此需要借助 $\tau^2(\theta^*)$ 这一项，让它不 $\equiv 0$。

2019.1.22

上节课讲了LASSO的例子，其中关键的一步：

Need $\exists\, K>0$:

$$\frac{\|X\Delta\|_2^2}{n} \geq K\|\Delta\|_2^2 \quad \forall \Delta \in \{\Delta : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$$

Thm: Suppose the rows of X are iid $N(0,I)$

Then whp (with high probability)

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Delta\|_2 - 9\sqrt{\frac{\log d}{n}}\,\|\Delta\|_1 \quad \forall \Delta \in \mathbb{R}^d$$

上节课讲了用这个thm证明上面那个Need.

本节课证明这个thm.

首先给出框架：

WLOG, assume $\|\Delta\|_2=1$ （因为这个不等式是齐次的，所以如果对 $\|\Delta\|_2=1$ 成立，则对所有 Δ 成立）

Step 1: Want LB on

$$\inf_{\|\Delta\|_2=1} \frac{\|X\Delta\|_2}{\sqrt{n}}$$

这里X是随机变量 ∴ 整个式子也是随机变量

For given $r>0$ define

$$V(r) \triangleq \{\Delta \in \mathbb{R}^d : \|\Delta\|_2=1,\ \|\Delta\|_1 \leq r\}$$

proposition 1:

$$\mathbb{E}\left[\inf_{\Delta\in V(r)}\frac{\|X\Delta\|_2}{\sqrt{n}}\right] \geq 3\left[\frac{1}{4}-\sqrt{\frac{\log d}{n}}\,r\right] \quad \text{whenever } V(r)\neq\phi$$

∵对不同的 X, $V(r)$ 不同, ∴ $\mathbb{E}$ 不能直接移到 inf 内

Step 2: Concentration property of

$$Q(r, X)=\inf_{\Delta\in V(r)}\frac{\|X\Delta\|_2}{\sqrt{n}}$$

proposition 2: let $r>0$, be s.t. $V(r)\neq\phi$. Then,

$$\Pr\left[\left|Q(r,X)-\mathbb{E}[Q(r,X)]\right|\geq\frac{1}{2}t(r)\right]\leq 2\exp\left(\frac{-nt^2(r)}{8}\right)$$

where, $t(r)=\frac{1}{4}+3\sqrt{\frac{\log d}{n}}\,r$

这个意思是 $Q(r, X)$ 偏离其中心的概率由一个指数函数控制

Step 3: Prop 1+2 implies w/ prob $\geq 1-2\exp\left(-\frac{nt^2(r)}{8}\right)$

$$Q(r,X)=\inf_{\Delta\in V(r)}\frac{\|X\Delta\|_2}{\sqrt{n}}\geq\mathbb{E}[Q(r,X)]-\frac{1}{2}t(r)$$

$$\geq 3\left[\frac{1}{4}-\sqrt{\frac{\log d}{n}}\,r\right]-\frac{1}{2}\left(\frac{1}{4}+3\sqrt{\frac{\log d}{n}}\,r\right)$$

$$=\frac{5}{8}-\frac{9}{2}\sqrt{\frac{\log d}{n}}\,r$$

Bound for all levels of r

下面来看要用到的工具

Tools

1) Comparison inequality of Gaussian processes.

随机过程视为随机变量+编号

2) Concentration inequality for Lipschitz functions of Gaussian RVs.

在高斯分布上加一个 Lipschitz 函数,

高斯分布的 concentration 性质仍能保留

3) "Pealing argument" from empirical process theorm.

今天 focus 在 proposition 1.

Pf (prop 1)

C-S 不等式

$$\tilde{Q}(r, X) \triangleq \inf_{\Delta \in V(r)} \|X\Delta\|_2 = \inf_{\Delta \in V(r)} \underbrace{\sup_{u : \|u\|_2 = 1}}_{\to u \in S^{n-1}} u^T X \Delta$$

Note: For each $(u, \Delta) \in S^{n-1} \times V(r)$

把每个 (u, Δ) pair 看成 index set

$Y_{u,\Delta} = u^T X \Delta$ (取一个 index (u,Δ) 就有个随机变量 $Y_{u,\Delta}$)

$= \sum_{i,j} u_i \Delta_j X_{ij}$

is a mean-zero Gaussian RV.

want: LB on $\mathbb{E}[\tilde{Q}(r,X)]$

$\Leftrightarrow$ UB on $\mathbb{E}[-\tilde{Q}(r,X)]$

$\Leftrightarrow$ UB on $\mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} (-u^T X \Delta)\right]$

注意到 $u \in S^{n-1}$ $\therefore$ $-u \in S^{n-1}$ 可以去掉负号

$=$ UB on $\mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} \underset{\substack{\| \\ Y_{u,\Delta}}}{u^T X \Delta}\right] \leq ?$

$\{Y_{u,\Delta}\}_{u,\Delta}$ is a GP (这个集合每个元素都是高斯的, 均值为0, 方差各不同)

Idea: construct another GP $\{Z_{u,\Delta}\}_{u,\Delta}$

s.t. (1) $\mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Z_{u,\Delta}\right]$ is "easy" to compute.

(2) related to $\mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Y_{u,\Delta}\right]$

Fact: (Gordon's inequality)

Let U, V be arbitary index sets. Consider $\{Y_{u,v}\}$, $\{Z_{u,z}\}$ families of zero-mean Gaussian RVs. Suppose

$$\sigma(Y_{u,v} - Y_{u',v'}) \leq \sigma(Z_{u,v} - Z_{u',v'})$$

$$\forall (u,v)\ (u',v') \in U \times V$$

$$\sigma(Y_{u,v} - Y_{u,v'}) = \sigma(Z_{u,v} - Z_{u,v'})$$

$$\forall u \in U,\ v, v' \in V$$

Then,

$$\mathbb{E}\left[\sup_{u\in U} \inf_{v\in V} Y_{u,v}\right] \leq \mathbb{E}\left[\sup_{u\in U} \inf_{v\in V} Z_{u,v}\right]$$

Z 更 spread out, 期望越大.

下面看如何来构造 Z.

In our setting,

$$\sigma^2(Y_{u,\sigma} - Y_{u',\sigma'}) = \mathbb{E}\left[(Y_{u,\sigma} - Y_{u',\sigma'})^2\right]$$

$$= \mathbb{E}\left[(u^T X \sigma - u'^T X \sigma')^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i,j} X_{ij}(u_i \sigma_j - u'_i \sigma'_j)\right)^2\right]$$

$\rightarrow$ 展开 $\sum_{i,j} \sum_{k,l} X_{i,j} X_{kl} (u_i \sigma_j - u'_i \sigma'_j)(u_k \sigma_k - u'_l \sigma'_l)$

$$= \sum_{ij} \sum_{kl} \mathbb{E}(X_{ij} X_{kl}) (\cdots)$$

这里若 $(i,j) \neq (k,l)$ 则 $\mathbb{E} = 0$, 否则 $=1$

$$= \sum_{i,j} (u_i \Delta_j - u_i' \Delta_j')^2$$

$$= \| u\Delta^T - u'\Delta'^T \|_F^2 \quad \text{(写成矩阵形式)}$$

$$= \| (u-u')\Delta^T + u'(\Delta - \Delta')^T \|_F^2$$

$$= \|\Delta\|_2^2 \cdot \|u-u'\|_2^2 + \|u'\|_2^2 \|\Delta - \Delta'\|_2^2$$

$$+ 2(u^T u' - \|u'\|_2^2)(\|\Delta\|_2^2 - \Delta^T\Delta')$$

$$\le \|u-u'\|_2^2 + \|\Delta - \Delta'\|_2^2$$

equality holds when $u=u'$ or $\Delta = \Delta'$

$\|\Delta\|_2^2 = 1$, $\|u'\|_2^2 = 1$

$\Delta^T\Delta' \le 1$ (C-S)

$\therefore u^T u' - \|u'\|_2^2 \le 0$

$\|\Delta\|_2^2 - \Delta^T \Delta' \ge 0$.

下面根据这个不等式来凑 Gordon inequality 的那个Z.
希望这个UB是某个RV的 varimle.

This suggests

$$Z_{u,\Delta} = g_1^T u + g_2^T \Delta$$

where g_1 $g_2 \sim \mathcal{N}(0, I)$ and independent.

check:

$$g_1^T(u-u') \sim \mathcal{N}(0, \|u-u'\|_2^2)$$

$$\| \quad \sum_i g_{1,i}(u_i - u_i')$$

$$g_2^T(\Delta-\Delta')\sim N(0,\|\Delta-\Delta'\|_2^2)$$

$$Z_{u,\Delta}-Z_{u',\Delta'}=g_1^T(u-u')+g_2^T(\Delta-\Delta')$$

这样以来 $\|u-u'\|_2^2+\|\Delta-\Delta'\|_2^2=\sigma^2(Z_{u,\Delta}-Z_{u',\Delta'})$

By Gordon's ineq.

$$\mathbb{E}\Big[\sup_{\Delta\in V(r)}\inf_{u\in S^{n-1}}\underbrace{(g_1^Tu+g_2^T\Delta)}_{Z_{u,\Delta}}\Big]$$

这个Z的好处是可以直接拆开了.

$$=\mathbb{E}\Big[\inf_{u\in S^{n-1}}g_1^Tu\Big]+\mathbb{E}\Big[\sup_{\Delta\in V(r)}g_2^T\Delta\Big]$$

$\downarrow$ C-S $\qquad\qquad$ $\downarrow$ Hölder

$$\le\mathbb{E}\big[-\|g_1\|_2\big]+\mathbb{E}\Big[\sup_{\Delta\in V(r)}\|\Delta\|_1\cdot\|g_2\|_\infty\Big]$$

由 $V(r)$ 中 $\|\Delta\|=r$

$$=\mathbb{E}\big[-\|g_1\|_2\big]+r\,\big[\|g_2\|_\infty\big]$$

Claim: $\mathbb{E}\big[\|g_2\|_\infty\big]\le 3\sqrt{\log d}$

$\mathbb{E}\big[\|g_2\|_2\big]\ge\frac{3}{4}\sqrt{n}$ for $n\ge 10$ say.

$$\Rightarrow\mathbb{E}\big[-\tilde{Q}(r,X)\big]\le 3r\sqrt{\log d}-\frac{3}{4}\sqrt{n}$$

$$\Rightarrow\mathbb{E}\big[\tilde{Q}(r,X)\big]\ge\frac{3}{4}\sqrt{n}-3r\sqrt{\log d}$$

这样就证完了 proposition 1

下面来看那个 Claim

$g_2 \sim N(0, I)$ $\quad E[\|g_2\|_\infty] = \int_0^\infty \Pr[\|g_2\| \geq t]\, dt$.

$\left(E(x) = \int_0^\infty \Pr[x > t]\, dt \quad \text{for } x > 0.\right)$

$\Pr[\|g_2\|_\infty > t] \Rightarrow \Pr[|g| > t] \quad g \sim N(0, I)$

存在一维大于 t.

$\Pr[|g| > t] = 2\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

$\leq$

下面看 Gordon 引证等式成立

$(u-u')\Delta^T \cdot (u-u')\Delta^T$

$A \cdot B = tr(AB^T) = tr(A^T B)$

$= tr\left((u-u')\Delta^T \Delta (u-u')^T\right)$

$= \|\Delta\|_2^2 \cdot \|u-u'\|_2^2$

回顾： 2019.1.28

Thm: Suppose $\underset{n<<d}{X\in\mathbb{R}^{n\times d}}$ has iid $N(0,1)$ entries

Then, whp

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Delta\|_2 \underbrace{- 9\sqrt{\frac{\log d}{n}}\|\Delta\|_1}_{\because n<<d \;\therefore 需要这项来修正} \quad \forall \Delta\in\mathbb{R}^d$$

Pf: Define

∵ 那个不等式是齐次的

$$V(r)=\{\Delta\in\mathbb{R}^d : \|\Delta\|_2=1, \|\Delta\|_1\leq r\}$$

为了控制 $Q(r,X)$ 看最小能到多小

$$Q(r,X)=\inf_{\Delta\in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}}$$

Step 1: $E[Q(r,X)] \geq 3\left[\frac{1}{4}-\sqrt{\frac{\log d}{n}}\right] r$

以上是上节课，证明了 mean 有一个下界，本节课讲 Q 大概率集中在其 mean 的周围

Step 2: Concentration

prop: for $r>0$ deviation of Q from its mean

$$P_r\left[\left|Q(r,X)-E[Q(r,X)]\right| \geq \frac{1}{2}t(r)\right] \geq 2\exp\left(-n t^2(r)/8\right),$$

where $t(r)=\frac{1}{4}+3\sqrt{\frac{\log d}{n}}\, r$

就是说 $Q(v,x)$ 离 $E[Q(v,x)]$ 的距离指数衰减.

Pf of prop:

Fact: (Concentration of Measure for Lipschitz Fns of Gaussians)

Let $g \sim \mathcal{N}(0, I_m)$ and $F: \mathbb{R}^m \to \mathbb{R}$ be an L-Lipschitz function i.e.

$$|F(x) - F(y)| \le L \|x - y\|_2 \quad \forall x, y \in \mathbb{R}^m$$

Then, $\forall t \ge 0$,

$$\Pr\Big[\big|F(g) - \mathbb{E}[F(g)]\big| \ge t\Big] \le 2\exp\left(-\frac{t^2}{2L^2}\right)$$

L 越小, 连续性越好, 则这个概率的UB越紧

例: $F(g) = g$

F 不能取二次函数, ∵二次函数不是 L-lipschitz

所以猜到, 这里 Q 应该是 L-Lipschitz 的, 我们再找到那个 L, 就好办了。那我们直接验证 Q 是不是 L-lipschitz 就好了.

Goal: verify $X \longrightarrow Q(r,X)$ is Lipschitz.

$$\sqrt{n}\left[Q(r,X) - Q(r,Y)\right] = \inf_{\Delta\in V(r)} \|X\Delta\|_2 - \inf_{\Delta\in V(r)} \|Y\Delta\|_2$$

$$= \inf_{\Delta\in V(r)} \|X\Delta\|_2 - \|Y\hat{\Delta}\|_2 \quad \text{where } \hat{\Delta} \in \arg\min_{\Delta\in V(r)} \|Y\Delta\|_2$$

(为什么这个minimizer $\hat{\Delta}$ 存在? ∵是最小化一个连续函数在紧集上, 由Weierstrass, 最小值存在)

$$\le \|X\hat{\Delta}\|_2 - \|Y\hat{\Delta}\|_2$$

($\hat{\Delta}$ 比 $\inf_{\Delta\in V(r)} \|X\Delta\|_2$ 的值大)

$$\le \sup_{\Delta\in V(r)} \|(X-Y)\Delta\|_2$$

$$\le \|X-Y\| \cdot \sup_{\Delta\in V(r)} \|\Delta\|_2$$

$\|Au\|_2 \le \|A\| \cdot \|u\|_2$

$\|A\| = \sup_{\|u\|_2=1} \|Au\|_2$

↑ spectral norm of A.

$$= \|X-Y\| \cdot 1$$ ← 由 $V(r)$ 定义.

$$\le \|X-Y\|_F$$

2-norm of vector is similiar to F-norm of a matrix.

综上:

$\sqrt{n}\,|Q(r,X)-Q(r,Y)| \le \|X-Y\|_F$

$\Rightarrow Q : \frac{1}{\sqrt{n}}$ Lipschitz.

spectual norm of vector is similar to $\|u\|_\infty$

$\|u\|_\infty \le \|u\|_2$ 类比.

$\|A\| \le \|A\|_F$

$\Rightarrow$ by fact, set $t = \frac{1}{2}t(r)$

$F(X) = Q(r,X)$

这里 $L = \frac{1}{\sqrt{n}}$ 代入即得:

$$\Pr\left[\,|Q(r,X) - E[Q(r,X)]| \ge \frac{1}{2}t(r)\right] \le 2\exp\left(-n t^2(r)/8\right)$$

以上证明了 step 2, 下面证 step 3.

Step 3: Bound uniformaly over r

For a fixed $r>0$, whp,

$$Q(r,X) \ge E[Q(r,X)] - \frac{1}{2}t(r)$$

$$\ge \frac{5}{8} - \frac{9}{2}\sqrt{\frac{\log d}{n}}\,r$$

Pf of prop:

Lemma1: Let $A \subseteq \mathbb{R}$ be non-empty $f: \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$

$h: \mathbb{R}^d \to \mathbb{R}_+$ are given. Consider r.v. $\quad \sup_{\substack{v\in A\\ h(v)\le r}} f(v,X)$ $\quad$ X is a random vector.

通过控制r来控制这个函数.

Suppose $\exists\, g:\mathbb{R}\to\mathbb{R}$ be non-negative and strictly increasing, s.t. (i) $g(r)\ge\mu\ \ \forall r\ge 0$ and

(ii) $\exists\, c,a>0$, $\forall r>0$, $\quad$ 这里c是常数, a是个函数.

$$(*)\quad \Pr\Big[\sup_{\substack{v\in A\\ h(v)\le r}} f(v,X)\ge g(r)\Big]\le 2\exp\left(-c\cdot a\cdot g^2(r)\right)$$

Then, $$\Pr[\mathcal{E}]\le\frac{2\exp(-4c\cdot a\cdot\mu^2)}{1-\exp(-4\cdot c\cdot a\cdot\mu^2)}$$

where $\mathcal{E}=\{\exists\, v\in A: f(v,X)\ge 2g(h(v))\}$

Then 后面的结论去掉了r, 和r无关。意思是之前对所有r的level都成立, 现在是对所有的, 概率都很小.

下面看这个lemma如何用到Step3上:

In our setting,

$A=\{\delta: \|\delta\|_2=1\}$ $\qquad h(v)=\|v\|$, $\qquad$ 就是把那个v(r)拆开了.

$f(\Delta, x) = 1 - \frac{\|x\Delta\|_2}{\sqrt{n}}$

下面看 g 是什么.

$t(r) = \frac{1}{4} + 3\sqrt{\frac{\log d}{n}}\, r$ $\qquad Q(r,x) \geq \frac{5}{8} - \frac{9}{2}\sqrt{\frac{\log d}{n}}\, r$

$= 1 - \frac{3}{2} t(r).$

下面凑形式来找 $g(r)$

(*) 变成了:

$$\Pr\left[\sup_{\Delta\in V(r)}\left(1 - \frac{\|x\Delta\|_2}{\sqrt{n}}\right) \geq \frac{3}{2} t(r)\right]$$

$$= \Pr\left[-\inf_{\Delta\in V(r)} \frac{\|x\Delta\|_2}{\sqrt{n}} \geq \frac{3}{2} t(r)\right]$$

$$= \Pr\left[Q(r,x) \leq 1 - \frac{3}{2} t(r)\right] \leq 2\exp(-c\cdot a \cdots)$$

$\therefore g(r) = \frac{3}{2}(tr) \geq \frac{3}{8} = \mu.$

有了 g, 下面套入 step 2:

$$\Pr\left[Q(r,x) \leq 1 - \frac{3}{2} t(r)\right] \leq 2\exp(-nt^2(r)/8)$$

$$= 2\exp(-ng^2(r)/18)$$

由 lemma 的结论有:

whp:

$\forall \|\Delta\|_2 = 1$

$$1-\frac{\|X\Delta\|_2}{\sqrt{n}} \le 2\cdot\frac{3}{2}\, t(\|\Delta\|_1)$$

$$= \frac{3}{4} + 9\sqrt{\frac{\log d}{n}}\,\|\Delta\|_1$$

$$\Rightarrow \frac{\|X\Delta\|_2}{\sqrt{n}} \ge \frac{1}{4} - 9\sqrt{\frac{\log d}{n}}\,\|\Delta\|_1$$

这就得出了那个Thm.

下面看那个lemma如何证.

Pf of Lemma 1.

Since $g(r)\ge\mu$ $\forall r\ge 0$, define, for $m=1,2,\dots$

$$A_m=\{v\in A:\ 2^{m-1}\mu\le g(h(v))\le 2^m\mu\}$$

分析中的常用技巧: divide-and-conquer.

Fix X, If $v\in A$ is s.t. $f(v,X)\ge 2g(h(v))$

Then $v\in A_m$ for some m.

$$\Rightarrow \Pr[\mathcal{E}] \le \Pr\Big[\bigcup_{m\ge1}\{\exists v\in A_m \text{ s.t. } f(v,X)\ge 2g(h(v))\}\Big]$$

$$\overset{\text{union bnd}}{\le} \sum \Pr[\exists v\in A_m, \text{ s.t. } f(v,X)\ge 2g(h(v))]$$

Now, $v \in A_m$ and $f(v,x) \geq 2g(h(v)) \Rightarrow$

(1) $f(v,x) \geq 2 \cdot 2^{m-1}\mu = 2^m \mu$

(2) $g(h(v)) \leq 2^m \mu \Rightarrow h(v) \leq g^{-1}(2^m \mu)$

$$\Rightarrow \Pr[\mathcal{E}] \overset{\circledast}{\leq} \sum_{m \geq 1} \Pr\Big[\sup_{\substack{h(v) \leq g^{-1}(2^m\mu) \\ v \in A}} f(v,x) \geq 2^m \mu \Big]$$

$$\leq 2 \sum_{m \geq 1} \exp\Big(-c \cdot a \cdot \big(g(\underbrace{g^{-1}(2^m \mu)}_{r})\big)^2\Big)$$

$$= 2 \sum_{m \geq 1} \exp\left(-c \cdot a\, 2^{2m} \mu^2\right)$$

用几何序列来做UB，然后几何序列求和即可

下节课讲

$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{\mathcal{L}(\theta) + \lambda R(\theta)\}$ 如何算法计算.

2019.1.29

本节讲如何求 $\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{\mathcal{L}(\theta) + \lambda R(\theta)\}$ (P)

Assumptions:

(A1) $\mathcal{L}$ is smooth and convex

(A2) R is a norm with $\lambda = 1$

($\lambda \neq 1$ 可以 scale R 让 $\lambda = 1$)

Obs: (P) is in general non-smooth convex optimization problem.

First-order optimality condition:

$$0 \in \nabla \mathcal{L}(\theta) + \partial R(\theta)$$ (nescessary & sufficient)

where $\partial R(\theta) = \{ s \in \mathbb{R}^d : R(\gamma) \geq R(\theta) + s^T(\gamma - \theta) \ \forall \gamma \}$

($\partial R(\theta)$ 是 sub-gradient 是一个集合,

$\because R(\theta)$ 不一定光滑, 但保证 convex,

$\therefore$ FOC 可以用 sub-gradient 来写)

$\because$ 是 convex $\therefore$ FOC 是充要

Aside: Exercise: when R is a norm,

$$\partial R(\theta)=\{s\in\mathbb{R}^d : R^*(s)\le 1,\ s^T\theta=R(\theta)\}$$

R^* is the dual norm of R

（R是 norm 的话，$\partial R(\theta)$ 可以用其 dual norm 更简单地表示）

注意到这里 FOC 是一个集合包含 0，比等式难解，下面看怎么弄.

Define the proximal map associated with R as:

$$\mathbb{R}^d \ni \underset{\text{给定}\,\theta}{\text{prox}_R(\theta)} = \underset{\gamma\in\mathbb{R}^d}{\arg\min}\Big\{R(\gamma)+\underbrace{\tfrac{1}{2}\|\theta-\gamma\|_2^2}_{\substack{\text{proximity term}\\ \text{希望}\gamma\text{离当前}\theta\text{不要太远}}}\Big\}$$

$R(\gamma)$：非光滑部分

希望找一个 γ 使得 $R(\gamma)$ 足够小，同时离 θ 不太远

Note: $\text{prox}_R(\theta)$ is well-defined (exists & unique)

due to the strongly convexity of

$$\gamma \longmapsto R(\gamma)+\tfrac{1}{2}\|\theta-\gamma\|_2^2$$

一个 strong convex + convex 还是 strong convex

strong convex 不存在一个平面，唯一.

claim: $\hat{\theta}$ is optimal for (P) iff

$$\hat{\theta} = \text{prox}_R\left(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})\right) \quad \text{(fixed-point equation)}$$

于是通过这个map将优化问题转化成了不动点问题

Pf: Consider first-order opt cond of $\text{prox}_R(\cdot)$

写FOC得：

$$0 \in \partial R(\gamma) + \gamma - \theta$$

$$0 \in \partial R(\text{prox}_R(\theta)) + \text{prox}_R(\theta) - \theta \quad \text{for any } \theta$$

Hence, $\hat{\theta}$ is optimal for (P)

$$\Leftrightarrow \quad 0 \in \nabla\mathcal{L}(\hat{\theta}) + \partial R(\hat{\theta})$$

$$\Leftrightarrow \quad 0 \in -(-\nabla\mathcal{L}(\hat{\theta}) + \hat{\theta}) + \hat{\theta} + \partial R(\hat{\theta})$$

$$\hat{\theta} = \text{prox}_R\left(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})\right)$$ 代入验证：

验证：

把$\hat{\theta}$和$\partial R(\hat{\theta})$中的$\hat{\theta}$换掉

$$0 \in -(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})) + \hat{\theta} + \partial R(\hat{\theta})$$

$$0 \in -(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})) + \text{prox}_R(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})) + \partial R(\text{prox}_R(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})))$$

满足上面那个FOC条件

以上证明了那个claim. 下面换个例子.

Example: $R(\gamma) = \mathbb{1}_C(\gamma) = \begin{cases} 0 & \text{if } \gamma \in C \\ +\infty & \text{o/w} \end{cases}$

C is closed convex.

$$\text{prox}_R(\theta) = \arg\min_{\gamma \in \mathbb{R}^d} \left\{ \mathbb{1}_C(\gamma) + \frac{1}{2}\|\theta - \gamma\|_2^2 \right\}$$

γ 肯定是 $\gamma \in C$ ∴

$$= \arg\min_{\gamma \in C} \left\{ \frac{1}{2}\|\theta - \gamma\|_2^2 \right\} = \Pi_C(\theta)$$

∴ 可见 proximal map 是 projection 的一个推广.

下面看这个 fixed-point equation 如何用算法求解.

Algorithm:

Idea: $H(\hat{\theta}) = \hat{\theta}$ 就是直接暴力迭代求不动点.

$$H(\theta^0) \overset{?}{=} \theta^0 \begin{cases} \xrightarrow{Y} \text{done} \\ \xrightarrow{N} \theta^1 = H(\theta^0) \end{cases}$$

问题: (1) 能否收敛

(2) 能否收敛到 fixed point.

Proximal gradient Method (PGM)

$$\theta^{k+1} \leftarrow \mathrm{prox}_{\alpha_k R}\left(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k)\right) \qquad \alpha_k > 0$$

Another interpretation of PGM: 代入定义得:

$$\theta^{k+1} = \underset{\gamma \in \mathbb{R}^d}{\arg\min} \left\{ \alpha_k R(\gamma) + \frac{1}{2} \| \theta^k - \alpha_k \nabla \mathcal{L}(\theta^k) - \gamma \|_2^2 \right\}$$

$$= \underset{\gamma \in \mathbb{R}^d}{\arg\min} \left\{ R(\gamma) + \frac{1}{2\alpha_k} \| \theta^k - \alpha_k \nabla \mathcal{L}(\theta^k) - \gamma \|_2^2 \right\}$$

$$= \underset{\gamma \in \mathbb{R}^d}{\arg\min} \left\{ \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T (\gamma - \theta^k) + \frac{1}{2\alpha_k} \| \gamma - \theta^k \|_2^2 + R(\gamma) + \underbrace{\alpha_k^2 \| \nabla \mathcal{L}(\theta^k) \|_2^2}_{\text{常数, 省略}} \right\}$$

前三项其实就是 $\mathcal{L}(\gamma)$ 的前三阶泰勒展开.

即: $\gamma \longmapsto \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T (\gamma - \theta^k) + \frac{1}{2\alpha_k} \| \gamma - \theta^k \|_2^2$

is a quadratic approx of $\mathcal{L}$ at θ^k.

∴ 也可以这么解释，每一步迭代都是用一个二次函数去估计 smooth 的项.

拓展：也可以用真正的泰勒二阶去估计，这里只是有点像泰勒，但还是有差别

下面来看：(1) 能否收敛 (2) 能否收敛到 fixed-point 之后再研究收敛速度.

加入另一个 Assumption

(A3) $\nabla \mathcal{L}$ is L-Lipschitz

下面套用经典算法分析框架

Proposition: Suppose $\{\alpha_k\}$ satisfies $0<\underline{\alpha}\le\alpha_k\le\bar{\alpha}<\frac{1}{L}$

Then:

(a) (Sufficient Descent) $\exists K_1>0$:

$$F(\theta)=\mathcal{L}(\theta)+\lambda R(\theta)$$

$$F(\theta^k)-F(\theta^{k+1})\geqslant K_1\|\theta^k-\theta^{k+1}\|_2^2$$

(b) (Safeguard) $\exists K_2>0$:

$$\|E(\theta^k)\|_2\le K_2\|\theta^k-\theta^{k+1}\|_2$$

where $E(\theta)=\operatorname{prox}_R(\theta-\nabla\mathcal{L}(\theta))-\theta$ is the residual function

(Rmk: $E(\theta)=0 \iff \theta$ is optimal for (P))

下面看 proposition 的应用.

(a) $\Rightarrow$ $\{F(\theta^k)\}_{k\geq 0}$ monitonically decreasing

since $F(\theta^k) \geq \hat{V} = F(\hat{\theta})$ $\forall k$.

$\Rightarrow$ $F(\theta^k) \longrightarrow V$ (单调有界序列收敛)
i.e. $\{F(\theta^k)\}$ converges.

$\Rightarrow$ $\theta^k - \theta^{k-1} \longrightarrow 0$ (by (a) again)

(b)
$\Rightarrow$ $\|E(\theta^k)\|_2 \longrightarrow 0$

(1) $\Rightarrow$ $V = \hat{V}$

(2) Every accumulation point of $\{\theta^k\}_{k\geq 0}$
is optimal for (P)

$\{\theta^k\}$ 可能不收敛, 但其中的一些
子序列可能收敛, 这些收敛的
点都是 (P) 的最优解
(Subsequential convergence)

下面证明 (a)

Pf (a)

$$\theta^{k+1} = \arg\min\left\{\frac{1}{2}\|\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k) - \gamma\|_2^2 + \alpha_k R(\gamma)\right\}$$

$$\Rightarrow \frac{1}{2}\|\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k) - \theta^{k+1}\|_2^2 + \alpha_k R(\theta^{k+1})$$

$$\leq \frac{1}{2}\alpha_k^2 \|\nabla \mathcal{L}(\theta^k)\|_2^2 + \alpha_k R(\theta^k)$$

$\because \theta^{k+1}$ 是 optimizer

Re arange:

(a.1) $$R(\theta^{k+1}) + \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \frac{1}{2\alpha_k}\|\theta^{k+1} - \theta^k\|_2^2 \leq R(\theta^k)$$

For L:

define $g(t) = \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k))$

$$\mathcal{L}(\theta^{k+1}) - \mathcal{L}(\theta^k) = g(1) - g(0)$$

$$= \int_0^1 g'(t)\,dt$$

$$= \int_0^1 \nabla \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k))^T(\theta^{k+1} - \theta^k)\,dt.$$

我们凑 L-Lipshitz

$$= \int_0^1 [\nabla \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k)) - \nabla \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)]^T(\theta^{k+1} - \theta^k)\,dt$$

先用一道 C-S不等式变式:

$[\nabla \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k)) - \nabla \mathcal{L}(\theta^k)](\theta^{k+1} - \theta^k)$

$\leq \|\nabla \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k)) - \nabla \mathcal{L}(\theta^k)\| \|\theta^{k+1} - \theta^k\|$

然后用 Lipshitz:

$\leq L\|\theta^{k+1} - \theta^k\| \|\theta^{k+1} - \theta^k\|$

$$\leq \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \int_0^1 tL\|\theta^{k+1} - \theta^k\|^2\,dt$$

(a.2) $$= \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|_2^2$$

Hence, $F(\theta^{k+1}) - F(\theta^k) = \mathcal{L}(\theta^{k+1}) + R(\theta^{k+1}) - \mathcal{L}(\theta^k) - R(\theta^k)$

$$\leq \underbrace{-\frac{1}{2}\left(\frac{1}{\alpha_k} - L\right)}_{k_1}\|\theta^{k+1} - \theta^k\|_2^2$$

(a.1) + (a.2) 即得.

以上证明了 (a)

2019.2.11

(a) + (b) $\Rightarrow$

① $F(\theta^k) \to \downarrow$

② Every accumulation point of $\{\theta^k\}$

is an optimal solution

(subsequential convergence)

每个聚点都是一个最优解.

以上是 Recamp.

Pf (Prop 1(b))

Lemma: Let θ, γ be arbitrary. Then,

$0 < \alpha \longmapsto g_1(\alpha) \triangleq \frac{1}{\alpha} \| \text{prox}_{\alpha R}(\theta - \alpha\gamma) - \theta \|_2$

is decreasing

(g_1 是关于 α 的单变量函数, 关于 $\alpha \downarrow$)

and

$0 < \alpha \longmapsto g_2(\alpha) \triangleq \| \text{prox}_{\alpha R}(\theta - \alpha\gamma) - \theta \|_2$

is increasing

先看怎么用这个 lemma,

Assume the lemma:

$$\|\theta^{k+1}-\theta^k\|_2 = \|\mathrm{prox}_{\alpha_k R}(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k)) - \theta^k\|_2$$

(根据 PGM 的定义)
这个第二阶段是 $g_2(\alpha)$
之前的假设 $0 < \underline{\alpha} \le \alpha_k \le \bar{\alpha} < \frac{1}{L}$

$$\ge \|\mathrm{prox}_{\underline{\alpha} R}(\theta^k - \underline{\alpha} \nabla \mathcal{L}(\theta^k)) - \theta^k\|_2$$

下面需要把这个 LHS 和
$\|\mathrm{prox}_R(\theta^k - \nabla \mathcal{L}(\theta^k)) - \theta^k\|_2$ 关联起来,
当 $\alpha \ge 1$ 时直接放
当 $\alpha < 1$ 时, 由 $g_1(\alpha) \downarrow$ ∴

$$\frac{1}{\underline{\alpha}}\|\mathrm{prox}_{\underline{\alpha} R}(\theta^k - \underline{\alpha} \nabla \mathcal{L}(\theta^k)) - \theta^k\|_2 \ge$$

$$\|\mathrm{prox}_R(\theta^k - \nabla \mathcal{L}(\theta^k)) - \theta^k\|_2$$

$$\ge \min\{1, \underline{\alpha}\} \cdot \underbrace{\|\mathrm{prox}_R(\theta^k - \nabla \mathcal{L}(\theta^k)) - \theta^k\|_2}_{E(\theta^k)\text{ 的定义}}$$

以上就证明了 (b) (safeguard).

下面看如何证明这个 Lemma.

Pf of Lemma: Define $h: \mathbb{R}_{++} \times \mathbb{R}^d \longrightarrow \mathbb{R}$

by $h(\alpha, \omega) = \gamma^T(\omega - \theta) + \frac{1}{2\alpha}\|\omega - \theta\|_2^2 + R(\omega)$

(γ, θ 由前面固定, α 是正参数, ω 是向量)

观察到 $h(\alpha, \omega)$ 后两项实际就是那个 minimizer, 只是多了前面一项.

and define Moreau envelope of h by

$$(*) \quad H(\alpha) = \inf_{\omega} h(\alpha, \omega)$$

Moreau envelope 是研究 partial optimization 的一个工具, How minimizer change its behaviour.

minimizer 由一个参数 α 控制

Claim: The opt. soln to $(*)$ given by

$$\omega^* = \operatorname{prox}_{\alpha R}(\theta - \alpha\gamma)$$

Pf: 观察(*)每一项关于 ω 都是 convex 的,

∴写 FO condition for (*):

$$0 \in \gamma + \frac{1}{\alpha}(w - \theta) + \partial R(w)$$

$w^* \triangleq \mathrm{prox}_{\alpha R}(\theta - \alpha\gamma)$ is opt soln to

$$\frac{1}{2}\|\underbrace{\theta - \alpha\gamma}_{\theta} - \underbrace{w}_{\gamma}\|_2^2 + \alpha R(w)$$

对应于prox中的 θ, γ

$$\left(\mathrm{prox}_R(\theta) = \arg\min\left\{\frac{1}{2}\|\theta - \gamma\|_2^2 + R(\gamma)\right\}\right)$$

FO condition:

$$0 \in w - (\theta - \alpha\gamma) + \alpha \cdot \partial R(w)$$

$$= \alpha\gamma + w - \theta + \alpha \cdot \partial R(w)$$

$$\Leftrightarrow 0 \in \gamma + \frac{(w - \theta)}{\alpha} + \partial R(w)$$

This is unique (∵是 strong convex ∴1阶有解,只有唯一的 minimizer)
($w^* = \mathrm{prox}_{\alpha R}(\theta - \alpha\gamma)$)

$$\Rightarrow H(\alpha) = h(\alpha, w^*)$$

Here, heuristically,

$$H'(\alpha) = \frac{\partial h(\alpha, w^*)}{\partial \alpha}$$

$$= -\frac{1}{2\alpha^2} \| w^* - \theta \|_2^2$$

$$= -\frac{1}{2} g_1(\alpha)^2$$

这样就把 $H(\alpha)$ 和 g_1 关联在一起

这里没有 $\frac{\partial w^*}{\partial \alpha}$，只是一个 heuristic 的结论.

$\because w^*$ 也和 α 相关.

同理，

Set $\tilde{H}(\alpha) = H(\frac{1}{\alpha})$

$$\tilde{H}'(\alpha) = -\frac{1}{\alpha^2} H'(\tfrac{1}{\alpha})$$

$$= -\frac{1}{\alpha^2}\left(-\frac{\alpha^2}{2}\right) \| \mathrm{prox}_{\frac{1}{\alpha}R}(\theta - \tfrac{1}{\alpha}\gamma) - \theta \|_2^2$$

$$= \frac{1}{2} g_2(\tfrac{1}{\alpha})^2$$

Note: $\tilde{H}(\alpha) = \inf_{w} h(\frac{1}{\alpha}, w)$

$$= \inf_{w} \left\{ \gamma^T (w - \theta) + \frac{\alpha}{2} \| w - \theta \|_2^2 + R(w) \right\}$$

This is a pointwise minimize of affine function of α, $\therefore \tilde{H}(\alpha)$ is concave function

pointwise of concave is concave

$\Rightarrow \tilde{H}'$ is decreasing (这里变量是concave前面的系数 α)

$\Rightarrow g_2(\alpha) \nearrow$

下面来看 g_1.

Envelope theorem

* deal differentiability of Moreau envelope

* Belongs to Danskin theorem.
(directional differentiability)

大概就是说 $\frac{dw^*}{d\alpha}$ 可以省掉?

下面看那些 accumulation point 能否是同一个点

Prop 2: Let Θ be the opt. soln set of our RLM problem, assumed to be non-empty under previous setting.

(c) (Cost-to-go Estimate)

$$F(\theta^{k+1}) - \hat{v} \le K_3\left[\operatorname{dist}(\theta^k, \Theta)^2 + \|\theta^k - \theta^{k+1}\|_2^2\right]$$

where $\operatorname{dist}(\theta, \Theta) = \inf_{\gamma \in \Theta} \|\theta - \gamma\|_2$ 就是投影

Rmk: dist is well defined

$\because$ Θ is close, convex set under our assumptions

To get convergence rate, we need estimate of $\operatorname{dist}(\theta^k, \Theta)$

Towards that end, consider

(A4) Assumption For any $v \ge \hat{v}$, $\exists\, \mu, \varepsilon$, s.t.

$$\operatorname{dist}(\theta, \Theta) \le \mu \|E(\theta)\|_2 \text{ for any } \theta \text{ satisfying}$$

$$F(\theta) \le v \text{ and } \|E(\theta)\|_2 \le \varepsilon$$

(Local error bound condition)

就是说这个不等式在一个小球中成立（放松了要求）

with (a) (b) (c) + (A4)

$$F(\theta^{k+1}) - \hat{v} \overset{(c, A_4)}{\le} K_4\left[\|E(\theta)\|_2^2 + \|\theta^k - \theta^{k+1}\|_2^2\right]$$

$$\overset{(b)}{\le} K_5 \|\theta^k - \theta^{k+1}\|_2^2$$

$$\overset{(a)}{\le} K_6\left[F(\theta^k) - F(\theta^{k+1})\right]$$

$$= K_6\left[F(\theta^k) - \hat{v} - (F(\theta^{k+1}) - \hat{v})\right]$$ 构造 recurrence

$$\Rightarrow F(\theta^{k+1}) - \hat{V} \le \underbrace{\frac{K_6}{1+K_6}}_{<1} \left(F(\theta^{k}) - \hat{V} \right)$$

Moreover: For the iterate:

by (a).

$$\|\theta^{k} - \theta^{k+1}\|_2^2 \le \frac{1}{K_1} \left(F(\theta^{k}) - \hat{V} \right) \qquad \because F(\theta^{k+1}) \ge \hat{V}$$

by (A4), (b):

$$\operatorname{dis}(\theta^{k}, \Theta) \le K \|\theta^{k} - \theta^{k+1}\|_2$$

$$\le K' \sqrt{F(\theta^{k}) - \hat{V}}$$

这就关联了 value 的收敛速度和 θ 的收敛速度.

至此证明收敛以及速度，还差 (c) 和 (A4) 没证明。

(a) (b) (c) 都是 PGM 的性质，都依赖于 PGM

而 (A4) 是与算法无关的。

2019. 2. 12

Recap.

$$(*) \quad \hat{v} = \min_{\theta} \{ F(\theta) \triangleq \mathcal{L}(\theta) + R(\theta) \}$$

PGM: $\theta^{k+1} \leftarrow \text{prox}_{\alpha_k R}(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k))$

where $\text{prox}_R(\theta) = \arg\min_{\gamma} \{ \frac{1}{2} \|\theta - \gamma\|_2^2 + R(\gamma) \}$

prop: Suppose $0 < \underline{\alpha} \leq \alpha_k \leq \bar{\alpha} < \frac{1}{L}$

(a) (Sufficient Decrease)

$$F(\theta^k) - F(\theta^{k+1}) \geq K_1 \|\theta^k - \theta^{k+1}\|_2^2$$

make it making progress

(b) (Safeguard)

$$\|E(\theta^k)\|_2 \leq K \|\theta^k - \theta^{k+1}\|_2$$

where

$$E(\theta) = \text{prox}_R(\theta - \nabla \mathcal{L}(\theta)) - \theta$$

(c) (Cost-to-Go Estimate)

$$F(\theta^{k+1}) - \hat{v} \leq K_3 [\text{dist}(\theta^k, \Theta)^2 + \|\theta^k - \theta^{k+1}\|_2^2]$$

where Θ is the opt soln set of $(*)$

How far it is from the optimal value

下面先证明 (c)

Pf of (c)

Let $\bar{\theta}^k$ be the projection of θ^k onto Θ (close, convex set)

Then, $dist(\theta^k, \Theta) = \|\theta^k - \bar{\theta}^k\|_2$

We compute:

$$F(\theta^{k+1}) - \hat{v} = F(\theta^{k+1}) - F(\bar{\theta}^k) \quad \leftarrow \because \bar{\theta}^k \in \Theta \text{ 就是 opt soln}$$

$$= \mathcal{L}(\theta^{k+1}) - \mathcal{L}(\bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k)$$

$$= \nabla \mathcal{L}(\hat{\theta}^k)^T(\theta^{k+1} - \bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k)$$

($\because$ mean-value theroam, $\hat{\theta}^k \in [\theta^{k+1}, \bar{\theta}^k]$)
就是某个中值定理

$$= \left.\underbrace{\nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \bar{\theta}^k)}_{\text{加一项减一项}} + R(\theta^{k+1}) - R(\bar{\theta}^k) \right\} \mathrm{I}$$

$$\left. + (\nabla \mathcal{L}(\hat{\theta}^k) - \nabla \mathcal{L}(\theta^k))^T(\theta^{k+1} - \bar{\theta}^k) \right\} \mathrm{II}$$

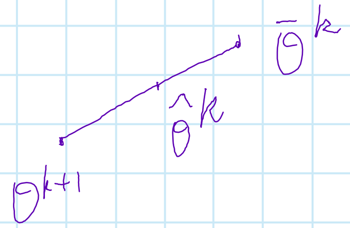
下面分别来 bound I 和 II.

(II): $\mathrm{II} \le \|\nabla \mathcal{L}(\hat{\theta}^k) - \nabla \mathcal{L}(\theta^k)\|_2 \cdot \|\theta^{k+1} - \bar{\theta}^k\|_2$ (C-S)

$$\le L \cdot \|\hat{\theta}^k - \theta^k\|_2 \cdot \|\theta^{k+1} - \bar{\theta}^k\|_2$$

Note: $\|\theta^{k+1} - \bar{\theta}^k\|_2 \le \|\theta^{k+1} - \theta^k\|_2 + \|\theta^k - \bar{\theta}^k\|_2 \leftarrow dist(\theta^k, \Theta)$
三角不等式.

下面看 $\|\hat{\theta}^k-\theta^k\|_2$



$$\|\hat{\theta}^k-\theta^k\|_2 \le \|\theta^{k+1}-\theta^k\|_2+\|\theta^{k+1}-\theta^k\|_2 \quad (三角不等式)$$

$$\le \|\theta^{k+1}-\theta^k\|_2+\|\theta^{k+1}-\bar{\theta}^k\|_2 \quad (由图中的距离知)$$

$$\le L\left[2\|\theta^{k+1}-\theta^k\|_2+\mathrm{dist}(\theta^k,\Theta)\right]\cdot\left[\|\theta^{k+1}-\theta^k\|_2+\mathrm{dist}(\theta^k,\Theta)\right]$$

$$\le 2L\left(\|\theta^{k+1}-\theta^k\|_2+\mathrm{dist}(\theta^k,\Theta)\right]^2$$

$$\le 4L\left[\|\theta^{k+1}-\theta^k\|_2^2+\mathrm{dist}^2(\theta^k,\Theta)\right]$$

$\because (a+b)^2 \le 2a^2+2b^2$

(II): By definition of θ^{k+1};

$$\frac{1}{2}\|\theta^k-\alpha_k\nabla\mathcal{L}(\theta^k)-\theta^{k+1}\|_2+\alpha_k R(\theta^{k+1})$$

(即: θ^{k+1} minimize $\frac{1}{2}\|\theta^k-\alpha_k\nabla\mathcal{L}(\theta^k)-\theta^{k+1}\|_2^2+\alpha_k R(\theta^{k+1})$)

$$\le \frac{1}{2}\|\theta^k-\alpha_k\nabla\mathcal{L}(\theta^k)-\bar{\theta}^k\|_2^2+\alpha_k R(\bar{\theta}^k)$$ (∵ θ^{k+1} 是 optimal，把 θ^{k+1} 换成 $\bar{\theta}^k$ 会变大)

化简后得:

$$R(\theta^{k+1})-R(\bar{\theta}^k)+\nabla\mathcal{L}(\theta^k)^T(\theta^{k+1}-\bar{\theta}^k)$$

$$\le \frac{1}{2\alpha_k}\|\bar{\theta}^k-\theta^k\|_2^2 \le \frac{1}{2\underline{\alpha}}\mathrm{dist}^2(\theta^k,\Theta)$$

∴ (I) 和 (II) 都小于 $K[\mathrm{dist}(\theta^k,\Theta)^2+\|\theta^k-\theta^{k+1}\|_2^2]$

下面来看上节课的 Assumption (A4)

To get convergence rate, we make the following assumption:

(A4) (Local) Error Bound:

For any $v \geq \hat{v}$, $\exists\, \mu, \varepsilon > 0$, s.t.

$$\mathrm{dist}(\theta, \Theta) \leq \mu \| E(\theta) \|_2$$

← 这个就是说我们知道 $E(\theta)=0$ 那么 $\theta \in \Theta$ 就是一个 opt. soln. 那么如果 $E(\theta) \neq 0$，但很小，我们离 opt soln set 是否足够接近。在这里是成立的，但不意味着对所有问题成立

for any Θ:

$$F(\theta) \leq v, \quad \| E(\theta) \|_2 \leq \varepsilon$$

This relates to the geometry of the optimization problem.

和算法性质(例如(a)(b)(c)无关)。

下面看成立的场景

Scenario I:

$\mathcal{L}$ strongly convex, $\nabla \mathcal{L}$ Lipschitz cont.

Strongly convex: $\mathcal{L}(\gamma) \geq \mathcal{L}(\theta) + \nabla \mathcal{L}(\theta)^T (\gamma - \theta) + \frac{K}{2} \| \gamma - \theta \|_2^2$

can find a quadratic support rather than just a linear support.

Fact:

$$(\nabla \mathcal{L}(\gamma) - \nabla \mathcal{L}(\theta))^T (\gamma - \theta) \geq K \| \gamma - \theta \|_2^2$$

wikipedia 上对 strongly convex 的第二种定义

Let $\hat{\theta}$ be the optimal solution to $(*)$. Then,
只有一个.

for any θ:

$$K \cdot \text{dist}(\theta, \Theta)^2 = K\|\theta - \hat{\theta}\|_2^2$$

$$\leq \left(\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\hat{\theta})\right)^T (\theta - \hat{\theta}) \quad \text{(由 strong convex 的 fact)}$$

下面需要和 $E(\theta)$ 来找关联.

By the FO condition, $0 \in \nabla \mathcal{L}(\hat{\theta}) + \partial R(\hat{\theta})$

$$\Rightarrow \boxed{-\nabla \mathcal{L}(\hat{\theta}) \in \partial R(\hat{\theta})} \quad ①$$

On the othe hand.

$$E(\theta) = \text{prox}_R(\theta - \nabla \mathcal{L}(\theta)) - \theta$$

$$\Rightarrow \theta + E(\theta) = \text{prox}_R(\theta - \nabla \mathcal{L}(\theta)) = \arg\min_{\gamma} \{\tfrac{1}{2}\|\theta - \nabla \mathcal{L}(\theta) - \gamma\|_2^2 - R(\gamma)\}$$

using the FO condition for the prox_R:

$$0 \in \theta + E(\theta) - \theta + \nabla \mathcal{L}(\theta) + \partial R(\theta + E(\theta))$$

(代入 prox 的定义, 这里 $\gamma = \theta + E(\theta)$)

$$\Rightarrow \boxed{-[\nabla \mathcal{L}(\theta) + E(\theta)] \in \partial R(\theta + E(\theta))} \quad ②$$

下面来看如何把 sub-diffential 转成不等式.

By def of subdiffential

$$① \Rightarrow R(\theta + E(\theta)) \geq R(\hat{\theta}) - \nabla \mathcal{L}(\hat{\theta})^T (\theta + E(\theta) - \hat{\theta})$$

$$② \Rightarrow R(\hat{\theta}) \geq R(\theta + E(\theta)) - [\nabla \mathcal{L}(\theta) + E(\theta)]^T (\hat{\theta} - (\theta + E(\theta)))$$

加起来 $\Rightarrow$

$$0 \geq \left(\nabla \mathcal{L}(\theta) + E(\theta) - \nabla \mathcal{L}(\hat{\theta})\right)^T \left(\theta + E(\theta) - \hat{\theta}\right)$$

$$\left(\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\hat{\theta})\right)^T (\theta - \hat{\theta}) + \|E(\theta)\|_2^2 \leq$$

$$E(\theta)^T \left(\nabla \mathcal{L}(\hat{\theta}) - \nabla \mathcal{L}(\theta) + \hat{\theta} - \theta\right)$$

$$\leq \|E(\theta)\|_2 \left[\|\nabla \mathcal{L}(\hat{\theta}) - \nabla \mathcal{L}(\theta)\|_2 + \mathrm{dist}(\theta, \Theta)\right]$$

C-S + 三角不等式.

$$\leq \|E(\theta)\|_2 \left(\underbrace{L \cdot \|\hat{\theta} - \theta\|_2}_{\text{Lipschitz 连续}} + \mathrm{dist}(\theta, \Theta)\right]$$

$$\leq (L+1) \cdot \mathrm{dist}(\theta, \Theta) \cdot \|E(\theta)\|_2$$

$$\Rightarrow \mathrm{dist}(\theta, \Theta) \leq \frac{L+1}{\kappa} \|E(\theta)\|_2$$

注: $\frac{L+1}{\kappa}$ 一般表征 condition number of the problem.

L 衡量 smoothness, κ 衡量曲率.

L 越小 κ 越大越好.

Scenario 2:

$\mathcal{L}$ takes the form $\mathcal{L}(\theta) = h(A\theta)$ for some $A \in \mathbb{R}^{m \times d}$. h is strongly convex on compact sets, smooth, (可导) w/ Lipschitz continuous grad. R has polyhedral epigraph.

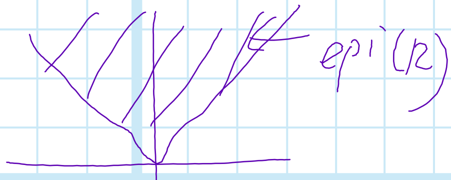
① strongly cvx on compact C $\exists K = K(C) > 0$

$$h(\gamma) \geq h(\theta) + \nabla h^T(\theta)(\gamma - \theta) + \frac{K}{2}\|\gamma - \theta\|_2^2 \quad \forall \gamma, \theta \in C$$

这时候 $\mathcal{L}$ 不一定是 strongly convex 了。∵如果 A 存在 null space，则 $\mathcal{L}$ 上有一个平面等值.

② $\text{epi}(R) = \{(\theta, t) \in \mathbb{R}^d \times \mathbb{R} : R(\theta) \leq t\}$

e.g. $R(\theta) = |\theta|$



polyhedral epigraph: epi(R) is a polyhedron

即 epigraph 是有限个半平面相交而成.

e.g. $R(\theta) = \|\theta\|_1$

$R(\theta) = \|\theta\|_\infty$

Application under Scenario 2:

$R(\theta) = \lambda \|\theta\|_1 \quad \lambda > 0.$

(1) least squares.

$$\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2 = h(X\theta)$$

$$h(u) = \frac{1}{2n}\|y - u\|_2^2$$

$$\nabla^2 h = \frac{1}{2n} I \quad \therefore \text{globally convex}$$

(2) logistic

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T \theta))$$

$= h(A\theta)$

$$h(u) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-b_i u_i))$$

(not convex over the entire space

Hessian 可能趋于0, ∴没有一个全局的 LB, 但如果限制在一个 compact set 上我们能有一个 LB.

2019.2.18

Recall

(Local) Error Bound

$$\hat{v} = \min_{\theta} \{\mathcal{L}(\theta) + R(\theta)\} \qquad (P)$$

Θ : optimal solution set

(EB) For any $v \geq \hat{v}$, $\exists\, \mu, \varepsilon > 0$

$$\operatorname{dist}(\theta, \Theta) \leq \mu \|E(\theta)\|_2 \quad \text{右边比左边好计算}$$

for any $\theta : F(\theta) \leq v$, $\|E(\theta)\|_2 \leq \varepsilon$.

Recall: $E(\theta) = \operatorname{prox}_R(\theta - \nabla \mathcal{L}(\theta)) - \theta$

$E(\theta)$ is the residual measure.

(S1) : $\mathcal{L}$ strongly convex, $\nabla \mathcal{L}$ is Lipschitz continuous $\Rightarrow$ (EB) hold.

(S2) : $\mathcal{L}(\theta) = h(A\theta)$, h: strongly convex on compact sets, A is a linear operator

h: smooth, Lipschitz continuous ∇h.

R has polyhedral epigraph. Θ is compact.

(i.e. $\operatorname{epi}(R) = \{(x,t) : R(x) \leq t\}$ is a polyhedron
(intersection of limited half-space)

e.g. ① LASSO

$$\mathcal{L}(\theta)=\frac{1}{2}\|y-A\theta\|_2^2 \quad R(\theta)=\|\theta\|_1$$

② Logistic regression

$$\mathcal{L}(\theta)=\frac{1}{n}\sum_i \log(1+\exp(-b_i a_i^T\theta))$$

$$R(\theta)=\|\theta\|_1$$

本节课主要来证 (S2):

先给出大概步骤:

Step 1. characterize Θ 先看 opt. soln set 长啥样.

Idea 1: $\Theta=\{\theta: 0\in\nabla\mathcal{L}(\theta)+\partial R(\theta)\}$ (FOC)

$$=\{\theta:\theta=\mathrm{prox}_R(\theta-\nabla\mathcal{L}(\theta))\}$$

This is true but not very useful

We should explore the structure by using the assumptions.

Idea 2:

prop: $\exists\ \bar{y}$ s.t. $\forall\ \hat{\theta}\in\Theta$, $A\hat{\theta}=\bar{y}$ and

$$\nabla\mathcal{L}(\hat{\theta})=A^T\nabla h(\bar{y})\triangleq\bar{g}$$

直观想法: 在之前 strongly convex 的 case 中, 有唯一解保证, 但是这里即使 $h(\cdot)$ 是 strongly convex, 但也不能保证 $h(Ax)$ 是 strongly convex 的。这样 θ 可能抗在一个 flat part 上。那这个 prop 的作用于在于即使 θ 不是唯一的，但 $A\hat{\theta}=\bar{y}$ 的 $\bar{y}$ 却是唯一的。

$$\mathcal{L}(\theta)=h(A\theta) \qquad \nabla\mathcal{L}(\theta)=A^T\nabla h(A\theta)$$

if $\hat{\theta}\in\Theta$: $\nabla\mathcal{L}(\hat{\theta})=A^T\nabla h(A\hat{\theta})=A^T\nabla h(\bar{y})$

$=\bar{g}$ 是唯一的！

In particular,

$$0\in\Theta \Rightarrow 0\in\nabla\mathcal{L}(\hat{\theta})+\partial R(\hat{\theta})$$
$$=\bar{g}+\partial R(\hat{\theta})$$

$$\Leftrightarrow -\bar{g}\in\partial R(\hat{\theta})$$

$\therefore$ 可以写 Θ: Idea 2, Idea 1 FOC

$$\Theta=\{\theta: A\theta=\bar{y},\ -\bar{g}\in\partial R(\theta)\}$$

以下来证明 $\bar{y}$ 的唯一性

Pf: Let $\theta_1,\theta_2\in\Theta$, Set $\bar{y}_1=A\theta_1$, $\bar{y}_2=A\theta_2$

By strong convexity of h,

$$h\left(\frac{\bar{y}_1+\bar{y}_2}{2}\right) < \frac{1}{2}h(\bar{y}_1)+\frac{1}{2}h(\bar{y}_2) \quad (\because \text{strong convex})$$

$$\parallel$$

$$h\left(\frac{A\theta_1+A\theta_2}{2}\right)$$

代入得：

① $\mathcal{L}\left(\frac{\theta_1+\theta_2}{2}\right) < \frac{1}{2}\mathcal{L}(\theta_1) + \frac{1}{2}\mathcal{L}(\theta_2)$

Also, by convexity of R,

② $R\left(\frac{\theta_1+\theta_2}{2}\right) \le \frac{1}{2}R(\theta_1) + \frac{1}{2}R(\theta_2)$

①+② $\Rightarrow$

$$F\left(\frac{\theta_1+\theta_2}{2}\right) < \underbrace{\frac{1}{2}F(\theta_1)}_{=\frac{1}{2}\hat{v}} + \underbrace{\frac{1}{2}F(\theta_2)}_{=\frac{1}{2}\hat{v}} = \hat{v}$$

$F\left(\frac{\theta_1+\theta_2}{2}\right) = \hat{v}$ (∵ θ_1, θ_2 都是 optimal)

∴ $\hat{v} < \hat{v}$ 矛盾, ∴ $\bar{y}_1 = \bar{y}_2$ ∎

In particular,

$\Theta = \Theta_{\mathcal{L}} \cap \Theta_R$ where

$\Theta_L = \{\theta : A\theta = \bar{y}\}$ $\Theta_R = \{\theta : -\bar{g} \in \partial R(\theta)\}$

这样就把 Θ 写成了两个集合的交集

下面看 $\mathrm{dist}(\theta, \Theta)$ 能不能拆. 若是能拆开就简单了.

$$\mathrm{dist}(\theta, \Theta) = \mathrm{dist}(\theta, \Theta_{\mathcal{L}} \cap \Theta_R) \geq \mathrm{dist}(\theta, \Theta_{\mathcal{L}}),\ \mathrm{dist}(\theta, \Theta_R)$$

Question:

dist to intersection $\overset{?}{\approx}$ dist to individual component

Step 2: Relationship between $dist(\theta, \Theta_{\mathcal{L}} \cap \Theta_R)$ and $dist(\theta, \Theta_{\mathcal{L}})$, $dist(\theta, \Theta_R)$

先放结论:

Plan: (1) show $\Theta_{\mathcal{L}}$ Θ_R are polyhedral

(2) $dist(\theta, \Theta_{\mathcal{L}} \cap \Theta_R) \le c\,[dist(\theta, \Theta_{\mathcal{L}}) + dist(\theta, \Theta_R)]$

Note:

$\Theta_{\mathcal{L}}$ is obviously polyhedral since it's solution to linear system.

Θ_R, consider the following facts:

(1) if R has polyhedral epigragh, so does its conjugate $\tilde{R}$, $\tilde{R}(\gamma) = \sup_{\theta}\{\theta^T \gamma - R(\theta)\}$

(2) $\partial \tilde{R} = (\partial R)^{-1}$, i.e.

$$\partial \tilde{R}(\gamma) = (\partial R)^{-1}(\gamma) \triangleq \{\theta : \gamma \in \partial R(\theta)\}$$

(3). If R has polyhedral epigraph and $R(\theta)$ is finite then $\partial R(\theta)$ is a polyhedron.

可以在 convex Wiki: RockFeller 上找

下面来看 Θ_R:

$$\Theta_R \underset{\text{by def'n}}{=} (\partial R)^{-1}(-\bar{g}) \underset{(2)}{=} \partial \tilde{R}(-\bar{g})$$

下面来检查是否 $\tilde{R}(-\bar{g}) \overset{?}{<} \infty$ 只要成立就可以用(3)来说明 $\partial \hat{R}(-\bar{g})$ 是polyhedral.

Note:

$\Theta_R \neq \emptyset$ $\because$ $\hat{\theta} \in \Theta$ will belong to Θ_R

Fact: If R is a norm, then

$$\tilde{R} = \begin{cases} 0 & \text{if } R^*(\gamma) \leq 1 \\ +\infty & \text{o/w} \end{cases}$$

$$\therefore \tilde{R}(-\bar{g}) < \infty$$

综上 Θ_L 和 Θ_R 都是polyhedral.

Next: Estimate point-to-polyhedron dist.

Fact: (Hoffman Error Bound)

Let $P = \{z : Cz \leq d\}$ be a non-empty polyhedron

Then, $\exists\, c > 0$ (which depends only on C) s.t.

$$\text{dist}(x, P) \leq c \|(Cx - d)^+\|_2 \quad \forall x$$

with this fact, we first prove the following

Corollary: Let $\{P_1, \cdots, P_M\}$ be a finite collection of polyhedron. s.t. $P = \bigcap_{i=1}^{M} P_i \neq \emptyset$ Then, $\exists\, \alpha > 0$:

$$\mathrm{dist}(x, P)^2 \le \alpha \sum_{i=1}^{M} \mathrm{dist}(x, P_i)^2 \quad \forall x.$$

i.e. $\{P_1, \dots, P_M\}$ is linearly regular

regular means it's somehow nice, since we can bound the distance by individual dist..

Apply the collary to $\{\Theta_L, \Theta_R\}$:

$$\mathrm{dist}(\theta, \Theta) \le C \cdot \left[\mathrm{dist}(\theta, \Theta_L) + \mathrm{dist}(\theta, \Theta_R) \right]$$

Also, $\mathrm{dist}(\theta, \Theta_L) \le c' \|A\theta - \bar{y}\|_2$ (By Hoffman BD)

目前的问题 ① $c'\|A\theta - \bar{y}\|_2$ 和最终目标 $\|E(\theta)\|_2$ 形式不太一样; ② $\mathrm{dist}(\theta, \Theta_R)$ 怎么用?

Key result to prove the corollary:

Let $H = \{ z : c^T z \le \delta \}$ Then,

$$\mathrm{dist}(x, H) = \frac{(c^T x - \delta)^+}{\|c\|_2}$$

Hint: $\mathrm{dist}(x, H)^2 = \min\{ \|x - z\|_2^2 : c^T z \le \delta \}$

write down KKT.

2019.2.19

Recamp:

$$\hat{v} = \min_{\theta} \{ F(\theta) \triangleq \underset{\displaystyle h(A\theta)}{\underset{\|}{\mathcal{L}(\theta)}} + R(\theta) \}$$

h: str. cvx on compact sets

∇h: Lipschitz cont.

R: norm, polyhedral epigraph

Θ: opt. set, assumed to be compact.

Claim: $\Theta = \Theta_{\mathcal{L}} \cap \Theta_R$ where

$\Theta_{\mathcal{L}} = \{\theta : A\theta = \bar{y}\}$ for some $\bar{y}$

$\Theta_R = \{\theta : -\bar{g} \in \partial R(\theta)\}$; $\bar{g} = A^T \nabla h(\bar{y})$

Claim: $\Theta_{\mathcal{L}}, \Theta_R$ are polyhedral

Corollary: $\text{dist}(\theta, \Theta) = \text{dist}(\theta, \Theta_{\mathcal{L}} \cap \Theta_R)$

$\leq c\left[\text{dist}(\theta, \Theta_{\mathcal{L}}) + \text{dist}(\theta, \Theta_R)\right] \quad \forall \theta$

(linear regularity of $\{\Theta_{\mathcal{L}}, \Theta_R\}$)

这里 Θ_L Θ_R 是 polyhedral

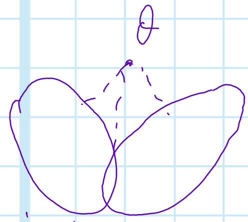
当它们不是 polyhedron 时，如何构造反例？

答：构造一个θ的序列

$$\mathrm{dist}(\theta^k, \Theta_L) \to 0, \quad \mathrm{dist}(\theta^k, \Theta_R) \to 0$$

$$\mathrm{dist}(\theta^k, \Theta) \to 0$$

同时让右边两个趋于0的速度比左边快，这样就没法找到那个c.



这个不行，∵Θ只有一个数.

EXERCISE

By Hoffman Bound

$$\mathrm{dist}(\theta, \Theta_L) \le c' \|A\theta - y\|_2$$

$$\therefore \mathrm{dist}(\theta, \Theta) \le c'\left[\|A\theta - y\|_2 + \mathrm{dist}(\theta, \Theta_R)\right]$$

今天的目标

$$\le \mu \|E(\theta)\|_2$$

$$E(\theta) = \mathrm{prox}_{\theta}(\theta - \nabla \mathcal{L}(\theta)) - \theta$$

Observe: $\Theta_R = (\partial R)^{-1}(-\bar{g})$

Idea: If we view $\theta \in (\partial R)^{-1}(g)$ for some g

then $\mathrm{dist}(\theta, \Theta_R) \approx \mathrm{dist}((\partial R)^{-1}(-g), (\partial R)^{-1}(-\bar{g}))$

$(\partial R)^{-1}$是一个映射，把一个向量映成一个集合.

如果它有某种Lipschitz性质，那么这个距离就可以用g来bound.

$(\partial R)^{-1}(g) \triangleq \{\theta : g \in \partial R(\theta)\}$ 把g映成集合.

下面来定义 $(\partial R)^{-1}$ 的连续性 OLC

Def: Outer Lipschitz continuity of $(\partial R)^{-1}$

We say $(\partial R)^{-1}$ is outer Lipschitz continuous if $\exists \beta > 0$, s.t. for any g', $\exists$ neighborhood $V_{g'}$ of g' s.t.

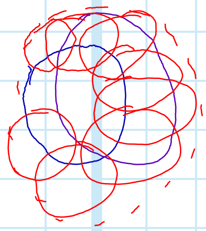
$$(\partial R)^{-1}(g'') \subseteq (\partial R)^{-1}(g') + \beta \|g' - g''\|_2 B(0,1) \quad \forall g'' \in V_{g'}$$

集合的 Lipschitz 连续：

改变输入，会得到一个新集合，新集合比原来的集合只改变了一点。

如何刻化？

把原来的集合的每个点扩大成一个圆，能盖住新的集合。



Outer 的意思就是外面把它包住。

e.g. $R(\theta) = |\theta|$

$$\partial R(\theta) = \begin{cases} \{1\} & \theta > 0 \\ [-1, 1] & \theta = 0 \\ \{-1\} & \theta < 0 \end{cases}$$

For $g' = 1$:

$$(\partial R)^{-1}(1) = \{\theta : 1 \in \partial R(\theta)\} = [0, \infty)$$

For $g' = -1$:

$$(\partial R)^{-1}(-1) = \{\theta : -1 \in \partial R(\theta)\} = (-\infty, 0]$$

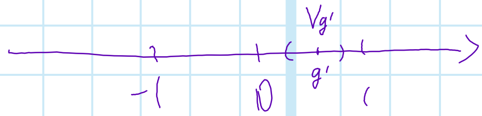
for $g' \in (-1,1)$:

$(\partial R)^{-1}(g') = \{0\}$

下面来看 Outer Lipschitz

take $g' \in (-1,1)$

$\{0\} \subseteq \{0\} + \beta \| g' - g'' \|_2 \, B(0,1) \quad \forall \beta > 0$



take $g' = 1$

$? \subseteq \mathbb{R}_+ + \beta \| g' - g'' \|_2 \cdot B(0,1)$

当 g'' 取到 -1 时 $\mathbb{R}_- \notin \mathbb{R}_+ + \beta \| g' - g'' \|_2 \, B(0,1)$

Fact: If R has polyhedral epigraph then $(\partial R)^{-1}$ is OLC.

$\Rightarrow$

$(\partial R)^{-1}(-g) \subseteq \underset{\uparrow}{(\partial R)^{-1}(-\bar{g})} + \beta \| g - \bar{g} \|_2 B(0,1) \quad \forall -g \in V_{-\bar{g}}$

这个就是 Θ_R

$\Rightarrow$ if $\theta \in (\partial R^{-1})(-g)$ then $\mathrm{dist}(\theta, \Theta_R) \leq \beta \| g - \bar{g} \|_2$

for $\theta \in (\partial R)^{-1}(-g)$, $-g \in V_{-\bar{g}}$

这样就得出了 $\mathrm{dist}(\theta, \Theta_R)$ 的 bound,

$\therefore$ 之前那个不等式可以继续写:

$\leq c' \left[\| A\theta - \bar{y} \|_2 + \mathrm{dist}(\theta, \Theta_R) \right] \quad \forall \theta$

$\le c''\left[\|A\theta-\bar{y}\|_2+\|g-\bar{g}\|_2\right]$ for $\theta\in(\partial R)^{-1}(-g)$, $-g\in V_{-\bar{g}}$

下面研究如何舍弃这两项. 注意到 θ 和 g 有对应关系 $\theta\in(\partial R)^{-1}(-g)$, $-g\in V_{-\bar{g}}$.

Now, consider, $-(\underbrace{\nabla\mathcal{L}(\theta)+E(\theta)}_{\text{"}g\text{"}})\in\partial R(\underbrace{\theta+E(\theta)}_{\text{"}\theta\text{"}})$

b/c $\operatorname{prox}_R(\theta-\nabla\mathcal{L}(\theta))=\arg\min_\gamma\left\{\frac{1}{2}\|\theta-\nabla\mathcal{L}(\theta)-\gamma\|_2^2+R(\gamma)\right\}$

写的 FOL 就是上面那个.

$$0\in\underbrace{\gamma-\theta}_{E(\theta)}+\nabla\mathcal{L}(\theta)+\partial R(\gamma)$$

$$\gamma=\operatorname{prox}_R(\theta-\nabla\mathcal{L}(\theta))$$

$\Rightarrow$ "θ" $\in(\partial R)^{-1}(-$"g"$)$

provided $-(\nabla\mathcal{L}(\theta)+E(\theta))\in V_{-\bar{g}}$, 把新的"$\theta$"和"$g$"代入:

$$\operatorname{dist}(\theta+E(\theta),\Theta)\le c''\left[\|A(\theta+E(\theta))-\bar{y}\|_2+\|\nabla\mathcal{L}(\theta)+E(\theta)-\bar{g}\|_2\right]$$

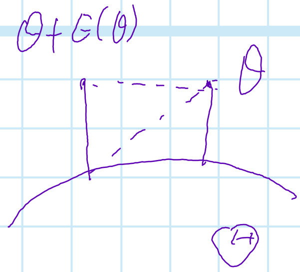
$$\le c''\left[\|A\theta-\bar{y}\|_2+\|\nabla\mathcal{L}(\theta)-\bar{g}\|_2+(\|A\|+1)\|E(\theta)\|_2\right]$$

$$\le c'''\left[\|A\theta-\bar{y}\|_2+\|E(\theta)\|_2\right]$$

$$\|\nabla\mathcal{L}(\theta)-\bar{g}\|_2=\|A^T\nabla h(A\theta)-A^T\nabla h(\bar{y})\|_2$$
$$\le\|A\|\cdot\|\nabla h(A\theta)-\nabla h(\bar{y})\|_2$$
$$\le L\cdot\|A\|\cdot\|A\theta-\bar{y}\|_2$$

$\therefore\ \operatorname{dist}(\theta,\Theta)\le\operatorname{dist}(\theta+E(\theta),\Theta)+\|E(\theta)\|_2$

$$\le\tau\left[\|A\theta-\bar{y}\|_2+\|E(\theta)\|_2\right]$$



Now, by strong cvx of h,

$\bar{y} = A\hat{\theta}$, $\hat{\theta} = \pi_{\Theta}(\theta)$

$$K \|A\theta - \bar{y}\|_2^2 \le \left(\nabla h(A\theta) - \nabla h(\bar{y})\right)^T (A\theta - \bar{y})$$

注: $h(y) \ge h(x) + \nabla h(x)^T (y-x) + \frac{K}{2}\|y-x\|_2^2$ (str cvx 定义)

$h(x) \ge h(y) + \nabla h(y)^T (\cancel{y} x - y) + \frac{K}{2}\|x-y\|_2^2$

加起来就得到上面那个.

$$= \left(A^T(\nabla h(A\theta) - \nabla h(\bar{y}))\right)^T (\theta - \hat{\theta})$$

$$= (\nabla \mathcal{L}(\theta) - \bar{g})^T (\theta - \hat{\theta})$$

Hence: $(a+b)^2 \le 2a^2 + 2b^2$

$$\mathrm{dist}^2(\theta, \Theta) \le \tau' \left[(\nabla \mathcal{L}(\theta) - \bar{g})^T (\theta - \hat{\theta}) + \|E(\theta)\|_2^2 \right]$$

$$\overset{\text{claim}}{\le} \tau'' \left[\mathrm{dist}(\theta, \Theta) \cdot \|E(\theta)\|_2 \right]$$

$$\mathrm{dist}(\theta, \Theta) \le \tau'' \left[\|E(\theta)\|_2 \right]$$

☑ 证毕.

下面来verify这个claim

$$-\nabla \mathcal{L}(\hat{\theta}) \in \partial R(\hat{\theta})$$

$$-(\nabla \mathcal{L}(\theta) + E(\theta)) \in \partial R(\theta + E(\theta))$$

$$\Leftrightarrow R(\theta + E(\theta)) \ge R(\hat{\theta}) - \nabla \mathcal{L}(\hat{\theta})^T (\theta + E(\theta) - \hat{\theta})$$

$$R(\hat{\theta}) \ge R(\theta + E(\theta)) - (\nabla \mathcal{L}(\theta) + E(\theta))^T (\hat{\theta} - \theta - E(\theta))$$

以上推记完了 Scenario 2.

还应看这种 Error Bound 分析能否扩展到 second order 的算法中,

2019.2.25

回顾：之前讲了各种一阶方法.

$$\min_{\theta \in \mathbb{R}^d} \{ F(\theta) = \mathcal{L}(\theta) + R(\theta) \} \quad (P)$$

PGM: $\theta^{k+1} \leftarrow \mathrm{Prox}_{\alpha_k R}(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k))$

$$\mathrm{prox}_R(\theta) = \arg\min_{\gamma} \{ \tfrac{1}{2} \|\theta - \gamma\|_2^2 + R(\gamma) \}$$

We only use the first order information

这里 $\mathrm{prox}_R(\theta)$ 也是一个优化问题，如果有解析解则能加快计算，如果没有：要用两重循环.

2 loop:

outer: prox-update

inner: computing the prox.

今天来讲 second order methods.

Second order method

- Assume f is smooth. Consider
$$\inf_x f(x)$$

classic Newton's method: 回顾牛顿法.

In each iteration, solve a local quadratic model of f.

i.e. At x^k: 用泰勒展开.

$$d^k = \arg\min_d \{ f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T \nabla^2 f(x^k) d \}$$

and set next iterate to be

$$x^{k+1} \leftarrow x^k + \alpha_k d^k, \quad \alpha_k > 0 \text{ step size}$$

Need d^k to be well-defined.

If f is strongly convex, then d^k is well-defined in each iteration (unique)

$\because$ if strongly cvx, $\nabla^2 f(x^k)$ 是 PD $\therefore$ 最值唯一

- SOM for (P) SOM 会有如下几个问题.
  - F is not twice differentiable. (因为有R)
    能不能设计一个sub-gradient 的二阶扩展?
  - Is SOM reasonable?

二阶的优势在哪？

收敛更快，之前一阶的收敛速度是 c^k，
二阶是 2^{c^k} ∴可能外层循环更少但
内部循环变多.

. fewer outer iterations
more expensive inner iterations.

Issue: Non-differentiability of F

Recall: PGM,

PGM的两种解释:

① 是一个求不动点的过程.

② $$\theta^{k+1} \leftarrow \underset{\gamma\in\mathbb{R}^d}{\mathrm{argmin}} \left\{ \mathcal{L}(\theta^k) + \nabla\mathcal{L}(\theta^k)^T(\gamma-\theta^k) + \overbrace{\frac{1}{2\alpha_k}\|\gamma-\theta^k\|_2^2}^{\frac{1}{2\alpha_k}(\gamma-\theta^k)^T I(\gamma-\theta^k)} + R(\gamma) \right\}$$

从②中可见PGM把其分析为两块，前面可以计算，算不出的R直接加.

To Generalize, consider

$$(*) \quad \theta^{k+1} \leftarrow \underset{\gamma\in\mathbb{R}^d}{\mathrm{argmin}} \left\{ \mathcal{L}(\theta^k) + \nabla\mathcal{L}(\theta^k)^T(\gamma-\theta^k) + \frac{1}{2\alpha_k}(\gamma-\theta^k)^T H_k(\gamma-\theta^k) + R(\gamma) \right\}$$

where H_k is an (approximate) Hessian at θ^k

Example of H_k:

① $H_k = I$ (PGM) 当 $H_k=I$，退化为 PGM

② $H_k = \nabla^2\mathcal{L}(\theta^k)$

③ H_k = quasi-Newton strategies.

例如对 $\nabla^2 \mathcal{L}(\theta^k)$ 的 low-rank approx

也可以是 subsample of $\nabla^2 \mathcal{L}(\theta^k)$.

BFGS

以上通过把 PGM 变形进行扩展，使能包含二阶信息

For simplicity, Consider $H_k = \nabla^2 \mathcal{L}(\theta^k)$

How to solve (*)?

- Do we need exact soln?

  if no, how exact we need?

- Is θ^{k+1} well-defined.

  例如 $\mathcal{L}(\theta) = \|y - \overset{\text{fat}}{A}\theta\|_2^2$，∴ Hessian $A^T A$

  一般都是 singular. ∴ 有可能不唯一，

  ∴ 不是 well-defined

  和我们前面的 regularizer 之前那部分为了引入 structure 理解为让问题变好

To fix this, use regularization:

$$\text{use } H_k = \underbrace{\underbrace{\nabla^2 \mathcal{L}(\theta^k)}_{\geq 0} + \mu_k I}_{> 0}, \quad \mu_k > 0$$

observe, $H_k \succ 0$, 这样,原问题就 strongly cvx

下面问题就变成了:

$$\theta^{k+1} \leftarrow \underset{\gamma \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T (\gamma - \theta^k) + \frac{1}{2\alpha_k} (\gamma - \theta^k)^T H_k (\gamma - \theta^k) + R(\gamma) \right\}$$

where $H_k = \nabla^2 \mathcal{L}(\theta^k) + \mu_k I$, $\mu_k > 0$.

Issues to address:

- Choice of α_k, μ_k
- Inexactness criterion for the sub-problem.

prop1: Define

$$\ell_k(\theta) = \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T (\theta - \theta^k) + R(\theta)$$

$$q_k(\theta) = \ell_k(\theta) + \frac{1}{2} (\theta - \theta^k)^T H_k (\theta - \theta^k)$$

Suppose θ^k is not optimal. Then,

$$q_k(\theta^k + \Delta^k) - q_k(\theta^k) \le \frac{1}{2} \left(\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k) \right)$$

where $\Delta^k = \underset{\Delta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T \Delta + \frac{1}{2} \Delta^T H_k \Delta + R(\theta^k + \Delta) \right\}$

这个 prop 给出了二阶的变化和一阶变化的关系.

prop2: For any $k \geq 0$ and $\beta \in (0,1)$, we have

$$F(\theta^k) - F(\theta^k + \alpha \Delta^k) \geq \beta\left[\ell_k(\theta^k) - \ell_k(\theta^k + \alpha\Delta^k)\right]$$

whenever $\alpha \in (0, \bar{\alpha}]$ for some $\bar{\alpha} > 0$.

Next Task: ① argue $\{F(\theta^k)\}$ converges.

② 由 prop 2, $\ell_k(\theta^k) - \ell_k(\theta^k + \alpha_k \Delta^k) \to 0$.

③ use prop 1 and defn of H_k to argue $\varepsilon(\theta^k) \to 0$, for some choice of μ_k

(借助 prop1, 关于 ℓ_k 和 q_k 的定义)

下面先证两个 prop.

Pf of Prop 1:

Observe, $\theta^k + \Delta^k$ minimize q_k. $\Rightarrow$

$$0 \in \partial q_k(\theta^k + \Delta^k)$$

$$\partial q_k(\theta) = \nabla \mathcal{L}(\theta^k) + H_k(\theta - \theta^k) + \partial h(\theta)$$

$$= \partial \ell_k(\theta) + H_k(\theta - \theta^k)$$

$$0 \in \partial q_k(\theta^k + \Delta^k) \iff$$

$$-H_k \Delta^k \in \partial \ell_k(\theta^k + \Delta^k)$$

$\Rightarrow \Delta^k \neq 0$, for o/w θ^k is optimal,
(check the FOC) 和假设矛盾.

Now, 用 sub-differential 的定义来写:

$$\ell_k(?) \geq \ell_k(\theta^k+\Delta^k) + (-H_k\Delta^k)^T(? - \theta^k - \Delta^k)$$

代入 $? = \theta^k \Rightarrow$

$$\ell_k(\theta^k) \geq \ell_k(\theta^k+\Delta^k) + (-H_k\Delta^k)^T(-\Delta^k)$$
$$= \ell_k(\theta^k+\Delta^k) + (\Delta^k)^T H_k \Delta^k$$
$$> \ell_k(\theta^k+\Delta^k) \quad (\because H_k \succ 0)$$

$$\Rightarrow q_k(\theta^k+\Delta^k) - q_k(\theta^k) = \ell_k(\theta^k+\Delta^k) + \frac{1}{2}(\Delta^k)^T H_k \Delta^k - \ell_k(\theta^k)$$
$$\leq \frac{1}{2}\left[\ell_k(\theta^k+\Delta^k) - \ell_k(\theta^k)\right]$$

2019.2.26

Recall

$$\min_{\theta \in \mathbb{R}^d} \{ F(\theta) \triangleq \mathcal{L}(\theta) + R(\theta) \}$$

SQM:

① Find $\Delta^k = \underset{\Delta}{\operatorname{argmin}} \{ \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T \Delta + \frac{1}{2} \Delta^T H_k \Delta + R(\theta^k + \Delta) \}$

where $H_k = \nabla^2 \mathcal{L}(\theta^k) + \mu_k I$, $\mu_k \geq 0$

② Find step size α s.t.

(*) $F(\theta^k) - F(\theta^k + \alpha \Delta^k) \geq \beta [l_k(\theta^k) - l_k(\theta^k + \alpha \Delta^k)]$

③ update $\theta^{k+1} = \theta^k + \alpha \Delta^k$

$$l_k(\theta) = \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)(\theta - \theta^k) + R(\theta)$$

$$q_k(\theta) = l_k(\theta) + \frac{1}{2}(\theta - \theta^k)^T H_k (\theta - \theta^k)$$

$$\Rightarrow \theta^k + \Delta^k \text{ minimize } q_k \iff 0 \in \underset{\underset{\partial l_k(\theta^k+\Delta^k) + H_k \Delta^k}{\parallel}}{\partial q_k(\theta^k + \Delta^k)}$$

Preliminary Results

① $\ell_k(\theta^k) \geq \ell_k(\theta^k + \Delta^k) + (\Delta^k)^T H_k (\Delta^k)$

② $q_k(\theta^k + \Delta^k) - q_k(\theta^k) \leq \frac{1}{2}\left[\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k)\right]$

prop 2: for any $\beta \in (0,1)$, $(*)$ holds whenever

$\alpha \in (0, \bar{\alpha}]$, $\bar{\alpha} = \min\left\{1, \frac{2\mu_k(1-\beta)}{L_G}\right\}$

L_G: Lipschtz const. for $\nabla \mathcal{L}$

Pf: By prelim result (1):

drop the ∇^2

$$\ell_k(\theta^k) - \ell(\theta^k + \Delta^k) \geq (\Delta^k)^T H_k \Delta^k \geq \mu_k \|\Delta^k\|_2^2 \quad (\Delta)$$

on the other hand,

利用 ℓ_k 的 cvx 性质.

$$\ell_k(\theta^k + \alpha\Delta^k) = \ell_k\left((1-\alpha)\theta^k + \alpha(\theta^k + \Delta^k)\right) \quad \alpha \in [0,1]$$

$$\leq (1-\alpha)\ell_k(\theta^k) + \alpha\, \ell_k(\Delta^k + \Delta^k)$$

$$\Rightarrow \ell_k(\theta^k) - \ell_k(\theta^k + \alpha\Delta^k) \geq \alpha\left[\ell_k(\theta^k) - \ell_k(\theta^k + \Delta^k)\right]$$

$$\geq \alpha \mu_k \|\Delta^k\|_2^2 \quad (\text{由}(\Delta))$$

利用 $\nabla\mathcal{L}$ Lipschitz cont.

$$\Rightarrow \mathcal{L}(\theta^k + \alpha\Delta^k) - \mathcal{L}(\theta^k) \leq \alpha \nabla\mathcal{L}(\theta^k)^T \Delta^k + \frac{\alpha^2 L_G}{2}\|\Delta^k\|_2^2$$

(这个结论不要求 $\mathcal{L}$ 是 convex)

记: $g(\alpha) = \mathcal{L}(\theta^k + \alpha\Delta^k)$

$\therefore \mathcal{L}(\theta^k + \alpha\Delta^k) - \mathcal{L}(\theta^k) = g(\alpha) - g(0)$

$= \int_0^\alpha g'(t)\,dt.$

Hence,

$$\underbrace{F(\theta^k)}_{\mathcal{L}(\theta^k)+R(\theta^k)} - \underbrace{F(\theta^k + \alpha\Delta^k)}_{\left(\mathcal{L}(\theta^k+\alpha\Delta^k)+R(\theta^k+\alpha\Delta^k)\right)}$$

$$\geq \ell_k(\theta^k) - \ell_k(\theta^k + \alpha\Delta^k) - \frac{\alpha^2 L_G}{2}\|\Delta^k\|_2^2$$

$\Rightarrow$ for any $\beta \in (0,1)$:

$$F(\theta^k) - F(\theta^k + \alpha\Delta^k) - \beta\left[\ell_k(\theta^k) - \ell_k(\theta^k + \alpha\Delta^k)\right]$$

$$\geq (1-\beta)\left[\ell_k(\theta^k) - \ell_k(\theta^k + \alpha\Delta^k)\right] - \frac{\alpha^2 L_G}{2}\|\Delta^k\|_2^2$$

$$\geq \left[(1-\beta)\alpha u_k - \frac{\alpha^2 L_G}{2}\right]\|\Delta^k\|_2^2$$

我们希望这项 ≥ 0 　就能解出那个 α

Convergence.

以上证明了 F 会单调下降，但是能否收敛到最优值?

① $E(\theta) = \operatorname{prox}_R(\theta - \nabla \mathcal{L}(\theta)) - \theta$

for the orig. problem.

② Note: 在我们的新问题中:

$$q_k(\theta) = \underbrace{\ell(\theta^k) + \nabla \ell(\theta^k)^T(\theta - \theta^k) + \frac{1}{2}(\theta-\theta^k)^T H_k(\theta - \theta^k)}_{\tilde{\mathcal{L}}(\theta)} + R(\theta)$$

$\tilde{\mathcal{L}}(\theta)$ strongly cvx.

$$\Rightarrow E_k(\theta) = \operatorname{prox}_R(\theta - \nabla \tilde{\mathcal{L}}(\theta)) - \theta$$

$$= \operatorname{prox}_R(\theta - \nabla \mathcal{L}(\theta^k) - H_k(\theta - \theta^k)) - \theta$$

$$= \operatorname{prox}_R((I - H_k)\theta - (\nabla \mathcal{L}(\theta^k) - H_k\theta^k)) - \theta$$

FACT: $E(\theta^k) = E_k(\theta^k)$

(*) $\Rightarrow \{F(\theta^k)\}$ monotonic decreasing $\Rightarrow F(\theta^k) \to \bar{v}$

Recall: $\ell_k(\theta^k) - \ell_k(\theta^k + \alpha_k \Delta^k) \geq \alpha_k \mu_k \|\Delta^k\|_2^2 \geq 0$

$$\Rightarrow \lim_{k \to \infty} \alpha_k \mu_k \|\Delta^k\|_2^2 = 0$$

我们希望能由此得出 $\Delta^k = 0$, 这样 θ^k 就是 opt.

Observe:

$$\|E(\theta^k)\|_2 = \|E_k(\theta^k) - E_k(\theta^k+\Delta^k)\|_2$$

$\because$ $\theta^k+\delta^k$ 就是当前问题的解 $\therefore$ 带入相当于减了个0

$$\le (L_G+\mu_k+2)\|\Delta^k\|_2 \quad (\text{由 } E_k \text{ 的 Lipschitz})$$

$\|prox_R(x) - prox_R(y)\|_2 \le \|x-y\|_2$

$\because$ prox 是 non-expensive operator.

类似的, projection 也有类似的性质, 经过 projection 后不会增长,

non-expensiveness of $prox_R(\cdot)$.

$\Rightarrow$

$$0 \le \frac{2\mu_k^2(1-\beta)\|E(\theta^k)\|_2^2}{L_G(L_G+\mu_k+2)^2} \le \alpha_k\mu_k\|\Delta^k\|_2^2$$

$\Rightarrow$

$$\lim_{k\to\infty} \frac{\mu_k^2\|E(\theta^k)\|_2^2}{(L_G+\mu_k+2)^2} = 0 \quad \text{两边夹逼}.$$

μ_k 是我们选的, 有很多选法, 如果选 μ_k 是个常数, 那么 $\|E(\theta^k)\|_2^2=0$, 这意味着, 尽管我们一直用一个有误差的 μ_k, 还是可以?

也可以 choose $\mu_k = c\cdot\|E(\theta^k)\|^\rho$, $\rho\ge 0$

$\Rightarrow \|E(\theta^k)\|_2^2$ goes to 0

Set $\mu_k = c\cdot\|E(\theta^k)\|_2^\rho$, $\rho\ge 0$, $c>0$.

以上证明了 convergence, 下面来看 convergence rate.

是不是把 ρ 取得越大, 收敛越快?

prop3: Under the Error Bound assumption, the above setting μ_k, then for large k, we can take $\alpha_k = 1$

(EB): $\mathrm{dist}(\theta, \Theta) \le \mu \|E(\theta)\|_2 \quad \forall \theta : \|E(\theta)\|_2 \le \varepsilon, \ldots$

μ_k: $\mu_k = C \cdot \|E(\theta^k)\|_2^{\rho}, \quad \rho \ge 0, \; C > 0$

Sketch:

Key inequality: $F(\theta^k + \Delta^k) - F(\theta^k) \le \frac{L_H}{6}\|\Delta^k\|_2^3 - \frac{\mu_k}{2}\|\Delta^k\|_2^2 + \frac{1}{2}\left[\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k)\right]$

L_H: Lip const. of $\nabla^2 \mathcal{L}$

$$\le \beta\left[\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k)\right] + \left(\frac{1}{2} - \beta\right)\underbrace{\left[\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k)\right]}_{\le -\mu_k \|\Delta^k\|_2^2} + \frac{L_H}{6}\|\Delta^k\|_2^3 - \frac{\mu_k}{2}\|\Delta^k\|_2^2$$

$$\le \beta\left[\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k)\right] + \underbrace{\frac{L_H}{6}\|\Delta^k\|_2^3 - \frac{\mu_k}{2}(2 - 2\beta)\|\Delta^k\|_2^2}$$

provided $\beta \in (0, \frac{1}{2})$

$$\le \beta\left[\ell_k(\theta^k + \Delta^k) - \ell_k(\theta^k)\right] \quad \text{provided} \le 0$$

$$\Leftrightarrow \|\Delta^k\|_2 \le \frac{6\mu_k(1-\beta)}{L_H}$$

先找一个UB，如果这个UB小于右边，则成立.

Lemma: $\forall k$: $\|\Delta^k\|_2 \le \left[\frac{L_H}{2\mu_k}\mathrm{dist}(\theta^k, \Theta) + 2\right]\mathrm{dist}(\theta^k, \Theta)$

Then, $\|\Delta^k\|_2 \le \left[\frac{L_H}{2C\|E(\theta^k)\|_2^{\rho}}\mu\|E(\theta^k)\|_2 + 2\right]\mu\|E(\theta^k)\|_2$

$$= \frac{\mu^2 L_H}{2C}\|E(\theta^k)\|_2^{2-\rho} + 2\mu\|E(\theta^k)\|_2 \overset{\text{Want}}{\le} \frac{6(1-\beta)}{L_H} \cdot C\|E(\theta^k)\|_2^{\rho}$$

分类讨论：
$$\begin{cases} p\in[0,1) & \text{成立。} \\ p=1, & \text{成立。左右都是一次方，可以调 } c \text{ 成立。} \\ p>1, & \text{不成立，无法求为 UB。} \end{cases}$$

$\therefore$ In order to guartee prop 3, p should be $\in[0,1]$

2019.3.4

(P) $\min_{\theta} \{ \mathcal{L}(\theta) + R(\theta) \}$

Second Order Method

① Solve

$$\Delta^k = \arg\min_{\Delta} \left\{ \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T \Delta + \frac{1}{2} \Delta^T H_k \Delta + R(\theta^k + \Delta) \right\}$$

where $H_k = \nabla^2 \mathcal{L}(\theta^k) + \mu_k I$, $\mu_k = c \| E(\theta^k) \|_2^{\rho}$

$E(\theta) = \mathrm{prox}_R (\theta - \nabla \mathcal{L}(\theta)) - \theta$ $\quad c > 0, \rho \in [0, 1]$

② Update: $\theta^{k+1} = \theta^k + \alpha_k \Delta^k$

prop: Under the error bound assumption, i.e.

$\mathrm{dist}(\theta, \Theta) \le \mu \cdot \| E(\theta) \|_2$ for $\theta : \| E(\theta) \|_2 \le \varepsilon$,

then for sufficiently large k, $\alpha_k = 1$, and descent condition can be satified.

这种情况下，α_k 直接取 1 就可以了。

Lemma: $\forall k$, $\| \Delta^k \|_2 \le \left[\frac{L_H}{2 \mu_k} \mathrm{dist}(\theta^k, \Theta) + 2 \right] \mathrm{dist}(\theta^k, \Theta)$

以上是回顾。

以上已知其可以 converge，下面看 convergence rate.

Local Convergence Rate Analysis.

Based on EB.　　　　说 local 需要 k 足够大.

For large k, $\Delta^k = \theta^{k+1} - \theta^k$ （当 k 足够大以后，$\alpha_k = 1$）
$E(\theta^k) \to 0$

$$\|\Delta^k\|_2 = \|\theta^{k+1} - \theta^k\|_2 \le \left[\frac{L_H}{2C\|E(\theta^k)\|_2^{\rho}} \operatorname{dist}(\theta^k, \Theta) + 2\right] \operatorname{dist}(\theta^k, \Theta)$$

这两项有个不等式关系.

by EB $\sim \operatorname{dist}^{\rho}(\theta^k, \Theta)$

主项

$$\sim \operatorname{dist}(\theta^k, \Theta)^{2-\rho} \quad \cdots\cdots \quad \operatorname{dist}(\theta^k, \Theta)^1$$

$$\le O(\operatorname{dist}(\theta^k, \Theta))$$

$$\operatorname{dist}(\theta^{k+1}, \Theta) \overset{EB}{\le} \mu \|E(\theta^{k+1})\|_2 = \mu \|E(\theta^{k+1}) - E_k(\theta^{k+1})\|_2$$

$$\begin{cases} E_k(\theta) = \operatorname{prox}_R(\theta - \nabla \tilde{\mathcal{L}}(\theta)) - \theta. \\ \tilde{\mathcal{L}}_k(\theta^k + \Delta) = \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T \Delta + \frac{1}{2} \Delta^T H_k \Delta. \end{cases}$$

$$\begin{cases} \because \theta^{k+1} \ \min \left\{ \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T (\theta - \theta^k) + \frac{1}{2} (\theta - \theta^k)^T H_k (\theta - \theta^k) + R(\theta) \right\} \\ \Rightarrow E_k(\theta^{k+1}) = 0 \end{cases}$$

$$\|E(\theta^{k+1}) - E(\theta^k)\|_2$$

(by def)

$$= \Big\| \operatorname{prox}_R\big(\theta^{k+1} - \nabla \mathcal{L}(\theta^{k+1})\big) - \theta^{k+1} - \Big[\operatorname{prox}_R\big(\theta^{k+1} - \overbrace{\nabla \mathcal{L}(\theta^k) - H_k(\theta^{k+1} - \theta^k)}^{\nabla \tilde{\mathcal{L}}(\theta^{k+1})}\big) - \theta^{k+1}\Big]\Big\|_2$$

(non-expansiveness of prox)

$$\le \|\nabla \mathcal{L}(\theta^{k+1}) - \nabla \mathcal{L}(\theta^k) - H_k(\theta^{k+1} - \theta^k)\|_2$$

(by def of H_k and Δ-ineq)

$$\le \|\nabla \mathcal{L}(\theta^{k+1}) - \nabla \mathcal{L}(\theta^k) - \nabla^2 \mathcal{L}(\theta^k)(\theta^{k+1} - \theta^k)\|_2 + \mu_k \|\theta^{k+1} - \theta^k\|_2$$

(Lipschitz cont of $\nabla^2 \mathcal{L}$)

$$\le \frac{L_H}{2}\|\theta^{k+1} - \theta^k\|_2^2 + c \cdot \|E(\theta^k)\|_2^{\rho} \cdot \|\theta^{k+1} - \theta^k\|_2$$

下面分析这一项看是什么阶的!

$$\|E(\theta^k)\|_2 = \|E(\theta^k) - E(\bar{\theta}^k)\|_2 \qquad \bar{\theta}^k = \Pi_{\Theta}(\theta^k)$$

其后项=0, 由定义可

$$= \Big\|\operatorname{prox}_R\big(\theta^k - \nabla \mathcal{L}(\theta^k)\big) - \theta^k - \Big[\operatorname{prox}_R\big(\bar{\theta}^k - \nabla \mathcal{L}(\bar{\theta}^k)\big) - \bar{\theta}^k\Big]\Big\|_2$$

non-expansiveness

$$\le \|\theta^k - \bar{\theta}^k\|_2 + \|\theta^k - \bar{\theta}^k\|_2 + \|\nabla \mathcal{L}(\theta^k) - \nabla \mathcal{L}(\bar{\theta}^k)\|_2$$

$$\le (L_G + 2)\, \mathrm{dist}(\theta^k, \Theta)$$

Hence, $\mathrm{dist}(\theta^{k+1}, \Theta) \le O\big(\mathrm{dist}(\theta^k, \Theta)^2\big) + O\big(\mathrm{dist}(\theta^k, \Theta)^{1+\rho}\big)$

$$= O\big(\mathrm{dist}(\theta^k, \Theta)^{1+\rho}\big)$$

superlinear when $\rho \in (0,1)$

quadratic if $\rho = 1$

linear if $\rho=0$, provided $c>0$ is chosen carefully

那么我们能否直接选 $\rho=1$?

不能! 注意到之前的那个 sub-problem,

$$\mathcal{L}(\theta^k)+\nabla\mathcal{L}(\theta^k)^T\Delta+\frac{1}{2}\Delta^T H_k\Delta+R(\theta^k+\Delta)$$

$$H_k=\nabla^2\mathcal{L}(\theta^k)+\mu_k I \qquad \mu_k=c\cdot\|E(\theta^k)\|_2^{\rho}$$

当 $\rho=0$ 时, μ_k 是常数 ∴ always guarteen

ρ 太大时, $\mu_k\downarrow$, H_k 变得 ill-conditioned.

∴ 要做一个权衡

big outer small inner v.s. small outer big inner,

下面看下一个问题.

(P) $\min\limits_{\theta}\{F(\theta)\triangleq\mathcal{L}(\theta)+\lambda R(\theta)\}$

Non-convex Instances of (P).

$\mathcal{L}$, R: both can be non-convex.

Example: Linear Model. w/ additively corrupted covariates

Recall $y=X\theta^*+\varepsilon$

$$X = \begin{bmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix}$$

X: Covariate vector.

ε: mean 0, iid

(E) $\hat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}$

Assumption: x_i is iid. (mean 0) w/ covariance $\Sigma_x > 0$ (known)

e.g.: $x_i \sim N(0, \Sigma_x)$

Then, an idealized version of the est. problem is:

$$\min_{\theta} \left\{ \frac{1}{2} \theta^T \underbrace{\Sigma_x}_{\Gamma} \theta - \theta^T \underbrace{\Sigma_x \theta^*}_{\gamma} \right\}$$

这个问题的FOC: $\Sigma_x \theta - \Sigma_x \theta^* = 0 \Rightarrow \theta = \theta^*$

如果知道θ^*就直接解出来了，这里可以把 Σ_x 和 $\Sigma_x \theta^*$ 当成参数.

How is this related to (E)?

$$\frac{1}{2n} \| y - X\theta \|_2^2 = \frac{1}{2n} \left[y^T y - 2\theta^T X^T y + \theta^T X^T X \theta \right]$$

扔掉常数.

$$\sim \frac{1}{2n} \left[\theta^T \underbrace{X^T X}_{\hat{\Gamma}} \theta - 2\theta^T \underbrace{X^T y}_{\hat{\gamma}} \right]$$

consider $\hat{\Gamma} = \frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$

$$\mathbb{E}_{(x,\varepsilon)} [\hat{\Gamma}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[x_i x_i^T] = \Sigma_x$$

$\therefore \hat{\Gamma}$ is a sample average of Σ_x

unbiased est. of Σ_x

consider $\hat{\gamma} = \frac{1}{n} X^T y$ $\quad y = X\theta^* + \varepsilon$

$$\mathbb{E}_{(x,\varepsilon)} [\hat{\gamma}] = \mathbb{E}_x \mathbb{E}_\varepsilon [\tfrac{1}{n} X^T y]$$

$$= \frac{1}{n} \mathbb{E}_x X^T \mathbb{E}_\varepsilon [y]$$

$$= \frac{1}{n} \mathbb{E}_x [X^T X \theta^*]$$

$$= \Sigma_x \theta^*$$

下面看如何根据这个性质来建模.

Consider: we do not observe x_i directly,

but through

现在 x_i 不能直接观察到, 也是有噪声.

$$z_i = x_i + w_i$$

w_i: random, mean 0, covariance Σ_w (assume to be known)

$$\frac{1}{n}\mathbb{E}_w\left[Z^TZ\right] = \frac{1}{n}X^TX + \Sigma_w$$

$$\frac{1}{n}\mathbb{E}_w\left[Z^Ty\right] = \frac{1}{n}X^Ty$$

In this setting: We can use:

$$\hat{\Gamma} = \frac{1}{n}Z^TZ - \Sigma_w$$

$$\hat{\gamma} = \frac{1}{n}Z^Ty$$

Problem:

两个PSD一减，就不知道是否PSD了.

$$\frac{1}{2}\left[\theta^T\left[\frac{1}{n}Z^TZ - \Sigma_w\right]\theta - \frac{1}{n}\theta^TZ^Ty\right]$$

$Z: n \times d$ $\Sigma_w: d \times d$

rank $\le n$

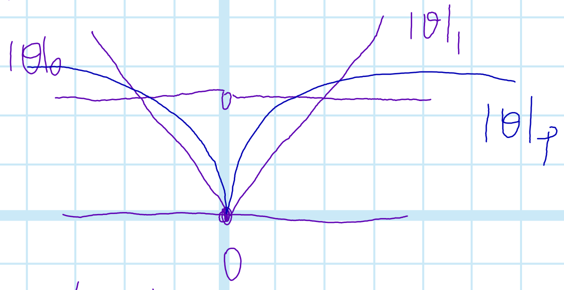
以上给出了一个 loss function 不是 convex 的例子

2019.3.5.

Non-convex Regularized loss

(P) $\min_{\theta}\{F(\theta) \triangleq \mathcal{L}(\theta) + \lambda R(\theta)\}$

$R(\theta) = \|\theta\|_1$ convex approx. of $\|\theta\|_0$ ← how many non-zeros



$\|\theta\|_p^p = \sum_i |\theta_i|^p \quad 0<p<1$

ℓ_p-quasi-norm (bridge-penalty)

Observe:

$\lim_{\theta\to 0} (|\theta|^p)' = +\infty$

$\Rightarrow \hat{\theta} = 0$ is a local min to (P)

$\Rightarrow$ cannot bound $\|\hat{\theta} - \theta^*\|_2$ for all local min in general

$\because \hat{\theta}=0$ 是一个最优解，但并不是 θ^*.

Consider a family of regularizers 形式化定义下 regularizer.

R_λ:

① separable: $R_\lambda(\theta) = \sum_{i=1}^{d} R_\lambda(\theta_i)$ 可以拆成每个维度

① $R_\lambda(0)=0$, $R_\lambda(t)=R_\lambda(-t)$ $\forall t\in\mathbb{R}$

② R_λ is non-decreasing on $\mathbb{R}_+$

③ For $t>0$, $t\to\dfrac{R_\lambda(t)}{t}$ is non-increasing in t

∵ 为了模仿 $\|\cdot\|_0$

④ R_λ is differentiable $\forall t\neq 0$, and subdifferentiable at $t=0$. with $\lim\limits_{t\downarrow 0} R_\lambda'(t)=\lambda L$ for some $L>0$

希望在0点处导数不为 $+\infty$

⑤ $\exists \mu>0$, s.t. $t\mapsto R_{\lambda,\mu}(t)=R_\lambda(t)+\dfrac{\mu}{2}t^2$ is convex

(weak-convexity of R_λ)

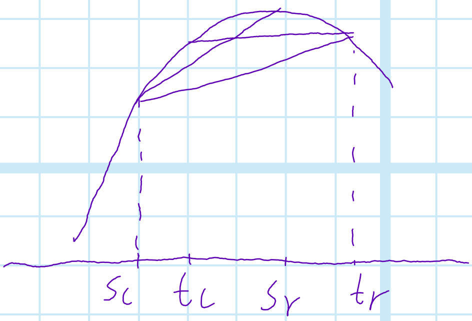
和 strong-convex 相反.

注意: ③: If R_λ is concave, Then it satisfies ③

由图容易看出不等式:

$$\frac{R_\lambda(s_r)-R_\lambda(s_l)}{s_r-s_l}\geq\frac{R_\lambda(t_r)-R_\lambda(s_l)}{t_r-s_l}$$

$$\geq\frac{R_\lambda(t_r)-R_\lambda(t_l)}{t_r-t_l}$$



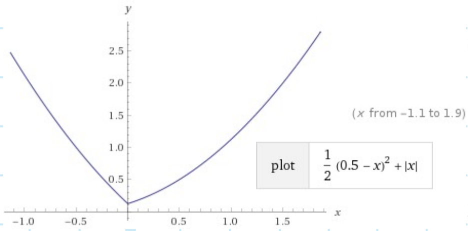
$s_l<s_r$, $t_l<t_r$

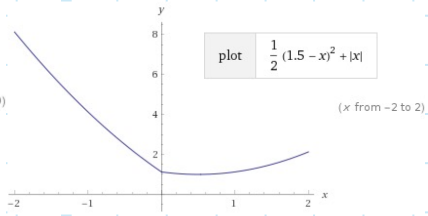
$s_l\leq t_l$, $s_r\leq t_r$

This family implements a Thresholding rule.

Consider:

$$\underset{t}{\operatorname{argmin}} \left\{ \frac{(z-t)^2}{2} + R_\lambda(t) \right\}$$





e.g. $R_\lambda(t) = \lambda|t|$

给定一个z，求出一个最小的t.

可见 $0 \in \underset{t}{\operatorname{argmin}} \left\{ \frac{(z-t)^2}{2} + R_\lambda(t) \right\}$ iff $|z| \le \lambda$

More generally, define

$$\lambda^* = \inf_{t>0} \left\{ \frac{t}{2} + \frac{R_t(t)}{t} \right\}$$

e.g. $R_\lambda(t) = \lambda|t|$ $\quad \lambda^* = \lambda$ (注意 $t>0$)

Claim: $0 = \underset{t}{\operatorname{argmin}} \left\{ \frac{(z-t)^2}{2} + R_\lambda(t) \right\}$ iff $|z| \le \lambda^*$

sketch:
$$\frac{(z-t)^2}{2} + R_\lambda(t) - \frac{z^2}{2} = t\left(\frac{t}{2} + \frac{R_\lambda(t)}{t} - z\right)$$

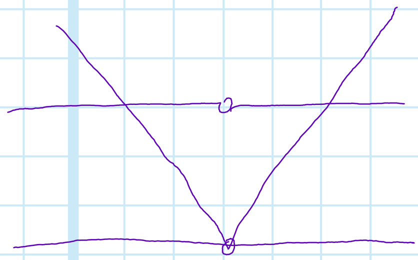
下面看一些 regularizer 的例子.

Examples:

① $R_\lambda(t) = \lambda|t|$

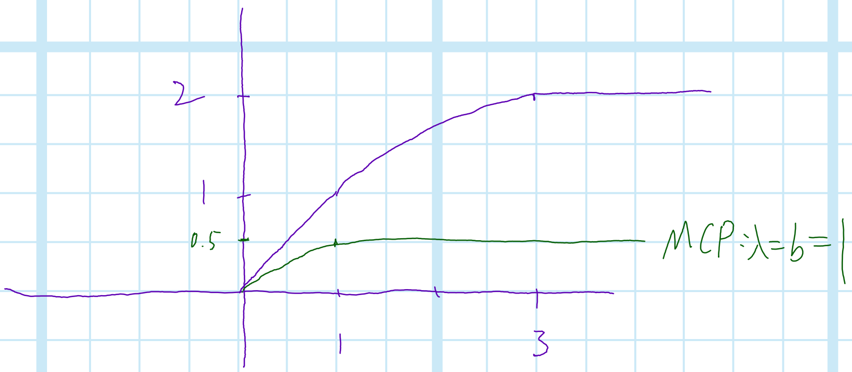
② Smoothly Clipped Absolute Deviation (SCAD)

$$R_\lambda(t) = \begin{cases} \lambda|t| & \text{for } |t| \le \lambda \\ -\dfrac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)} & \text{for } \lambda < |t| \le a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{for } |t| > a\lambda \end{cases}$$



here, $a > 2$ is fixed

SCAD: $\lambda = 1, a = 3$



以下验证 condition ④:

$$R_\lambda'(t) = \mathrm{sgn}(t) \cdot \begin{cases} \lambda & 0 < t < \lambda \\ -\dfrac{2t - 2a\lambda}{2(a-1)} = \dfrac{a\lambda - t}{a-1} & \lambda < t < a\lambda \\ 0 & t > a\lambda \end{cases}$$

t左右对称

写成 compact 形式:

$$= \mathrm{sgn}(t)\left[\lambda \mathbb{1}_{\{|t| \le \lambda\}} + \frac{(a\lambda - t)_+}{a-1} \mathbb{1}_{\{|t| > a\lambda\}}\right]$$

for $t \ne 0$, For $t = 0$,

$$\partial R_\lambda(0) = [-\lambda, \lambda] \Rightarrow \lim_{t \downarrow 0} R'(t) = \lambda \quad (L = 1)$$

以下验证 condition ⑤

Take $\mu = \frac{1}{a-1}$. Then, $R_{\lambda,\mu}(t) = R_\lambda(t) + \frac{\mu}{2}t^2 = \begin{cases} \lambda|t| + \frac{\mu}{2}t^2 \\ \frac{a\lambda}{a-1}|t| - \frac{\lambda^2}{2(a-1)} \\ \frac{(a+1)\lambda^2}{2} + \frac{\mu}{2}t^2 \end{cases}$

③ Minimax Concave Penalty (MCP)

$$R_\lambda(t) = \lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz, \quad b>0 \text{ is fixed}$$

下面来画图.

For $t>0$

$$R_\lambda(t) = \lambda \int_0^{t} \left(1 - \frac{z}{\lambda b}\right)_+ dz$$

$$= \lambda \int_0^{\min\{t, \lambda b\}} \left(1 - \frac{z}{\lambda b}\right) dz$$

$$= \lambda \left[\min\{t, \lambda b\} - \frac{1}{2\lambda b} \left(\min\{t, \lambda b\}\right)^2 \right]$$

画图: MCP: $\lambda = b = 1$ 见上一页

见上一页

下面来验证 MCP 是否满足条件.

Condition ④:

$$R_\lambda'(t) = \operatorname{sgn}(t) \cdot \lambda \cdot \left(1 - \frac{|t|}{\lambda b}\right)_+ \quad \text{for } t \neq 0.$$

For $t=0$, $\lim_{t \downarrow 0} R_\lambda'(t) = \lambda \quad (L=1)$

Condition ⑤

Take $\mu = \frac{1}{b}$

下面继续讨论.

① stat error bound $\|\hat{\theta} - \theta^*\|_2$ 现在不是 convex 的了.

② alg. conv analysis.

2019.3.11

Recamp.

感觉方法就是为了做non-convex的分析，去掉了Regalizer的cvx性质，需引入regalarer的其它性质以便分析

Non-Convex Regularities

$R_\lambda : \mathbb{R} \to \mathbb{R}$

(1) $R_\lambda(0) = 0$, $R_\lambda(t) = R_\lambda(-t)$ $\forall t$.

(2) R_λ is non-decreasing on $\mathbb{R}_+$

(3) $\forall t$, $t \longrightarrow \dfrac{R_\lambda(t)}{t}$ is non-increasing

(4) R_λ differentiable $\forall t \neq 0$, sub-differentiable at $t=0$, with $\lim_{t \downarrow 0} R_\lambda'(t) = \lambda L$ for some L

(5) $\exists \mu > 0$, s.t. $t \longmapsto R_{\lambda,\mu}(t) \triangleq R_\lambda(t) + \frac{\mu}{2} t^2$ is cvx

不要非凸得太严重.

上节课讲的两个例子都是piece-wise quadratic

$\therefore$ 加上一个二次项就能convex.

$$\hat{\theta} \in \min_{\|\theta\|_1 \le R} \{ F(\theta) = \mathcal{L}(\theta) + R_\lambda(\theta) \} \qquad (P)$$

R: radius is chosen s.t. $\|\theta^*\|_1 \le R$.

Goal: Estimation/Statistical error

$$\|\hat{\theta} - \theta^*\|_2$$

for any first-order point $\hat{\theta}$?

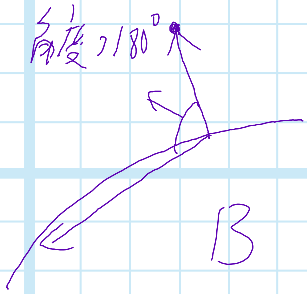
满足FOC的点，也可能是最大点

先写FOC:

$\tilde{\theta}$ is a first order point if:

$$\tilde{\theta} = \Pi_B\left(\tilde{\theta} - \nabla F(\tilde{\theta})\right)$$

$$B = \{\theta : \|\theta\|_1 \le R\}$$



或者写成不等式：(projection不好往下写，写不等式)

$$\left(\tilde{\theta} - \nabla F(\tilde{\theta}) - \tilde{\theta}\right)(\theta - \tilde{\theta}) \le 0 \quad \forall \theta \in B$$

$$\iff \left(\nabla \mathcal{L}(\tilde{\theta}) + \nabla \ell_\lambda(\tilde{\theta})\right)^T (\theta - \tilde{\theta}) \ge 0 \quad \forall \theta \in B$$

下面先定义一下性质。 (restricted strong convex)

Def: $\mathcal{L}$ satisfies the following RSC:

$$\left(\nabla \mathcal{L}(\theta^* + \Delta) - \nabla \mathcal{L}(\theta^*)\right)^T \Delta$$

$$\ge \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log d}{n} \|\Delta\|_1^2 & \text{if } \|\Delta\|_2 \le 1 \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log d}{n}} \|\Delta\|_1 & \text{if } \|\Delta\|_2 \ge 1 \end{cases}$$

n是样本数
d是维度.

下面看如何从 strong convexity 来得出.

$$g(y) \geq g(x) + \nabla g(x)^T (y-x) + \frac{\alpha}{2}\|x-y\|_2^2$$

$$+)\quad g(x) \geq g(y) + \nabla g(y)^T (x-y) + \frac{\alpha}{2}\|x-y\|_2^2$$

$$0 \geq (\nabla g(x) - \nabla g(y))(y-x) + \alpha\|x-y\|_2^2$$

$$\qquad -\Delta \qquad\qquad \Delta$$

下面给出 Thm.

Thm: Under the above setting ($\mathcal{L}$ 满足RSC, R 满足之前那三条.), with:

$$\alpha_1 > \frac{\mu}{2},\quad \theta^* \text{ is feasible for (P)} \leftarrow \theta^* \in B$$

$$\frac{4}{L}\max\left\{\|\nabla\mathcal{L}(\theta^*)\|_\infty,\ \alpha_2\sqrt{\frac{\log d}{n}}\right\} \leq \lambda \leq \frac{\alpha_2}{6LR}$$

希望λ尽量大, 这样能更符合 sparsity.

and $n \geq \frac{16R^2\max\{\tau_1^2, \tau_2^2\}}{\alpha_2^2}\log d$ every first-order point

$\tilde{\theta}$ satisfies

$$\|\tilde{\theta} - \theta^*\|_2 \leq \frac{3\lambda L\sqrt{k}}{2\alpha_1 - \mu} \qquad \text{where } k = \|\theta^*\|_0$$

Pf: Define $\tilde{\Delta} = \tilde{\theta} - \theta^*$

Claim: $\|\tilde{\Delta}\|_2 \leq 1$ 注意到RSC是分两段的.

By RSC,

$$(\nabla\mathcal{L}(\tilde{\theta}) - \nabla\mathcal{L}(\theta^*))^T\tilde{\Delta} \geq$$

$\geq \alpha_1 \|\tilde{\Delta}\|_2^2 - \tau_1 \frac{\log d}{n} \|\tilde{\Delta}\|_1^2$ 下面把不等式两边往 $\|\tilde{\Delta}\|_2^2$ 凑.

$R_{\lambda,\mu}(t)$

Since $R_\lambda(t) + \frac{\mu}{2} t^2$ is convex, (Regalizer ⑤)

$$R_{\lambda,\mu}(\theta^*) - R_{\lambda,\mu}(\tilde{\theta}) \geq \nabla R_{\lambda,\mu}(\tilde{\theta})^T (\theta^* - \tilde{\theta})$$

$$= (\nabla R_\lambda(\tilde{\theta}) + \mu \tilde{\theta})^T (-\tilde{\Delta})$$

$$\Rightarrow \nabla R_\lambda(\tilde{\theta})^T (\theta^* - \tilde{\theta}) \leq R_\lambda(\theta^*) - R_\lambda(\tilde{\theta}) + \frac{\mu}{2} \|\tilde{\theta} - \theta^*\|_2^2$$

$$\therefore \alpha_1 \|\tilde{\Delta}\|_2^2 - \tau_1 \frac{\log d}{n} \|\tilde{\Delta}\|_1^2$$

$$\leq \underbrace{\nabla \mathcal{L}(\tilde{\theta})^T \tilde{\Delta}}_{\text{FOC}} - \nabla \mathcal{L}(\theta^*)^T \tilde{\Delta} \quad \text{(def. of RSC)}$$

$$\leq -\nabla R_\lambda(\tilde{\theta})^T \tilde{\Delta} - \nabla \mathcal{L}(\theta^*)^T \tilde{\Delta}$$

$$\leq -\nabla \mathcal{L}(\theta^*)^T \tilde{\Delta} + R_\lambda(\theta^*) - R_\lambda(\tilde{\theta}) + \frac{\mu}{2} \|\tilde{\Delta}\|_2^2$$

C-S 不等式

$$\leq \underbrace{\|\nabla \mathcal{L}(\theta^*)\|_\infty}_{\leq \frac{\lambda L}{4}} \|\tilde{\Delta}\|_1 + \underbrace{R_\lambda(\theta^*) - R_\lambda(\tilde{\theta})}_{\text{需要和 } \|\tilde{\Delta}\|_1 \text{ 产生联系}} + \frac{\mu}{2} \|\tilde{\Delta}\|_2^2$$

为了让 R 和 $\|\tilde{\Delta}\|_1$ 产生联系, 引入如下:

Claim2 Let $\theta \in \mathbb{R}^d$ and S be index set of the k largest entries (in magnitude) of θ. Then $R_\lambda(\theta_S) - R_\lambda(\theta_{S^c})$, Then,

$$R_\lambda(\theta_S) - R_\lambda(\theta_{S^c}) \le \lambda L(\|S_c\|_1 - \|\theta_{S^c}\|_1)$$

Also, if θ^* is k-sparse. then, $\forall\, \theta$:

$$R(\theta^*) - R_\lambda(\theta) \le \lambda L(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1),$$

$\Delta = \theta - \theta^*$, S: index set of k largest entries of Δ.

$$\le \frac{\lambda L}{4} \underbrace{\|\tilde\Delta\|_1}_{\|\tilde\Delta_S\|_1 + \|\tilde\Delta_{S^c}\|_1} + \lambda L(\|\tilde\Delta_S\|_1 - \|\tilde\Delta_{S^c}\|_1) + \frac{\mu}{2}\|\tilde\Delta\|_2^2$$

下面合并一下同类项.

$$\left(\alpha_1 - \frac{\mu}{2}\right)\|\tilde\Delta\|_2^2 \le \frac{5\lambda L}{4}\|\tilde\Delta_S\|_1 - \frac{3\lambda L}{4}\|\tilde\Delta_{S^c}\|_1 + \tau_1 \frac{\log d}{n}\|\tilde\Delta\|_1^2$$

右边有的 $\|\cdot\|_1$ 有平方, 有的没有. 可以用那个 Ball 来控制平方.

$$\tau_1 \frac{\log d}{n}\|\tilde\Delta\|_1^2 \le 2R\tau_1 \frac{\log d}{n}\|\tilde\Delta\|_1 \text{ (三角不等式)}$$

$$\le \frac{5\lambda L}{4}\|\tilde\Delta_S\|_1 - \frac{3\lambda L}{4}\|\tilde\Delta_{S^c}\|_1 + 2R\tau_1\frac{\log d}{n}\|\tilde\Delta\|_1$$

$$\le \cdots + \alpha_2\sqrt{\frac{\log d}{n}}\|\tilde\Delta\|_1$$

这里是用 $n \ge \frac{16R^2\max\{\tau_1^2, \tau_2^2\}}{\alpha_2^2}\log d$ 那个假设

$$\le \cdots + \frac{\lambda L}{4}\|\tilde\Delta\|_1$$

(利用前面的 Assumption.

注意到 $\|\tilde{\Delta}\|_1 = \|\tilde{\Delta}_S\|_1 + \|\tilde{\Delta}_{S^c}\|_1$

$$\le \frac{3\lambda L}{2}\|\tilde{\Delta}_S\|_1 - \frac{\lambda L}{2}\|\tilde{\Delta}_{S^c}\|_1$$

$\Rightarrow$

$$2\left(\alpha_1 - \frac{\mu}{2}\right)\|\tilde{\Delta}\|_2^2 \le 3\lambda L\|\tilde{\Delta}_S\|_1 \le 3\lambda\sqrt{k}\|\tilde{\Delta}_S\|_2$$

$$\le 3\lambda L\sqrt{k}\|\tilde{\Delta}\|_2$$

下面来证那个Claim 2.

Pf of Claim 2: Define $f(t) = \dfrac{t}{R_\lambda(t)}$ for $t > 0$,

$$\|\theta_{S^c}\|_1 = \sum_{j\in S^c} R_\lambda(\theta_j)\underbrace{\frac{|\theta_j|}{R_\lambda(|\theta_j|)}}_{f(|\theta_j|)} \overset{\text{由}R_\lambda\text{的(3)}}{\le} \sum_{j\in S^c} R_\lambda(\theta_j) f(\|\theta_{S^c}\|_\infty)$$

$$= R_\lambda(\theta_{S^c}) f(\|\theta_{S^c}\|_\infty)$$

Similarly

$$R_\lambda(\theta_S)\cdot f(\|\theta_{S^c}\|_\infty) \le \|\theta_S\|_1 \qquad \underline{\text{EXERCISE}}$$

注意 θ_{S^c} 中的任一维都小于 θ_S 中的任一维

$$R_\lambda(\theta_s) - R_\lambda(\theta_{s^0}) \le \frac{1}{f(\|\theta_{s^c}\|_\infty)}\left(\|\theta_s\|_1 - \|\theta_{s^c}\|_1\right)$$

用 (4)

For $t \geq s > 0$

$$f(t) \geq f(s) = \frac{s-0}{R_\lambda(s) - R_\lambda(0)}$$ 取个 lim 也成立.

$$\Rightarrow f(t) \geq \lim_{s \downarrow 0} f(s) = \frac{1}{\lambda L}$$

代回去即证完了 Claim 2.

下节课讲如何计算 first order point.

2019.3.12

本节主要讲算法.

Recamp:

(P) $\min\{\mathcal{L}(\theta) + R_\lambda(\theta)\}$

$\mathcal{L}$: smooth non-convex

R_λ: Regularizer non-smooth

$\rightarrow$ (P) is non-convex non-smooth.

Algorithmily,

Find descent direction at x

For smooth problems, e.g.

$$\min f(x)$$

a descent direction $-\nabla f(x)$

For non-convex, non-smooth:

NP-hard.

下面来证NP-hard.

Consider $c \in \mathbb{Z}_+^n$ s.t. $r = \sum_i c_i \geq 1$ (不全为0)

Define $f(x) = (1 - \frac{1}{r}) \max_i |x_i| - \min_i |x_i| + |c^T x|$

<u>Note</u>: $f(0) = 0$

Claim (Nesterov) It is NP-hard to decide

if $\exists x$, s.t. $f(x) < 0$

Pf: This is equivalent to deciding if

$\exists \sigma \in \{\pm 1\}$ s.t. $c^T \sigma = 0$ (就是把集合分成两个和相等的问题)

(Partition Problem, NP-hard)

下面证这两个问题等价.

($\Leftarrow$) Suppose $\exists \sigma \in \{\pm 1\}^n$ s.t. $c^T \sigma = 0$

Then $f(\sigma) = -\frac{1}{r} < 0$

($\Rightarrow$) Suppose $\exists x$, s.t. $f(x) < 0$.

Normalize: $\max |x_i| = 1$

Set $\delta = |c^T x|$, Then

$$0 > f(x) = 1 - \frac{1}{r} - \min_i |x_i| + \delta$$

$$\Leftrightarrow \quad |x_i| > 1 - \frac{1}{r} + \delta \quad \forall i.$$

Set $\sigma_i = \mathrm{sgn}(x_i)$. Then,

$$|x_i| = \sigma_i x_i > 1 - \frac{1}{\gamma} + \delta$$

$$\Rightarrow |\sigma_i - x_i| = |\sigma_i| \cdot |\sigma_i - x_i|$$

$$= 1 - \sigma_i x_i < \frac{1}{\gamma} - \delta$$

$$\Rightarrow |c^T\sigma| = |c^T x| + |c^T(\sigma - x)| \quad \text{(triangle inequality)}$$

$$\le \delta + \|c\|_1 \cdot \|\sigma - x\|_\infty$$

$$< \delta + \gamma\left(\frac{1}{\gamma} - \delta\right) \quad (-S$$

$$= 1 + (1-\gamma)\delta \le 1$$

$$\Rightarrow c^T\sigma = 0 \quad \text{b/c} \quad c \in \mathbb{Z}_+^n, \ \sigma \in \{\pm\}^n$$

如果$|c^T\sigma| < 0$ 它只能 $=0$.

∴我们找到了，只要取 $\sigma_i = \mathrm{sgn}(x_i)$ 那么就能直接 $\Rightarrow c^T\sigma = 0$.

根据 ($\Leftarrow$) 和 ($\Rightarrow$) 推出了两个问题可以在多项式时间内相互转换，因而是等价的复杂度.

Use structure of (P):

$$\mathcal{L}(\theta) + R_\lambda(\theta)$$

利用$R_\lambda(\theta)$是weakly cvx.

$$= \underbrace{\left(\mathcal{L}(\theta) - \frac{\mu}{2}\|\theta\|_2^2\right)}_{\substack{\text{smooth} \\ \text{non-convex}}} + \underbrace{\left(R_\lambda(\theta) + \frac{\mu}{2}\|\theta\|_2^2\right)}_{\substack{\text{convex} \\ \text{non-smooth}}}$$

Nestrov: composite gradient

"converge to a point with no descent direction".

下面来看算法:

Want to solve:

$$\min_{\bar{R}_{\lambda,\mu}(\theta)\le R} \left\{ \underbrace{\left(\mathcal{L}(\theta) - \frac{\mu}{2}\|\theta\|_2^2 \right)}_{\bar{\mathcal{L}}(\theta)} + \lambda \bar{R}_{\lambda,\mu}(\theta) \right\}$$

where $\bar{R}_{\lambda,\mu}(\theta) = \frac{1}{\lambda} R_{\lambda,\mu}(\theta) = \frac{1}{\lambda}\left[R_\lambda(\theta) + \frac{\mu}{2}\|\theta\|_2^2 \right]$

<u>Alg</u>:

(A) $$\theta^{t+1} = \operatorname*{argmin}_{\bar{R}_{\lambda,\mu}(\theta)\le R} \left\{ \frac{1}{2}\left\| \theta - \left(\theta^t - \frac{1}{\eta}\nabla\bar{\mathcal{L}}(\theta^t)\right)\right\|_2^2 + \frac{\lambda}{\eta}\bar{R}_{\lambda,\mu}(\theta) \right\}$$

是一个 strongly cvx problem, w.r.t cvx region

Compare:

$$\operatorname{prox}_{\frac{1}{\eta}(\lambda\bar{R}_{\lambda,\mu})}\left(\theta^t - \frac{1}{\eta}\nabla\bar{\mathcal{L}}(\theta^t)\right) \triangleq \operatorname*{argmin}_{\theta}(\cdots)$$

也可以用两步:

2-stage approach

Step 1: Compute

$$\hat{\theta}^{t+1} = \operatorname{prox}_{\frac{1}{\eta}(\lambda\bar{R}_{\lambda,\mu})}\left(\theta^t - \frac{1}{\eta}\nabla\bar{\mathcal{L}}(\theta^t)\right)$$

Step 2:

If $\bar{R}_{\lambda,\mu}(\tilde{\theta}^{t+1}) \le R$ then

$$\theta^{t+1} \leftarrow \tilde{\theta}^{t+1}$$

Else, set

$$\theta^{t+1} = \underset{\bar{R}_{\lambda,\mu}(\theta)\le R}{\arg\min} \left\{ \frac{1}{2} \left\| \theta - \left(\theta^t - \frac{\nabla \tilde{\mathcal{L}}(\theta^t)}{\eta} \right) \right\|_2^2 \right\}$$

$$= \Pi_{\bar{R}_{\lambda,\mu}(\theta)\le R} \left[\theta^t - \frac{\nabla \tilde{\mathcal{L}}(\theta^t)}{\eta} \right]$$

下面来verify 用两步的是等价的

Correctness:

If $\bar{R}_{\lambda,\mu}(\tilde{\theta}^{t+1}) \le R$, Then ✓

Else, observe:

有两个优化问题, 一个有约束, 一个没有.

$x^* = \arg\min f(x)$, $\hat{x} = \underset{x\in C}{\arg\min} f(x)$

若 $x^* \notin C$, $\Rightarrow \hat{x} \in \mathrm{bd}(C)$

如果不在C内, 那么$\hat{x}$应该在C的边界上.

$\bar{R}_{\lambda,\mu}(\theta^{t+1}) = R$ (在边界上.)

Claim: θ^{t+1} solves step 2.

Pf: Suppose not. Then, $\exists\, \bar{\theta}$, s.t.

$$\bar{R}_{\lambda,\mu}(\bar{\theta}) \le R \quad \text{and}$$

$$\frac{1}{2}\left\| \bar{\theta} - \left(\theta^t - \frac{\nabla \bar{\mathcal{L}}(\theta^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta}\bar{R}_{\lambda,\mu}(\bar{\theta})$$

$$< \frac{1}{2}\left\| \theta^{t+1} - \left(\theta^t - \frac{\nabla \bar{\mathcal{L}}(\theta^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta}\bar{R}_{\lambda,\mu}(\theta^{t+1})$$

下面那项更大.

Contradicts optimality of θ^{t+1} wrt (A)

Sometimes, step 1 can be computed in closed form.

Recall in step 1:

$$\arg\min_{\theta}\left\{\frac{1}{2}\left\|\theta - \left(\theta^t - \frac{\nabla \bar{\mathcal{L}}(\theta^t)}{\eta}\right)\right\|_2^2 + \frac{1}{\eta}R_\lambda(\theta) + \frac{\mu}{2\eta}\|\theta\|_2^2\right\}$$

$$= \arg\min_{\theta}\left\{\frac{1}{2}\left(1+\frac{\mu}{\eta}\right)\|\theta\|_2^2 - \theta^T\left(\theta^t - \frac{\nabla \bar{\mathcal{L}}(\theta^t)}{2}\right) + \frac{1}{\eta}R_\lambda(\theta)\right\}$$

$$= \arg\min_{\theta}\left\{\frac{1}{2}\left\|\theta - \frac{1}{1+\frac{\mu}{\eta}}\left(\theta^t - \frac{\nabla \bar{\mathcal{L}}(\theta^t)}{\eta}\right)\right\|_2^2 + \frac{1/\eta}{1+\mu/\eta}R_\lambda(\theta)\right\}$$

The last expression takes the form

$$\frac{1}{2}(x-c)^2 + \nu R_\lambda(x)$$

in each coordinate

e.g. SCAD.

SCAD:

$$R_\lambda(t)=\begin{cases}\lambda|t| & |t|\le\lambda\\ -\dfrac{t^2-2a\lambda|t|+\lambda^2}{2(a-1)} & \lambda<|t|\le a\lambda\\ \dfrac{(a+1)\lambda^2}{2} & |t|>a\lambda\end{cases}$$

FOC:

$$0\in\partial\left[\tfrac{1}{2}(x-c)^2+\nu R_\lambda(x)\right]$$

$$=x-c+\nu\,\partial R_\lambda(x)$$

$$=\begin{cases}-c+\nu[-\lambda,\lambda] & x=0\\ x-c+\nu\lambda & 0<x\le\lambda\\ x-c+\dfrac{\nu(a\lambda-x)}{(a-1)} & \lambda<x\le a\lambda\\ x-c & x>a\lambda\\ \cdots\cdots\ \text{负半轴对称} & \end{cases}$$

$$\hat{x}=\begin{cases}0 & \text{if } |c|\le\nu\lambda\\ c-\operatorname{sgn}(c)\nu\lambda & \text{if } \nu\lambda\le|c|\le(\nu+1)\lambda\\ \left(1-\dfrac{\nu}{a-1}\right)^{-1}\left(c-\operatorname{sgn}(c)\dfrac{a\nu\lambda}{a-1}\right) & \text{if } (\nu+1)\lambda\le|c|\le a\lambda\\ c & \text{if } |c|\ge a\lambda\end{cases}$$

以上是SCAD的结论，来计算 step 1.

本节讲了算法(A)和其等价算法.

方法在于转化为能求导和convex的部分.

Recall 2019.3.18.

$$\min_{\theta}\{\mathcal{L}(\theta)+R_\lambda(\theta)\}$$

$$=\min_{\theta}\Big\{\overbrace{\Big(\mathcal{L}(\theta)-\frac{\mu}{2}\|\theta\|_2^2\Big)}^{\bar{\mathcal{L}}}+\Big(R_\lambda(\theta)+\frac{\mu}{2}\|\theta\|_2^2\Big)\Big\}$$

smooth convex — non-smooth convex

把 non-smooth, non-convex 拆成了两个.

Algorithm:

$$\theta^{t+1}=\underset{\bar{R}_{\lambda,\mu}(\theta)\text{是}}{\operatorname{argmin}}\Big\{\frac{1}{2}\Big\|\theta-\Big(\theta^t-\frac{\nabla\bar{\mathcal{L}}(\theta^t)}{\eta}\Big)\Big\|_2^2+\frac{\lambda}{\eta}\bar{R}_{\lambda,\mu}(\theta)\Big\}$$

$$\bar{R}_{\lambda,\mu}=\frac{1}{\lambda}\Big(R_\lambda(\theta)+\frac{\mu}{2}\|\theta\|_2^2\Big)$$

Def: $\mathcal{L}$ satisfies RSC if

$$\mathcal{L}(\theta_1)-\mathcal{L}(\theta_2)-\nabla\mathcal{L}(\theta_2)^T(\theta_1-\theta_2)\geq$$

$$\begin{cases}\alpha_1\|\theta_1-\theta_2\|_2^2-\tau_1\dfrac{\log d}{n}\|\theta_1-\theta_2\|_1^2 & \text{if } \|\theta_1-\theta_2\|_2\leq C\\ \alpha_2\|\theta_1-\theta_2\|_2^2-\tau_2\sqrt{\dfrac{\log d}{n}}\|\theta_1-\theta_2\|_1 & \text{if } \|\theta_1-\theta_2\|\geq C\end{cases}$$

Def: $\mathcal{L}$ satisfies restricted smoothness (RS) if

$$\mathcal{L}(\theta_1) - \mathcal{L}(\theta_2) - \nabla\mathcal{L}(\theta_2)^T(\theta_1 - \theta_2) \le$$

$$\alpha_3 \|\theta_1 - \theta_2\|_2^2 + \tau_3 \frac{\log d}{n} \|\theta_1 - \theta_2\|_1^2$$

如果右边没有后面那个修正项，就变成了普通的 Lipschitz continuous gradient.

这两条提供了局部收敛速度的保证.

Thm (Informal)

Suppose $\mathcal{L}$ satisfies RSC & RS, Under appropriate choice of parameters, for sufficiently large t,

$$\max\{\|\theta^t - \hat{\theta}\|_2, \overbrace{\|\hat{\theta} - \theta^*\|_2}^{\text{statistic error}}\} = O\left(\sqrt{\frac{k \log d}{n}}\right)$$

where $k = \|\theta^*\|_0$ 代表 sparsity

it estibish a bound on $\|\theta^t - \hat{\theta}\|_2$, the bound does not converge to 0.

Even we found the optimality $\hat{\theta}$, we still have statistical error.

下面看另一类问题.

Phase Synchronization.

(有着更广的类叫 group synchronization).

Goal:

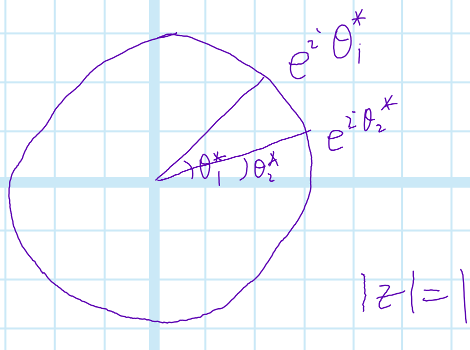
measurements: 观测到的信号

$$e^{i(\theta_j^* - \theta_k^*)} \quad \forall j,k, \quad 这里 i 是虚数.$$

recover: 希望 recover 每个 j 的数据.

$$e^{i\theta_j^*} \quad \forall j=1,\dots,n$$

类似于一个 localization 问题.
已知距离求坐标.



Rmk: There's a "group" (群) of interpretation of this problem

一个群是一个代数结构, 定义了 identity operation, and inverse.

$g_j^* = e^{i\theta_j^*}$ $\qquad g_j^* (g_k^*)^{-1}$

group element.

问题是用群来解释就是：已知两个群元素的组合，能否还原每个元素。

☆ 应用：有两组点云，能否组合到一起，

查下 group synchronization.

Let $z^* \in \mathbb{T}^n = \{ w \in \mathbb{C}^n : |w_i| = 1 \ \forall i \}$ be the ground truth.

$\mathbb{T}$ 表示 Tores ?

measurement model:

$$C_{jl} = z_j^* \bar{z}_l^* + \Delta_{jl} \qquad 1 \le j < l \le n.$$

实际上 $z_k^* = e^{-i\theta_k^*}$

Δ_{jl} 是噪声，加了噪声后可能就不是 group element 了

$z_j^* \bar{z}_l^* \Delta_{jl} \in \mathbb{T}$

Least-squares formulation:

$$\hat{z} \in \underset{z \in \mathbb{T}^n}{\operatorname{argmin}} \sum_{j,l} |C_{jl} - z_j \bar{z}_l|^2$$

compact, 用 Weierstrass

Assume: $\Delta_{jj}=0 \Rightarrow C_{jj}=1$ 我们知道模为1

把那个平方打开，利用 $|z_i \bar{z}_i|^2=1$ EXERCISE

问题转化为：

$$\hat{z} \in \arg\max_{z \in \mathbb{T}^n} z^H C z \quad (P)$$

z^H: Hermitian transpose of z.

Assume $\Delta=\Delta^H$, then $C=C^H$

这里如果 $C=C^H$ 那么 $z^H C z \in \mathbb{R}$

证：$(z^H C z)^H = z^H C^H z = z^H C z$

Observe: If $\hat{z}$ is optimal, then so is $e^{i\theta}\hat{z}$ for any $\theta \in [0, 2\pi)$

如果 $\hat{z}$ 是最优的，把 $\hat{z}$ 转一定的角度，仍然是最优的。

可以依此来定义 optimal 还有多远。

$\Rightarrow$ distance measure:

随意转，找个最近的

$$d_2(z, w) \triangleq \min_{\theta \in [0, 2\pi)} \|z - e^{i\theta} w\|_2$$

$z, w \in \mathbb{T}^n$.

SDP

一种方法是用 semi-definite approximation

这里我们直接求解。

Q: Estimation error $d_2(\hat{z}, z^*)$

Optimalization error $d_2(z^t, \hat{z})$

$\Leftarrow$ Relaxation of $\mathbb{T}^n$

prop: Let $z \in \mathbb{C}^n$ be s.t. $\|z\|_2^2 = n$ and

$(z^*)^H C z^* \le z^H C z$ Then,

$$d_2(z, z^*) \le \frac{4\|\Delta\|}{\sqrt{n}}$$

Pf: $d_2(z, z^*)^2 = \min\limits_{\theta \in [0, 2\pi)} \|z - e^{i\theta} z^*\|_2^2$

利用 $\|v\|_2^2 = v^H v$

$\|z\|_2^2 = \|z^*\|_2^2 = 1$

$$= 2\left[n - \max_{\theta \in [0, 2\pi)} \mathrm{Re}(e^{i\theta} z^H z^*)\right]$$

$e^{i\theta} z^H z^* = |z^H z^*|$ 时最大.

$$= 2(n - |z^H z^*|)$$

$$C_{jl} = z_j^* \bar{z}_l^* + \Delta_{jl}$$

$$C = z^* (z^*)^H + \Delta$$

$$z^H C z = |z^H z^*|^2 + z^H \Delta z \geq (z^*)^H C z^*$$

$$= n^2 + (z^*)^H \Delta z^*$$

$$\Rightarrow \quad n^2 - |z^H z^*|^2 \le z^H \Delta z - (z^*)^H \Delta z^*$$

$$(n - |z^H z^*|)(n + |z^H z^*|) \le z^H \Delta z - (z^*)^H \Delta z^*$$

$$n - |z^H z^*| \le \frac{1}{n}\left(z^H \Delta z - (z^*)^H \Delta z^*\right)$$

加上 $z^H \Delta z^* - (z^*)^H \Delta z$ (一定是虚数) 然后只保留实部.

$$= \frac{1}{n} \mathrm{Re}\left[(z - z^*)^H \Delta (z + z^*) \right]$$

$\|z\|_2^2 = n$

$\|z + z^*\| \le \|z\|_2 + \|z^*\|_2 = \sqrt{n} + \sqrt{n}$

$$\le \frac{1}{n} \|\Delta\| \cdot \|z - z^*\|_2 \cdot \|z + z^*\|_2$$

$$\le \frac{2}{\sqrt{n}} \|\Delta\| \cdot \underbrace{\|z - z^*\|_2}_{= d_2(z, z^*)}$$

EXERCISE

2019.3.19

Phase Synchronization

$$C = z^*(z^*)^H + \Delta$$

$$z^* \in \mathbb{T}^n = \{ w \in \mathbb{C}^n : |w_i| = 1 \}$$

$$\hat{z} \in \underset{z \in \mathbb{T}^n}{\operatorname{argmax}} \; z^H C z$$

$\uparrow$ non-convex, ∵只在边界上.
而不是边界内.

Prop 1: (Est. Error) Let $z \in \mathbb{C}^n$ be s.t. $\|z\|_2^2 = n$ and $(z^*)^H C z^* \le z^H C z$, Then, $d_2(z, z^*) \le \frac{4\|\Delta\|}{\sqrt{n}}$

$$d_2(z, w) = \min_{\theta \in [0, 2\pi)} \| z - e^{i\theta} w \|$$

来看传统SDP怎么做,

SDP:

$$Z = z z^H$$

Semedifinite Relaxation

$$\max \; C \cdot Z$$

$$\text{s.t. } \operatorname{Diag}(Z) = e,$$

$$Z \succeq 0 \quad (w^H Z w \ge 0, \forall w \in \mathbb{C}^n)$$

我们研究另一种方法:

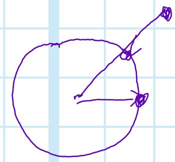
Projected Gradient Method

$$w^k \leftarrow z^k + \frac{\alpha_k}{n} C z^k \qquad \text{梯度上升}$$

$$z^{k+1} \leftarrow \frac{w^k}{|w^k|} \qquad \text{往 feasible 集合上投影.}$$

when given $w \in \mathbb{C}^n$

$$\left(\frac{w}{|w|}\right)_j = \begin{cases} \frac{w_j}{|w_j|} & \text{if } w_j \neq 0 \\ 1 & \text{if } w_j = 0 \end{cases}$$



圆心往圆周上投影
可取任一点，这里取1

那么关于这个算法我们可以研究：

Convergence ?

initialization?

下面研究 initialization.

Consider spectral estimation:

(spectral estimator 就是用某个最大 eigen value 来估 .)

Let $u \in \mathbb{C}^n$ be a leading eigenvector of C

$u^H C u = \|u\|^2$ ∴ eigen vector is real.

$v_c = \frac{u}{|u|} \in \mathbb{T}^n$ 这里是 component wise normalization 如上面定义.

prop 2: $d_2(v_c, z^*) \leq \frac{8\|\Delta\|}{\sqrt{n}}$

和 prop 1 只差了个常数，现在有了更加紧的 bound

能否不降若干系来满足 prop 1.

Thm (Est. Error of PG)

Suppose $\|\Delta\| \leq \frac{n}{16}$, If $z^0 = V_c$, $\alpha_k \geq 2$, Then,

$$d_2(z^{k+1}, z^*) \leq \underbrace{\mu^{k+1} d_2(z^0, z^*)}_{①} + \underbrace{\frac{\nu}{1-\mu} \frac{8\|\Delta\|}{\sqrt{n}}}_{②} \quad \forall k$$

where, $\mu = \frac{16(\alpha\|\Delta\|+n)}{(7\alpha+8)n} < 1 \quad \nu = \frac{2\alpha}{7\alpha+8}$

①: 由 optimization 的误差，随着迭代 $\downarrow$

②: 由 noise 引入的误差

prop 3: For any $w \in \mathbb{C}^n$ and $z \in \mathbb{T}^n$,

$$\left\| \frac{w}{|w|} - z \right\|_2 \leq 2\|w - z\|_2$$

在 convex 的情况下，我们有 non-expansiveness:

$\|\pi(x) - \pi(y)\|_2 \leq \|x - y\|_2$

这里没有 convex, we pay a factor 2.

下面用 prop 3 来证 prop 2.

Pf (prop 2). Choose u, s.t. $\|u\|_2^2 = n$ and

$u^H z^* = |u^H z^*|$, (w/o g) By def'n of u,

$$(z^*)^H C (z^*) \le u^H C u$$

$$d_2(v_c, z^*) \le \|v_c - z^*\|_2$$ 这里固定了 $\theta=0$ ∴放大了.

$$\le 2\|u - z^*\|_2 \quad (\text{prop } 3)$$

$$= 2 d_2(u, z^*)$$

$$d_2^2(u, z^*) = \min_\theta \|u - e^{i\theta} z^*\|_2^2$$

$$= 2\left[n - \max_\theta \mathrm{Re}(e^{i\theta} u^H z^*)\right]$$

$$\le \frac{8\|\Delta\|}{\sqrt{n}} \quad (\text{prop } 1)$$

$\blacksquare$

Pf (prop 3) It suffices to prove $\left|\frac{w_j}{|w_j|} - z_j\right| \le 2|w_j - z_j|$

即：如果每一维都成立，即总体也成立

WLOG, assume $z_j = 1$,

If $w_j = 0 \Rightarrow$ trivial

Consider $w_j \ne 0$, let $\frac{w_j}{|w_j|} = e^{i\phi}$ for some $\phi \in [0, 2\pi)$

Claim: $|e^{i\phi} - 1| \le 2|re^{i\phi} - 1|$ for any $\phi \in [0, 2\pi)$ and $r \ge 0$.

要证这个不等式，只用找到 r 让右边最小时也成立即可.

Pf (Claim):

$$|re^{i\phi} - 1|^2 = r^2 - 2r\cos\phi + 1 \triangleq g(r)$$

Thus,

$$g'(r) = 2r - 2\cos\phi$$

$$\Rightarrow \arg\min_{r \geq 0} |re^{i\phi} - 1|^2 = \begin{cases} 0 & \text{if } \phi \in [\frac{\pi}{2}, \frac{3\pi}{2}] \\ \cos\phi & \text{if } \phi \in [0, \frac{\pi}{2}] \cup [\frac{3}{2}\pi, 2\pi) \end{cases}$$

$$\Rightarrow \min_{r \geq 0} |re^{i\phi} - 1|^2 = \begin{cases} 1 & \text{if } \phi \in [\frac{\pi}{2}, \frac{3\pi}{2}], \\ \sin^2\phi & \text{o/w} \end{cases}$$

For $\phi \in [\frac{\pi}{2}, \frac{3\pi}{2}]$

$$|e^{i\phi} - 1| \leq 2 \leq 2|re^{i\phi} - 1|$$

o/w, $|e^{i\phi} - 1| = \sqrt{g(1)} = \sqrt{2(1 - \cos\phi)}$

$$= 2|\sin \phi/2| \quad \text{(half-angle formula)}$$

$$\leq 2|\sin\phi| \leq 2|re^{i\phi} - 1|$$

$\blacksquare$

Prop 4: Let $\{z^k\}$ be generated by PG, w/ $\alpha_k = \alpha \geq 0$.

Define $\theta_k = \arg\min_{\theta \in [0, 2\pi)} \| z^k - e^{i\theta} z^* \|_2$

$$\varepsilon_k = e^{-i\theta_k}(z^k - e^{i\theta_k} z^*)$$

$$\beta_k = 1 + \alpha + \frac{\alpha}{n}(z^*)^H \varepsilon^k$$

Then for any $r \in \mathbb{C}$, $k \geq 0$,

$$d_2(z^{k+1}, z^*) \leq 2\| r g^k - z^* \|_2$$

where,

$$g^k = \beta_k z^* + \left(I + \frac{\alpha}{n} \Delta\right) \varepsilon^k + \frac{\alpha}{n} \Delta z^*$$

Pf: By def'n,

$$w^k = \left(I + \frac{\alpha}{n} C\right) z^k$$

$$= \left(I + \frac{\alpha}{n}\left(z^*(z^*)^H + \Delta\right)\right) z^k$$

$$= e^{i\theta_k} \left(I + \frac{\alpha}{n}\left(z^*(z^*)^H + \Delta\right)\right)(z^* + \varepsilon^k)$$

来凑 g_k <u>EXERCISE</u>

$$= \left[\left(1 + \alpha + \frac{\alpha}{n}(z^*)^H \varepsilon^k\right) z^* + \left(I + \frac{\alpha}{n}\Delta\right)\varepsilon^k + \frac{\alpha}{n}\Delta z^*\right] e^{i\theta_k}$$

$$= g^k e^{i\theta_k}$$

$$\Rightarrow z^{k+1} = \frac{w^k}{|w^k|} = e^{i\theta_k} \frac{g^k}{|g^k|}$$, Hence, by Prop 3,

$$d_2(z^{k+1}, z^*) \leq \left\| \frac{g^k}{|g^k|} - z^* \right\|_2 = \left\| \frac{r g^k}{|r g^k|} - z^* \right\|_2$$

$$\leq 2\| r g^k - z^* \|_2$$

<u>EXERCISE</u> 加了 r 对于复数 同样成立

2019.3.25

继续:

Phase Sychronization

The generiative model: $C = z^*(z^*)^H + \Delta$

$$\hat{z} \in \underset{z \in \mathbb{T}^n}{\arg\max}\ z^H C z$$

$$\mathbb{T}^n = \{ z \in \mathbb{C}^n ; \ |z_i| = 1 \}$$

Projected Grad.

$$w^k \leftarrow z^k + \frac{\alpha_k}{n} C z^k$$

$$z^{k+1} \leftarrow \frac{w^k}{|w^k|}$$

让 $z^H C z$ 上升, 然后投影回 $\mathbb{T}^n$

$$\left(\frac{w}{|w|} \right)_j = \begin{cases} \frac{w_j}{|w_j|} & \text{if } w_j \neq 0 \\ 1 & \text{if } w_j = 0 \end{cases}$$

Spectral Estimator

- u: leading eigenvector of C
- $v_c \triangleq \frac{u}{|u|}$

Fact: $d_2(V_c, z^*) \leq \frac{8\|\Delta\|}{\sqrt{n}}$

does not depend on the distribution of the noise.

Thm: Suppose $\|\Delta\| \leq \frac{n}{16}$, $\alpha_k = \alpha \geq 2$, Then

$$d_2(z^{k+1}, z^*) \leq \mu^{k+1} d_2(z^0, z^*) + \frac{\nu}{1-\mu} \cdot \frac{8\|\Delta\|}{\sqrt{n}},$$

$$\mu = \frac{16(\alpha\|\Delta\| + n)}{(7\alpha + 8)n} < 1, \quad \nu = \frac{2\alpha}{7\alpha+8}$$

Recall: $d_2(z, w) = \min_{\theta \in [0, 4\pi)} \| z - e^{i\theta} w\|_2$

prop 4: Let $\{z^k\}$ be the iterates, Define

$$\theta_k = \arg\min_{\theta \in [0, 2\pi)} \| z^k - e^{i\theta} z^*\|_2$$

$$\varepsilon^k = e^{-i\theta_k}(z^k - e^{i\theta_k} z^*)$$

$$\beta_k = 1 + \alpha + \frac{\alpha}{n}(z^*)^H \varepsilon^k$$

Then, for any $y \in \mathbb{C}$, $k \geq 0$

$$d_2(z^{k+1}, z^*) \leq 2\| y g^k - z^*\|_2$$

where, $g^k = \beta_k z^* + (I + \frac{\alpha}{n}\Delta)\varepsilon^k + \frac{\alpha}{n}\Delta z^*$

Today:

Pf: (Thm 1)

We prove by induction, that

(i) $\|\varepsilon^k\|_2 \le \frac{\sqrt{n}}{2}$

(ii) $d_2(z^{k+1}, z^k) \le \mu d_2(z^k, z^*) + \nu \cdot \frac{8\|\Delta\|}{\sqrt{n}}$

unrole and calc the sum.

Base case: $k=0$

$$\|\varepsilon^0\|_2 = \| z^0 - e^{i\theta_0} z^* \|_2$$

by Assumption $\|\Delta\| \le n/16$

$$= d_2(z^0, z^*) \le \frac{8\|\Delta\|}{\sqrt{n}} \le \frac{\sqrt{n}}{2}$$

这里 $z^0 = V_c$

$$|\beta_0| \ge \left| 1 + \alpha + \frac{\alpha}{n} Re(z^{*H} \varepsilon^0) \right|$$

prop 4中, 扔掉虚部.

$$= \left| 1 + \alpha + \frac{\alpha}{2n} \left(\|z^* + \varepsilon^0\|_2^2 - \|z^*\|_2^2 - \|\varepsilon^0\|_2^2 \right) \right|$$

$$= \left| 1 + \alpha + \frac{\alpha}{2n} \left(\|z^0\|_2^2 - \|z^*\|_2^2 - \|\varepsilon^0\|_2^2 \right) \right|$$

模长相同, 都是 n.

$$\ge 1 + \frac{7\alpha}{8}$$

利用前面 ε_0 的 bound $\|\varepsilon^0\|_2 \le \frac{\sqrt{n}}{2}$

Take $y = \frac{1}{\beta_0}$ in prop 4:

$$d_2(z^1, z^*) \le 2\left\| \frac{1}{\beta_0}\left(I + \frac{\alpha}{n}\Delta\right)\varepsilon^0 + \frac{1}{\beta_0}\frac{\alpha}{n}\Delta z^* \right\|_2$$

$$\le 2\frac{1}{|\beta_0|}\left(\left\|\left(I + \frac{\alpha}{n}\Delta\right)\varepsilon^0\right\|_2 + \frac{\alpha}{n}\|\Delta z^*\|_2 \right)$$

$$\le \frac{16}{7\alpha+8}\left(\left(1 + \frac{\alpha}{n}\|\Delta\|\right) \underbrace{\|\varepsilon^0\|_2}_{\overset{\|}{d_2(z^0, z^*)}} + \frac{\alpha}{\sqrt{n}}\|\Delta\| \right)$$

$$= \mu\, d_2(z^0, z^*) + \frac{\nu}{1-\mu}\frac{8\|\Delta\|}{\sqrt{n}}$$

Inductive Step

$$\|\varepsilon^{k+1}\|_2 = d_2(z^{k+1}, z^*)$$

$$\le \mu \cdot \underbrace{d_2(z^k, z^*)}_{\overset{\|}{\|\varepsilon^k\|_2 \le \frac{\sqrt{n}}{2}}} + \nu\frac{8\|\Delta\|}{\sqrt{n}}$$

$$\le \frac{8(\alpha\|\Delta\| + n)}{(7\alpha+8)\sqrt{n}} + \frac{16\alpha}{7\alpha+8}\cdot\frac{\|\Delta\|}{\sqrt{n}}$$

$$\le \frac{\sqrt{n}}{2}$$

$$|\beta_{k+1}| \ge 1 + \frac{7\alpha}{8}$$

$$\Rightarrow d_2(z^{k+1}, z^*) \le \mu\, d_2(z^{k+1}, z^*) + \nu\frac{8\|\Delta\|}{\sqrt{n}}$$

by Prop 4.

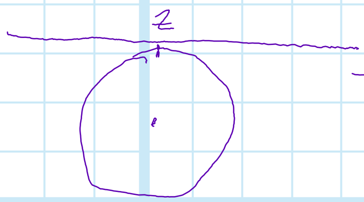
Optimization aspect

First-order optimality condition

obs: $\mathbb{T}$ is a smooth manifold.

For every point, a neighbour can be mapped into a Euclid-space.

It locally looks like a Euclidian space



$T_z\mathbb{T}$ Tangent space.

$$T_z\mathbb{T} = \{ w \in \mathbb{C} : Re(z\bar{w}) = 0 \}$$

例: 若 $z = e^{i\theta}$,

$$T_z\mathbb{T} \ni w = re^{i(\theta \pm \frac{\pi}{2})}$$

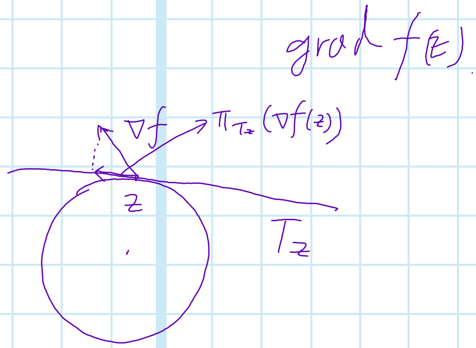
$$z\bar{w} = re^{i(\pm\frac{\pi}{2})}$$

$$\mathbb{T}^n = \underbrace{\mathbb{T} \times \cdots \times \mathbb{T}}_{n\,\mathbb{T}} \Rightarrow T_z\mathbb{T}^n = \{ w \in \mathbb{C}^n : Re(z_i\bar{w}_i) = 0, \forall i \}$$

This is a linear subspace.

Def: z is first order critical if

$$\Pi_{T_z}(\nabla f(z))=0.$$



对比无约束的FOC:

$$\min_z f(z)$$

FOC: $\nabla f(z)=0$,

现在变成投影 $\Pi_{T_z}(\nabla f(z))=0$.

$\Pi_{T_z}(\nabla f(z))$ is called the <u>Riemanian gradient</u> of f at z and is denoted by $\text{grad} f(z)$

$\therefore$ FOC: $\quad \text{grad} f(z)=0$.

For $\mathbb{T}^n$: (<u>EXERCISE</u>)

$$\Pi_{T_z\mathbb{T}^n}(w)= w-\text{Diag}(\text{Re}(z_j\bar{w}_j))z$$

$\therefore$ In original problem:

$$f(z)=-z^H C z$$

$$\Rightarrow \text{grad} f(z)=\Pi_{T_z\mathbb{T}^n}(\nabla f(z))$$

$$=2\left(Cz-\text{Diag}(\text{Re}((Cz)_j\bar{z}_j))z\right)$$

Define $S(z) = \mathrm{Diag}(\mathrm{Re}((Cz)_j \bar{z}_j)) - C$

$\Rightarrow$ 1^{st} order optimality condition is

$$S(z)\, z = 0.$$

上面定义了黎曼梯度，下面来定义 Riemannian Hessian.

对于普通的函数 f，若知道 $\nabla^2 f(z) w$，对所有 w，便能确定 $\nabla^2 f(z)$

Riemanian Hessian: Projection of the directional derivative of the Riemannian grad onto the tangent space.

$$\mathrm{Hess}\, f(z)(w) = \Pi_{T_z \mathbb{T}^n}\left(D\, \mathrm{grad}\, f(z)[w]\right)$$

$$= \Pi_{T_z \mathbb{T}^n}\left(2 S(z) w\right)$$

这里是隐式定义的 Hessian，而是 define its action on each direction

Def: 2^{nd} optimality condition

对比普通 Euclidian 的情况：

$$\min_z f(z)$$

$$\nabla f(z) = 0$$

$$\nabla^2 f(z) \succeq 0$$

mainfold 情况下：

$$w^H(\mathrm{Hess}\, f(z))(w)$$

$$= 2w^H S(z)\, w \geq 0 \quad \forall w \in T_z \Pi^n$$

$$w^H \Pi_{T_z \Pi^n}(2S(z)w)$$

投影 自伴算子

$$= w^H P(2S(z)w) \qquad \because P \text{ is self-adjoint}$$

$$= (Pw)^H(2S(z)w) \qquad \therefore P = P^H$$

2019.3.26

$$\hat{z} \in \underset{z \in \mathbb{T}^n}{\operatorname{argmax}} \{ f(z) \triangleq z^H C z \}$$

C is generated by:
$C = z^* z^{*H} + \Delta$

$$\mathbb{T}^n = \{ z \in \mathbb{C}^n : |z_i| = 1 \ \forall i \}$$

$$\operatorname{grad} f(z) = 2[C - \operatorname{Diag}(\operatorname{Re}((Cz)_j \bar{z}_j))] z$$

$$\operatorname{Hess} f(z)(w) = \Pi_{T_z \mathbb{T}^n}(2S(z)w),$$

where $S(z) = \operatorname{Diag}(\operatorname{Re}((Cz)_j \bar{z}_j)) - C$

Optimality Conditions:

1^{st} order: $S(z) z = 0$

2^{nd} order: $w^H S(z) w \geq 0, \ \forall w \in T_z \mathbb{T}^n$

where,

$$T_z \mathbb{T}^n = \{ w \in \mathbb{C}^n, \operatorname{Re}(z_j \bar{w}_j) = 0, \ \forall j \}$$

Fact: $\hat{z}$ satisfies the 1^{st} + 2^{nd} order conditions

Prop: (i) $z \in \mathbb{T}^n$ satisfies the 1^{st} order condition iff

$(Cz)_j \bar{z}_j$ is real $\forall j$

(ii) If $\mathrm{diag}(C)\geq 0$ and $z\in\mathbb{T}^n$ satisfies $1^{st}+2^{nd}$ order condition, then

$$(Cz)_j\,\bar{z}_j\geq 0,\ \forall j$$

$$\Rightarrow (Cz)_j\,\bar{z}_j=|(Cz)_j|$$

Pf: (i)

$$0=S(z)z$$

$$\Leftrightarrow \mathrm{Re}\left((Cz)_j\bar{z}_j\right)z_j=(Cz)_j \quad \text{同乘 } \bar{z}_j\text{, 用 } |z_j|=z_j\bar{z}_j$$

$$\Leftrightarrow \mathrm{Re}\left((Cz)_j\bar{z}_j\right)=(Cz)_j\,\bar{z}_j$$

(ii) 观察到我们对 2^{nd} order 定义是任取 w, 使得 $w^H S(z)\,w\geq 0$ 我们可以取 $w=e_j$

$$0\leq e_j^H S(z)\,e_j$$

$$=\mathrm{Re}\left((Cz)_j\bar{z}_j\right)-C_{jj}$$

$$=(Cz)_j\bar{z}_j-C_{jj}$$

还需验证 $e_j\overset{?}{\in}T_z\mathbb{T}^n$

Verify $\mathrm{Re}(z_j)\overset{?}{=}0$

This cannot be verified.

那么 How to fix this argument?

需满足如下条件:

(i) $\mathrm{Re}(z_j\bar{w}_j)=0$

(ii) $w_j = \eta \mathbb{1} \Rightarrow (\eta e_j)^H S(z) (\eta e_j)$

$$= (Cz)_j \bar{z}_j - C_{jj}$$

$$\Rightarrow |\eta| = 1$$

$$\eta = i z_j$$

Let $w = (i z_j) e_j \in T_z \mathbb{T}^n \quad \forall j$

Then,

$$0 \le w^H S(z) w = (Cz)_j \bar{z}_j - \underbrace{C_{jj}}_{\ge 0} \Rightarrow (Cz)_j \bar{z}_j \ge 0$$

Corollary:

consider $\hat{z}$, It satisfies both 1^{st} + 2^{nd} order condition. If $\mathrm{diag}(C) \ge 0$, then

$$\boxed{\underbrace{\left(\mathrm{Diag}(|Cz|) - C\right)}_{S(\hat{z})} \hat{z} = 0}$$

这个结论是在 $\mathrm{diag}(c) \ge 0$ 下成立的。能否删掉这个条件

Obs:

(i) $S(z) = \mathrm{Diag}[\mathrm{Re}((Cz)_j \bar{z}_j)] - C$

$$C \to C + \frac{n}{\alpha} I \triangleq \tilde{C}$$

$$(\tilde{C} z)_j = (Cz)_j + \frac{n}{\alpha} z_j$$

$\therefore \mathrm{Diag}[\mathrm{Re}((\tilde{C} z)_j \bar{z}_j)] = \mathrm{Diag}[\mathrm{Re}((Cz)_j \bar{z}_j + \frac{n}{\alpha})]$

$$= \mathrm{Diag}[\mathrm{Re}((Cz)_j \bar{z}_j)] + \frac{n}{\alpha} I$$

$$\therefore \mathrm{Diag}[\mathrm{Re}((\tilde{C}z)_j \bar{z}_j)] - \tilde{C}$$

$$= \mathrm{Diag}[\mathrm{Re}((Cz)_j \bar{z}_j + \tfrac{n}{\alpha})] - \tilde{C}$$

$$= \mathrm{Diag}[\mathrm{Re}((Cz)_j \bar{z}_j)] + \cancel{\frac{n}{\alpha} I} - (C + \cancel{\tfrac{n}{\alpha} I})$$

(ii) For $\hat{z}$,

$$\boxed{\left[\mathrm{Diag}(|\tilde{C}\hat{z}|) - \tilde{C}\right]\hat{z} = 0}$$

Convergence Analysis of PG

Alg: PG

$$w^k \leftarrow z^k + \frac{\alpha}{n} C z^k$$

$$z^{k+1} \leftarrow \frac{w^k}{|w^k|}$$

High-Level Ideas:

① Error Bound (Depend only on the problem, not related to Alg)

之前作业题.

$$\mathrm{dist}(x, \mathcal{X}) \le \mu \cdot \|R(x)\|_2$$

$$\|R(x)\| \leq \mathrm{dist}(0, \partial F(x))$$

② Alg properties.

(i) sufficient ascent

(ii) cost-to-go — how far away from the optimal value

(iii) Safeguard. — if residual is 0 we should stop.

在这个问题中, Error Bound 和之前不一样.

下面来看每一条.

① Error Bound. (e.g. $\mathrm{dist}(x, X) \leq \mu \|R(x)\|_2$)

如何确定 LHS 和 RHS — what to put here.

LHS:

$$d_2(z, \hat{z})$$

RHS: 在 Corollary (ii) 中, 如果 $\hat{z}$ 是最优, 那么等于0, 如果等于一个很小的数, 那么我们是否足够接近?

Candidate:

$$\Sigma(z) = \mathrm{Diag}(|\tilde{C} z|) - \tilde{C}$$

Define $\rho(z) = \|\Sigma(z) z\|_2$

Note: $\rho(\hat{z}) = 0$

注意，我们已知 $\hat{z}$ is optimal $\Rightarrow \rho(\hat{z})=0$
但反过来不知道.

$$\boxed{d_2(z,\hat{z}) \le c\cdot \rho(z)}$$ ← 我们尝试证这个.

② Alg properties.

(a) (Sufficient Ascent)

$$f(z^{k+1}) - f(z^k) \ge a_0 \|z^{k+1} - z^k\|_2^2$$

(b) (Cost-to-go) How far you are

$$f(\hat{z}) - f(z^k) \le a_1\, d_2(z^k, \hat{z})^2$$

(c) (Safeguard)

$$\rho(z^k) \le a_2 \|z^{k+1} - z^k\|_2$$

Using these properties,

$$f(\hat{z}) - f(z^{k+1}) = [f(\hat{z}) - f(z^k)] - [f(z^{k+1}) - f(z^k)]$$

(cost-to-go) $\le a_1 \cdot d_2(z^k, \hat{z})^2 - [f(z^{k+1}) - f(z^k)]$

(EB) $\le a' \cdot \rho(z^k)^2 - [f(z^{k+1}) - f(z^k)]$

(safeguard) $\le a'' \cdot \|z^{k+1} - z^k\|_2^2 - [f(z^{k+1}) - f(z^k)]$

$$\le \underbrace{(a''' - 1)}_{>0}[f(z^{k+1}) - f(\hat{z}) + f(\hat{z}) - f(z^k)]$$

Rearange

$$a''\left[f(\hat{z})-f(z^{k+1})\right] \le (a''-1)\left[f(\hat{z})-f(z^{k})\right]$$

$$\left[f(\hat{z})-f(z^{k+1})\right] \le \frac{a''-1}{a''}\left[f(\hat{z})-f(z^{k})\right]$$

$\Rightarrow$ linear convergence.

$\exists\ \lambda \in (0,1)$:

$$f(\hat{z})-f(z^{k+1}) \le \lambda^{k}\left(f(\hat{z})-f(z^{0})\right)$$

How about the iterates.

$$\begin{aligned}
d_2(z^k,\hat{z})^2 &\le c\cdot \rho(z^k)^2 && \text{(EB)}\\
&\le c'\|z^{k+1}-z^k\|_2^2 && \text{(Safeguard)}\\
&\le c''\left(f(z^{k+1})-f(z^k)\right) && \text{(Sufficient ascent)}\\
&\le c''\left(f(\hat{z})-f(z^k)\right)\\
&\le \lambda^{k-1}\left[f(\hat{z})-f(z^k)\right]\cdot c'''
\end{aligned}$$

注意这是 $d_2(z^k,\hat{z})^2$ $\therefore$ $d_2(z^k,\hat{z})$ 的收敛速率

大约是 $\lambda^{\frac{k-1}{2}}$，比 f 的速度慢.

这个分析和之前类似，唯一区别是

这里是非凸的.

2019.4.2

Generative model:

$$C = z^*(z^*)^H + \Delta$$

Estimator:

$$\hat{z} \in \underset{z \in \mathbb{T}^n}{\operatorname{argmax}} \; z^H C z$$

PGD

$$w^k \leftarrow \left(I + \frac{\alpha}{n} C\right) z^k$$

$$z^{k+1} \leftarrow \frac{w^k}{|w^k|}$$

Def: $\tilde{C} = C + \frac{n}{\alpha} I$

$$\Rightarrow \; w^k = \frac{\alpha}{n} \tilde{C} z^k$$

$$\Rightarrow \; z^{k+1} = \frac{\tilde{C} z^k}{|\tilde{C} z^k|}$$

也叫 (GPM): Generalized Power Method

Facts:

(1) $d_2(\hat{z}, z^*) \leq \frac{4\|\Delta\|}{\sqrt{n}}$

(2) $d_2(V_c, z^*) \leq \frac{8\|\Delta\|}{\sqrt{n}}$

(3) define

$$\Sigma(z) = \mathrm{Diag}(|\tilde{C} z|) - \tilde{C}$$

Then

$$\Sigma(\hat{z})\hat{z} = 0 \quad \left(1^{st} \text{ opt.} + 2^{nd} \text{ cond}\right)$$

Define residual measure.

$$\rho(z) = \|\Sigma(z) z\|_2$$

<u>Error Bound</u>

Thm. For any $z \in \mathbb{T}^n$, satisfying $d_2(z, z^*) \leq \frac{\sqrt{n}}{2}$

+ additional assumptions (on $\|\Delta\|$, α, ...)

and any opt soln $\hat{z}$

$$d_2(z, \hat{z}) \leq \frac{8}{n} \rho(z)$$

这是一个 local error bound，但是如果起点在这个 region 内，整个序列都在。

<u>Pf</u>:

$$\rho(z) = \|\Sigma(z) z\|_2 \geq \underbrace{\|\Sigma(\hat{z}) z\|_2}_{\text{need LB}} - \underbrace{\|(\Sigma(z) - \Sigma(\hat{z})) z\|_2}_{\text{need UB}}$$

通常 UB 更好算

UB:

看如何把 d_2 关联起来.

$$\|(\Sigma(z) - \Sigma(\hat{z})) z\|_2$$

$$= \left\| \left[\mathrm{Diag}(|\tilde{C} z|) - \mathrm{Diag}(|\tilde{C} \hat{z}|) \right] z \right\|_2$$

$$= \left(\sum_j \left| \left(|(\tilde{C} z)_j| - |(\tilde{C} \hat{z})_j| \right) z_j \right|^2 \right)^{1/2}$$

z_j 模长始终为1, 可以扔掉

加了这项, 模长不变

$$= \left\| |\tilde{C} e^{-i\hat{\theta}} z| - |\tilde{C} \hat{z}| \right\|_2 \qquad \hat{\theta} = \arg\min_{\theta \in [0, 2\pi]} \|z - e^{i\theta} \hat{z}\|_2$$

$$\leq \| \tilde{C} (e^{-i\hat{\theta}} z - \hat{z}) \|_2 \qquad \text{(tri-inequality)}$$

$$\leq \|z^* (z^*)^H (e^{-i\hat{\theta}} z - \hat{z})\|_2 + \|\Delta (e^{-i\hat{\theta}} z - \hat{z})\|_2 + \frac{n}{\alpha} \|e^{-i\hat{\theta}} z - \hat{z}\|_2 \qquad (\because \tilde{C} = C + \frac{n}{\alpha} I)$$

$$\leq \sqrt{n} \cdot |(z^*)^H (e^{-i\hat{\theta}} z - \hat{z})| + (\|\Delta\| + \frac{n}{\alpha}) d_2(z, \hat{z})$$

$$|(z^*)^H (e^{-i\hat{\theta}} z - \hat{z})| \leq \left| (z^* - e^{-i\theta^*} \hat{z})^H (e^{-i\hat{\theta}} z - \hat{z}) \right| + \left| (e^{-i\theta^*} \hat{z})^H (e^{-i\hat{\theta}} z - \hat{z}) \right|$$

$$\left(\theta^* = \arg\min_{\theta} \|\hat{z} - e^{i\theta} z^*\|_2 \right)$$

$$\le \frac{4\|\Delta\|}{\sqrt{n}} \cdot d_2(z,\hat z) + \left|(e^{-i\theta^*}\hat z)^H(e^{-i\hat\theta} z - \hat z)\right|$$

Note: $d_2(z,\hat z)^2 = \| e^{-i\hat\theta} z - \hat z\|_2^2$

$= 2\left(n - |\hat z^H z|\right)$

$-\max_\theta 2Re\left((e^{-i\theta} z)^H \hat z\right)$
当 $e^{-i\theta} z$ is real, get max

$$\le \frac{4\|\Delta\|}{\sqrt{n}} \cdot d_2(z,\hat z) + \left| e^{i\theta^*}\left(|\hat z^H z| - n\right)\right|$$

← 取模长可以隐掉

$$\le \frac{4\|\Delta\|}{\sqrt{n}} d_2(z,\hat z) + \frac{1}{2} d_2(z,\hat z)^2$$

∴ UB:

$$\|(\Sigma(z) - \Sigma(\hat z))z\|_2$$

$$\le \left(5\|\Delta\| + \frac{n}{\alpha}\right) d_2(z,\hat z) + \frac{\sqrt{n}}{2} d_2(z,\hat z)^2$$

下面来看 $\|\Sigma(z) z\|_2$ 的 LB，

如果 z 在 $\Sigma(z)$ 的 null space 中，那么就是 0，

但观察到 $\Sigma(z)$ 由 z 决定，z 是否能证其不会在 Null space 中。

Want: $\Sigma(\hat z) \succeq 0$, $\Sigma(\hat z)$ pd on certain subspace

∵ $\Sigma(\hat z)\hat z = 0$ (opt condition) ∴ suppose is $(\mathrm{span}(\hat z)^\perp)$

Define: $\hat u = \left(I - \frac{1}{n}\hat z \hat z^H\right)\left(e^{-i\hat\theta} z - \hat z\right)$

be projection of $e^{-i\hat\theta}z - \hat z$ onto $\text{span}(\hat z)^\perp$

Note: ① $\hat u^H \hat z = 0$ ∴ $\hat u$ is the projector

② $\|\Sigma(\hat z) z\|_2 = \|\Sigma(\hat z)(e^{-i\hat\theta}z - \hat z)\|_2$

$= \|\Sigma(\hat z)\hat u\|_2$ ← 看这个是否 LB by sth positive

注意到 $\lambda_{min}(\Sigma(\hat z)) = \min_{\hat u} \dfrac{\hat u^H \Sigma(\hat z)\hat u}{\|\hat u\|^2}$

那么 $\hat u^H \Sigma(\hat z)\hat u \geq c\|\hat u\|_2$?

$\hat u^H \Sigma(\hat z)\hat u = \hat u^H[\text{Diag}(|\tilde C \hat z|) - \tilde C]\hat u$

$= \hat u^H[\text{Diag}(|C\hat z|) - C]\hat u$ $\quad |(\tilde C \hat z)_j| = (\tilde C \hat z)_j \bar{\hat z}_j$

(prop 5 of prev notes)

$= \sum_j |(C\hat z)_j| \cdot |\hat u_j|^2 - |(z^*)^H \hat u|^2 - \hat u^H \Delta \hat u$

$\geq (|(z^*)^H \hat z| - \|\Delta \hat z\|_\infty)\|\hat u\|_2^2$ $\quad C\hat z = z^*(z^*)^H\hat z + \Delta\hat z$

$- |\hat u^H(z^* - e^{-i\theta^*}\hat z)|^2 - \|\Delta\| \cdot \|\hat u\|_2^2$

∵ $\hat u^H \hat z = 0$ ∴ 可以随便加系数

$\geq [n - \|\Delta \hat z\|_\infty - \frac{3}{2}d_2(\hat z, z^*)^2 - \|\Delta\|]\|\hat u\|_2^2$

$\geq (n - \|\Delta \hat z\|_\infty - \|\Delta\| - \frac{24\|\Delta\|^2}{n})\|\hat u\|_2^2$

下面利用这个来分析 Bound.

$$\|\hat{u}\|_2 \geq \| e^{-i\hat{\theta}} z - \hat{z} \|_2 - \frac{1}{n} \| \hat{z}\hat{z}^H (e^{-i\hat{\theta}} z - \hat{z}) \|_2$$

$$= d_2(z, \hat{z}) - \frac{1}{\sqrt{n}} \left| \hat{z}^H (e^{-i\hat{\theta}} z - \hat{z}) \right|$$

$$= d_2(z, \hat{z}) - \frac{1}{2\sqrt{n}} d_2(z, \hat{z})^2$$

Hence,

$$\| \Sigma(\hat{z}) z \| = \| \Sigma(\hat{z}) \hat{u} \|_2 \geq \lambda_{min}(\Sigma(\hat{z})) \|\hat{u}\|_2$$

$$\geq \left[n - \|\Delta \hat{z}\|_\infty - \|\Delta\| - \frac{24\|\Delta\|^2}{n} \right] \cdot \left(d_2(z, \hat{z}) - \frac{1}{2\sqrt{n}} d_2(z, \hat{z})^2 \right)$$

Note: $d_2(z, \hat{z}) \leq d_2(z, z^*) + d_2(\hat{z}, z^*)$

$$\leq \frac{\sqrt{n}}{2} + \frac{4\|\Delta\|}{\sqrt{n}}$$

$$\Rightarrow d_2(z, \hat{z})^2 \leq \left(\frac{\sqrt{n}}{2} + \frac{4\|\Delta\|}{\sqrt{n}} \right) \cdot d_2(z, \hat{z})$$

Note: $\|\Delta \hat{z}\|_\infty \leq \|\Delta z^*\|_\infty + \|\Delta (e^{-i\theta^*} \hat{z} - z^*)\|_\infty$

$$\leq \|\Delta z^*\|_\infty + \|\Delta\| \cdot d_2(\hat{z}, z^*)$$

$$\leq \|\Delta z^*\|_\infty + \frac{4\|\Delta\|^2}{\sqrt{n}}$$

∴ Assumptions: 加个 Assumptions 来保证结论.

$$\|\Delta\| = O(n^{3/4})$$

$$\|\Delta z^*\|_\infty = O(n)$$

2019.4.8.

今天换一个问题.

信号处理问题.

MIMO Detection

$$y = Hx^* + v \quad \text{(Generative model)}$$

$H \in \mathbb{C}^{m \times n}$

m: # of output

n: # of input

Channel matrix

$x^* \in \mathbb{C}^n$: vector of transmitted symbols

$y \in \mathbb{C}^m$: received vector

$v \in \mathbb{C}^m$: additive noise

Typically,

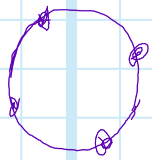
each x_i^* draw from a discate discate constellation $\mathcal{S}$

↖ 就是星座的意思

Two example:

(1) M-ary Phase shift keeping (MPSK)

$$\mathcal{S}_M = \left\{ \exp\left(\frac{2\pi i k}{M}\right),\ k = 0, 1, \cdots, M-1 \right\}$$



$M=4$

(2) $(4u^2)$-Quadraturm Amplitude Modulation (QAM)

$$Q_u = \left\{ z \in \mathbb{C} : Re(z), Im(z) = \pm 1, \pm 3, \dots, \pm(z-1) \right\}$$

v_i: Complex Gaussian $\mathcal{CN}(0, \sigma_i^2)$ circular symmetry

Problem:

Given y and H, goal is recover x^*

ML: $\hat{x} \in \underset{x \in \mathcal{S}^n}{\arg\min} \|y - Hx\|_2^2$ 与线性回归相比.

x有了离散约束

和Phase Sync相同, 有两种方法:

Approach 1: SDR

在(1)中, 直接用 $\|x\|=1$ 然后rounding

(2) 模长不一样怎么做?

Approach 2: PG projected gradient

这里我们不但是投影到非凸集合上，
而是投到离散集合上.

<u>PG:</u>

$$\tilde{w}^k \leftarrow 2H^H(Hx^k - y)$$

$$x^{k+1} \leftarrow \Pi_{\mathcal{S}^n}\left(x^k - \frac{\alpha_k}{m}\tilde{w}^k\right)$$

我们为什么能这么做:

Rmk: ① Projection step is easy

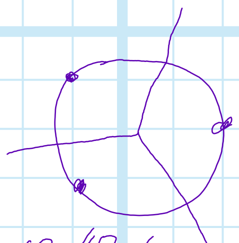
— projection can be done coordinate-wise

比如，只有四个点，那就全算出距离，然后取最小的那个.

Then, ① brute-force computation

② Voronoi diagram (算最近邻点)

例:



3-PSK

② 在传统的问题中, convergence很重要, 可以无限.
但在离散问题中, 不存在这种情况. (不能无限)

want: Finite convergence

preferrably, less than $|\mathcal{S}^n|$ iterations.

至少比暴力搜要简单些.

Thm: Let $c = \dfrac{4}{\min\limits_{s \neq s' \in \mathcal{S}} |s - s'|} < \infty$ 任两个点不能无限接近

Suppose:

(i) $\dfrac{2\alpha_k}{m} \| H^H v \|_\infty < \dfrac{1}{c}$ control noise

(ii) $\| I - \dfrac{2\alpha_k}{m} H^H H \| \leq \beta < \dfrac{1}{4}$ conditioning of the channel

Then,

$$\| x^{k+1} - x^* \|_2 \leq 4\beta \| x^k - x^* \|_2 \leq \cdots \leq (4\beta)^{k+1} \| x^0 - x^* \|_2 \leq \frac{c}{2}$$

这里没有限制起始点，只要控制了 noise 和 conditioning，就有 linear convergence

注意，这也意味着迭代次数是有限的，imply finite convergence. ∵ 离散空间是有距离的，不能两个点距离太小。。。

In particular, after at most

$$k^* = \left\lceil \log_{4\beta} \left(\frac{2}{c \| x^0 - x^* \|_2} \right) \right\rceil \text{ iterations,}$$

$$x^k = x^* \quad \forall k \geq k^*$$

Proof: By defin of PG

$$z^k \triangleq x^k - \frac{\alpha_k}{m}\tilde{w}^k$$ (what we are going to project)

$$= x^k - \frac{2\alpha_k}{m} H^H (H x^k - y)$$

$$= x^* + \underbrace{\left(I - \frac{2\alpha_k}{m} H^H H\right)}_{\text{用Assumption来bound}}(x^k - x^*) + \underbrace{\frac{2\alpha_k}{m} H^H v}_{\text{用Assumption来bound}}$$

$$\left[\because H^H y = H^H (H x^* + v)\right.$$

and $x^{k+1} = \Pi_{\mathcal{S}^n}(z^k)$

Let $w^k = \left(I - \frac{2\alpha_k}{m} H^H H\right)(x^k - x^*)$

$$J_k = \left\{ j : |w_j^k| \geq \frac{1}{c} \right\}$$ ← key tricks

这里 J_k 让 w_j^k 足够小，因此 z^k 就变成了 x^* 加两个小项，一投影就是 x^* 了。

Note: Coordinate-wise

$$z_i^* = x_i^* + w_i^k + \frac{2\alpha_k}{m}(H^H v)_i$$

$$\Rightarrow |z_i^k - x_i^*| \leq |w_i^k| + \frac{2\alpha_k}{m}|(H^H v)_i|$$

$$\left(\leq \frac{1}{c} + \frac{1}{c} = \frac{2}{c} \quad \text{if } i \notin J_k\right)$$

$$\Rightarrow x_i^* = \Pi_{\mathcal{S}}(z_i^k) = x_i \quad - - - -$$

<u>Fact</u>:

$x_i^{k+1} = x_i^* \quad \forall i \notin J_k$ 只要不在 I 中，就不变这一项.

prop: Let $z \in \mathbb{C}^n$ and $x \in \mathcal{S}^n$ be given, where $\mathcal{S} = \mathcal{S}_m$ or $\mathbb{C}_u$.

Then, $\|\Pi_{\mathcal{S}^n}(z) - x\|_2 \le 2\|z - x\|_2$

先假设成立.

Assume this. ∵只要不在 J_k 中的维，取值相等

$$\|x^{k+1} - x^*\|_2 = \|x^{k+1}_{J_k} - x^*_{J_k}\|_2 \le 2\|z^*_{J_k} - x^*_{J_k}\|_2$$

由那个 prop

$$= 2\left\|w^k_{J_k} + \left(\frac{2\alpha_k}{m} H^H v\right)_{J_k}\right\|_2.$$

$$\left\|\frac{2\alpha_k}{m} H^H_v\right\|_\infty < \frac{1}{c} \le |w_j^k| \quad \text{for } j \in J_k$$

Assumption (i)

$$\le 4\|w^k_{J_k}\|_2 \le 4\|w^k\|_2$$

$$\le 4\beta\|x^k - x^*\|_2 \quad \text{(Assumption 2)}$$

以上就证明了 Thm，下面来看 prop.

Pf of Prop. Focus on $\mathcal{S}_M$

① $\displaystyle \min_{r\geq 0} |re^{i\phi}-1|^2 = \begin{cases} 1 & \text{if } \phi \in [\frac{\pi}{2}, \frac{3\pi}{2}] \\ \sin^2\phi & \text{if } \phi \in [0, \frac{\pi}{2}) \cup (\frac{3\pi}{2}, 2\pi) \end{cases}$

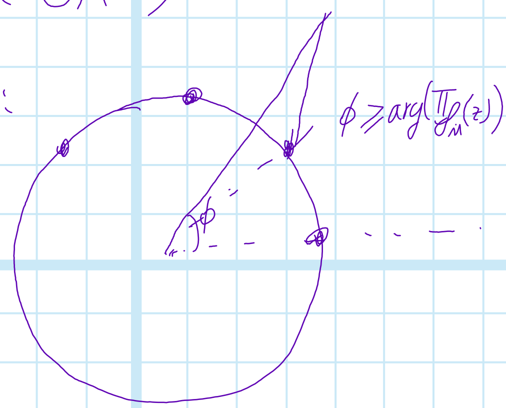
② Let: $z = re^{i\phi}$ if $\phi \in [\frac{\pi}{2}, \frac{3\pi}{2}]$.

$$|\Pi_{\mathcal{S}_M}(z) - 1| \leq 2 \leq 2|re^{i\phi} - 1|$$

③ Say $\phi \in [0, \frac{\pi}{2}]$ (a) If $\phi \geq \arg(\Pi_{\mathcal{S}_M}(z))$

$$|\Pi_{\mathcal{S}_M}(z) - 1| \leq |e^{i\phi} - 1|$$

如图:



$\leq 2|\sin\phi|$

← Half-angle formula.

$\leq 2|re^{i\phi} - 1|$

(b) $\phi < \arg(\Pi_{\mathcal{S}_M}(z))$

和(a)一样, 我们需要ϕ的LB.

Obs $\phi \geq \frac{1}{2}\arg(\Pi_{\mathcal{S}_M}(z))$

$$|\Pi_{\mathcal{S}_M}(z) - 1| = 2\left|\sin\frac{\arg(\Pi_{\mathcal{S}_M}(z))}{2}\right|$$

$\Pi_{\mathcal{S}_M}(z) = e^{i|\arg(\Pi_{\mathcal{S}_M}(z))|}$

$\leq 2|\sin\phi| \leq \cdots$

下面来看那两个Assumption能否容易满足.

Thm 2: Suppose

(i) entries of H iid standard complex Gaussian (i.e. $H_{ij} \sim \mathcal{CN}(0,1)$)

complex normal

$$H_{ij} = g_{ij}^R + i g_{ij}^I$$

$$g_{ij}^R, g_{ij}^I \sim \mathcal{N}(0, 1/2)$$

(ii) entries of v have variance

$$\sigma_v^2 \le \frac{m}{4c^2 \log n}$$

(iii) aspect ratio: $\gamma \triangleq \frac{m}{n} \geq \frac{20}{\beta^2} > 1$

(因此输出比输入多)

(iv) step size $\alpha_k = \frac{1}{2}$

Then, whp, (i)+(ii) in Thm hold

with high probability

Comments:

$\gamma \le 1$ 是否可能存在, open problem.

下节将主要看(ii).

2019.4.9

上节课的 Assumption 2

$$\| I - \frac{1}{m} H^H H \| \le \beta < \frac{1}{4}$$

spectral norm

eigen values:

$$\max_i \left| I - \frac{1}{m} \lambda_i (H^H H) \right|$$

可能取:

$\sigma_{max}(H)$ $\quad \left| 1 - \frac{1}{m} \lambda_{max}(H^H H) \right|$

$\sigma_{min}(H)$ $\quad \left| 1 - \frac{1}{m} \lambda_{min}(H^H H) \right|$

下面来研究随机矩阵的奇异值.

Larggest singular value of Gaussian Random Matrix

$A \in \mathbb{R}^{m \times n}$ $\quad m \ge n$

$$A_{ij} \sim \mathcal{N}(0,1)$$

$$\sigma_{max}(A) = \sup_{\substack{\|u\|_2=1 \\ \|v\|_2=1}} u^T A v \quad \text{(Courant-Fischer)}$$

Idea: Fix u, v, Consider

$$u^T A v = \sum_{ij} u_i v_j A_{ij}$$

$$= \sum_i u_i \left(\sum_j v_j A_{ij} \right) \longrightarrow \sim \mathcal{N}\left(0, \sum_j v_j^2\right) = \mathcal{N}(0,1)$$

$$\sim \sum_i u_i g_i \quad g_i \sim \mathcal{N}(0,1)$$

$$\sim \mathcal{N}\left(0, \sum_i u_i^2\right) = \mathcal{N}(0,1)$$

$\therefore u^T A v \sim \mathcal{N}(0,1)$ $\therefore$ 可以 bound

<u>Fact:</u> $\Pr(|u^T A v| \geq t) \leq \frac{1}{\sqrt{2\pi}\, t} \exp(-t^2/2)$

[这里还不能用 union bound，$\because u\,v$ 是无限a个，不能取完 u,v 来取最小。

<u>obs:</u> We cannot take union bound over $S^{m-1} \times S^{n-1}$ ($S^{m-1} = \{x \in \mathbb{R}^m : \|x\|_2 = 1\}$)

<u>Idea:</u> Discretize $S^{m-1} \times S^{n-1}$

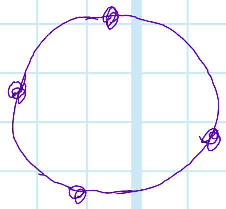
(Find representative points and take union bount over this finite subset. We should control the error of the discretiz)

$\longrightarrow$ ε-net

<u>Def</u> A set $N \subseteq S^{n-1}$ is an ε-net if $\forall u \in S^{n-1}$ $\exists v \in N$, s.t. $\|u - v\|_2 \leq \varepsilon$

就是集合中任一点到 ε-net 的距离都小于 ε

The main question is how large $|N|$ will be.

$S^1 =$ 

Prop: Let $\varepsilon > 0$ be given, There exists an ε-Net of size $|N| \le 2n(1+2/\varepsilon)^{n-1}$

看这个怎么用在之前那个奇异值上.

Thm: Let M be δ-net of S^{m-1} and N be an ε-net of S^{n-1}, Them,

$$\sigma_{max}(A) \le \frac{1}{(1-\varepsilon)(1-\delta)} \sup_{u \in M, v \in N} |u^T A v|$$

Here is similar to Corant-Fischer, Difference is M, N is finite.

Pf Take $z \in S^{n-1}$, We can write $z = v + h$, where $v \in N$, $\|h\|_2 \le \varepsilon$

Note: $\sigma_{max}(A) = \sup_{\substack{u \in S^{m-1} \\ v \in S^{n-1}}} u^T A v = \sup_{v \in S^{n-1}} \|Av\|_2$

$$\therefore \|Az\|_2 \le \|Av\|_2 + \|Ah\|_2 \le \|Av\| + \varepsilon \underbrace{\|A\|}_{\sigma_{max}(A)}$$

$\therefore$ 两边同取sup

$$\sigma_{max}(A) \triangleq \sup_{z \in S^{n-1}} \|Az\|_2 \le \sup_{v \in N} \|Av\|_2 + \varepsilon \sigma_{max}(A)$$

$$\sigma_{max}(A) \le \frac{1}{1-\varepsilon} \sup_{v\in N} \|Av\|_2$$

Next $\|Av\|_2 = \sup_{w\in S^{n-1}} |w^T A v|$ $\quad w = u+q,\ u\in M,\ \|q\|_2 \le \delta$

$$\le \sup_{u\in M} |u^T A v| + \delta \|Av\|_2$$

Consequently, since for each u, v:

$$P_r(|u^T A v| \ge t) \le \frac{1}{\sqrt{2\pi}\, t} \exp\left(-\frac{t^2}{2}\right)$$

So, use union bound

$$P_r\left[\begin{array}{l} |u^T A v| \ge t \\ \text{for some } u\in M, \\ v\in N \end{array}\right] \le \underbrace{|M|\cdot|N|}_{O(m^2\cdot c^{2m})} \cdot \frac{1}{\sqrt{2\pi}\, t} \exp\left(-\frac{t^2}{2}\right)$$

由前面那个 $|N| \le 2n\left(1+\frac{2}{\varepsilon}\right)^{n-1}$

Choose $t = O(\sqrt{m})$, then exp small prob.

以上就得到了奇异值的 UB

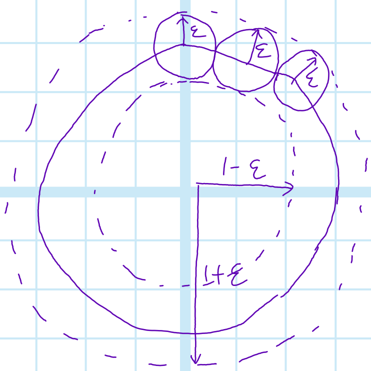
下面补充Prop的证明.

<u>Pf of Prop</u>

a useful trick

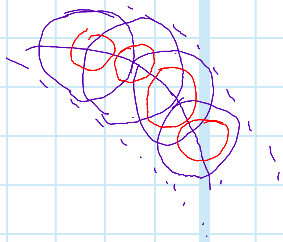
By volume argument

数有多少球，直接用那个环的体积除以每个小球的体积.



Let N be a maximal cardinality subset of S^{n-1}, s.t.

$u, v \in N \Rightarrow \|u-v\|_2 > \varepsilon$



$\Rightarrow$ by construction, N is ε-net

Obs: $u, v \in N$:

$$B(u,\tfrac{\varepsilon}{2}) \cap B(v,\tfrac{\varepsilon}{2}) = \phi$$

$$\bigcup_{u\in N} B(u,\tfrac{\varepsilon}{2}) \subseteq B(0,1+\tfrac{\varepsilon}{2}) \setminus B(0,1-\tfrac{\varepsilon}{2})$$

$$|N| \cdot Vol(B(u,\tfrac{\varepsilon}{2})) \le Vol(B(0,1+\tfrac{\varepsilon}{2})) - Vol(B(0,1-\tfrac{\varepsilon}{2}))$$

Fact:

$\because Vol(B(0,r)) = r^n \cdot Vol(B(0,1))$

$\Rightarrow$ divide by $Vol(B(0,1))$

$$|N| \cdot \left(\tfrac{\varepsilon}{2}\right)^n \le \left(1+\tfrac{\varepsilon}{2}\right)^n - \left(1-\tfrac{\varepsilon}{2}\right)^n$$

Ineq: $(1+x)^l - (1-x)^l \le 2lx(1+x)^{l-1}$

Q.E.D

注意到H是复矩阵,

$$u^H Q u = \left(u^{R^T} - u^{I^T}\right) \begin{bmatrix} Re(Q) & Im(Q) \\ Im(Q) & Re(Q) \end{bmatrix} \begin{bmatrix} u^R \\ u^I \end{bmatrix}$$

看另一种写法，利用Lipschitz性质.

$\sigma_{max}(A)$: 1-Lipschitz

$$\left(|\sigma_{max}(A) - \sigma_{max}(B)| \leq \|A-B\|_2 \right)$$

$$P_r\left[\sigma_{max}(A) \geq \mathbb{E}[\sigma_{max}(A)] + t\right] \leq O(e^{-t^2})$$

(Concentration ineq for Lipschitz fns of Gaussian RVs)

$$\mathbb{E}[\sigma_{max}(A)] = \mathbb{E}\left[\sup_{\substack{\|u\|_2=1 \\ \|v\|=1}} u^T A v\right]$$

一个由 (u,v) index的高斯过程 可以由另一个高斯过程来bound.

Slepian lemma

新问题:　　　　　　　　　　　　　　　　　　2019.4.15

Community Detection

- Goal: identify groups of densely connected nodes in a network.

- Stochastic block model (SBM)

Data:

undirected network of n nodes,

- specified by $n \times n$ adjacent matrix A

$$A_{ij} = \begin{cases} 0 & \text{if } (i,j) \text{ is not an edge} \\ 1 & \text{o/w} \end{cases}$$

- K communities in network, each node belong to one community

→ membership vector z_i for node i

$z_{ik} = 1$ if node i belongs to the community k

→ $z_i \in \{0,1\}^K$, not observed

- SBM is parametrized by Ψ ($K \times K$ symmetric)

Ψ_{kr} = prob of an edge forming between a pair of nodes from community k and r

For simplicity, assume

$$n = mK, \quad m \geq 1 \text{ Integer.}$$

这里researching说每个 community 的 member 相同.

- network generation

Given $\{z_i\}$, Ψ. A_{ij} ($i<j$ (∵对称)) independent Bernoulli RVs,

s.t. $\mathbb{E}[A_{ij} | z_i, z_j] = z_i^T \Psi z_j$ 属于某类的概率乘以类间的概率.

$\left(= \Psi_{kr} \text{ if } \begin{array}{l} z_{ik}=1 \\ z_{jr}=1 \end{array}\right)$ ← 举例

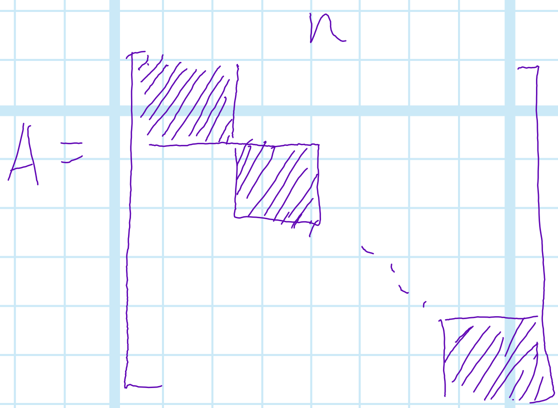
More compactly

$$M_Z = \mathbb{E}[A|Z] = Z \Psi Z^T$$

Z: i^{th} row is z_i^T

Intuitively, if $\Psi_{kr} \simeq 0$ whenever $k \neq r$

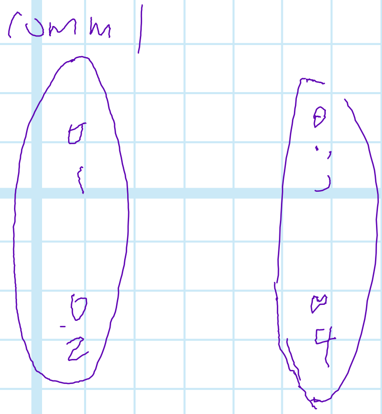
(Between communities have no talking)



Diagnal block will be very dense, otherwise, sparse.

我们要找的其实是 A 的 permute.

e.g. 4 nodes, 2 communities



$n=4, \quad K=2,$

$Z_{11} = Z_{21} = Z_{32} = Z_{42} = 1$

$$\mathbb{E}[A|Z] = Z \Psi Z^T$$

$$= \left[\begin{array}{cc|cc} \Psi_{11} & \Psi_{11} & \Psi_{12} & \Psi_{12} \\ \Psi_{11} & \Psi_{11} & \Psi_{12} & \Psi_{12} \\ \hline \Psi_{12} & \Psi_{12} & \Psi_{22} & \Psi_{22} \\ \Psi_{12} & \Psi_{12} & \Psi_{22} & \Psi_{22} \end{array}\right] \qquad \text{rank} = K = 2$$

generally, rank-K matrix.

Models of Ψ （实际上Ψ可以是任意对称阵）

① planted partition $PP(p,q)$

$$\Psi = \begin{bmatrix} p & & q \\ & p & \\ q & & \ddots & p \end{bmatrix} \qquad p > q$$

community中连接的概率比community间的大

$$= qE_q + (p-q)I_K$$

② Balance PP, $PP_b(p,q)$

every community has the same size $m = \frac{n}{k}$

问题描述：

Given A generated from SBM, estimate Z

log-likelihood function: $\longrightarrow \max$

$$\ell(Z,\Psi) = \sum_{i<j} A_{ij} \log (M_Z)_{ij} + (1-A_{ij}) \log(1-(M_Z)_{ij})$$

$$= \sum_{i<j} A_{ij} (f \circ M_Z)_{ij} + (g \circ M_Z)_{ij}$$

apply f component-wise

$$f(x) = \log \frac{x}{1-x} \qquad g(x) = \log(1-x)$$

Obs: $f \circ M_Z = f \circ (Z \Psi Z^T) = Z(f \circ \Psi) Z^T$

Under the pp model, $PP(p,q)$

$$f \circ \Psi = f(q) E_K + (f(p) - f(q)) I_K$$

$$\Rightarrow \ell(Z,\Psi) = \sum_{i<j} A_{ij}\left[f(q)\, Z E_k Z^T + (f(p)-f(q))\, Z Z^T\right]_{ij}$$

$$+ \sum_{i<j}\left[g(q)\, Z E_k Z^T + (g(p)-g(q))\, Z Z^T\right]_{ij}$$

$$\underset{k\times k}{Z E_k} \underset{k\times n}{Z^T} = \underset{n\times k}{Z}\, E_{k\times n} = E_n$$ 看起来是依赖于Z，但实际并不是

$$= \sum_{i<j}\left[(f(p)-f(q)) A_{ij} (ZZ^T)_{ij} + g(p)-g(q)\,(ZZ^T)_{ij}\right]$$

+ constant.

obs: $[ZZ^T]_{ii} = 1$

$$2\ell(Z,\Psi) = (f(p)-f(q))\langle A, ZZ^T\rangle +$$
$$(g(p)-g(q))\langle E_n, ZZ^T\rangle +$$
const.

Divide $f(p)-f(q)$: we obtain:

优化问题:

$$\hat{Z} \in \underset{Z\in\mathcal{Z}}{\arg\max}\ \langle A, ZZ^T\rangle - \lambda\langle E_n, ZZ^T\rangle \qquad (P)$$

$\mathcal{Z}$: set of admissible membership matrices.

constrain is discrete

non-convex

也可以用 SDP relaxation 来解.

Consider $X = ZZ^T \in \{0,1\}^{n\times n}$

Note : $X_{ij} = \begin{cases} 1 & \text{if } i,j \text{ in same community} \\ 0 & \text{o/w} \end{cases}$

$\longrightarrow$

$$(\mathrm{Q}) \quad \hat{X} \in \underset{X \in \mathcal{X}}{\arg\max} \ \langle A, X \rangle - \lambda \langle E_n, X \rangle$$

$\mathcal{X}$ = set of admissiable clustering matrices

假设类中的member个数相同

Under the PP_b model, all Z take the form up to permutation,

$$Z_0 = n \underbrace{\begin{bmatrix} \vdots & & & \\ & \vdots & & \\ & & \vdots & \\ & & & \vdots \end{bmatrix}}_{K}$$

相同类的放在一起

$$Z_0 = I_K \otimes \mathbb{1}_m$$

$$\Rightarrow \mathcal{Z} = \{ P Z_0 Q^T : P, Q : \text{perm matrices} \}$$

Corespondingly

$$\mathcal{X} = \{ P X_0 P^T, P : \text{perm matrices} \}$$

$$X_0 = Z_0 Z_0^T = I_K \otimes E_m$$

e.g.:

$\boxed{\begin{matrix}1 & 0\\ \\ 2 & 0\end{matrix}}$ $\boxed{\begin{matrix} & 0\\ & 3\\ 4 & 0\end{matrix}}$

$$Z_0 = \begin{bmatrix} 1 & \\ 1 & \\ & 1 \\ & 1 \end{bmatrix}$$

$$X_0 = \left[\begin{array}{cc|cc} 1 & 1 & & \\ 1 & 1 & & \\ \hline & & 1 & 1 \\ & & 1 & 1 \end{array}\right]$$

以上就是问题的 formulation

Idea: Semidefinite relaxation of (Q) SDR

obs:

1°: For any $X \in \mathcal{X}$, $X \succeq 0$ ($\because X = ZZ^T$),

$0 \le X_{ij} \le 1$

$X_{ii} = 1$ (i 和 i 是否在一个 community, 显然在)

2°: $\langle \underset{\text{元全为1}}{E_n}, X \rangle = \mathbb{1}_n^T X \mathbb{1}_n$ ($\because E_n = \mathbb{1}_n \mathbb{1}^T$)

$$X\mathbb{1}_n = PX_0P^T\mathbb{1}_n = PX_0\mathbb{1}_n = mP\mathbb{1}_n = m\mathbb{1}_n$$

$\Rightarrow \langle E_n, X\rangle = mn$ is const.!

∴ 只需要优化第一项了.

(SDR) of (Q):

$$\max \langle A, X\rangle$$
$$\text{s.t.} \quad X\mathbb{1}_n = m\mathbb{1}_n$$
$$\operatorname{diag}(X) = \mathbb{1}_n$$
$$X \succeq 0 \quad (\text{PSD})$$
$$X \geq 0$$

$\because X \succeq 0 \quad \therefore \begin{bmatrix} 1 & x_{ij} \\ x_{ij} & 1 \end{bmatrix} \succeq 0$ ∴不用写 $X \leq 1$，自动蕴含。

Question: Under what conditions would (SDR) give X_0 as the optimal soln?

Idea: Construct a primal-dual certificate.

我们希望 X_0 是 opt. soln. 能否建一个 dual?

Derivation of the dual

Obs: $2(X\mathbb{1}_n)_i = \langle X, \Phi_i\rangle$

$$\Phi_i = e_i\mathbb{1}_n^T + \mathbb{1}_n e_i^T$$

$$\left(\langle X, \mathbb{1}_n e_i^T\rangle = e_i^T X \mathbb{1}_n\right)$$

$$X\mathbb{1}_n = m\mathbb{1}_n \iff \mathcal{L}(X) = 2m\mathbb{1}_n.$$

where $\mathcal{L}(X) = (\langle X, \Phi_1\rangle, \dots, \langle X, \Phi_n\rangle)$

下面来写对偶.

原问题(SDP) $\max \langle A, X\rangle$

$$
\begin{aligned}
\text{s.t. } & X\mathbb{1}_n = m\mathbb{1}_n & (\mu) \\
& \mathrm{diag}(X) = \mathbb{1}_n, & (\nu) \\
& X \succeq 0 \quad X \geq 0 & (\Gamma) \\
& (\Lambda)
\end{aligned}
$$

$\rightarrow$ (SDD) $\min 2m\mathbb{1}_n^T\mu + \mathbb{1}_n^T\nu$

$$
\begin{aligned}
\text{s.t. } & \Lambda \triangleq \mathcal{L}^*(\mu) + \mathrm{diag}^*(\nu) - A - \Gamma \succeq 0 \\
& \Gamma \geq 0.
\end{aligned}
$$

$\mathcal{L}^*$: adjoint operator of $\mathcal{L}$, defined by

$$\mathcal{L}(X) = (\langle X, \Phi_1\rangle, \dots, \langle X, \Phi_n\rangle)$$

$\mu^T\mathcal{L}(X) = \langle X, \mathcal{L}^*(\mu)\rangle$ 类似于转置

$$\mu^T\mathcal{L}(X) = \sum_i \mu_i \langle X, \Phi_i\rangle = \langle X, \underbrace{\sum_i \mu_i \Phi_i}_{\mathcal{L}^*(\mu)}\rangle$$

2019.4.16

SBM:

$$\mathbb{E}[A|Z] = Z\Psi Z^T$$

A : $n \times n$, Z : $n \times K$

membership matrix

Ψ : $K \times K$ $m = n/k$

balanced $PP_b(p, q)$

$$\Psi = \begin{bmatrix} p & & q \\ & p & \\ q & & \ddots \ p \end{bmatrix}$$

$$= q E_k + (p-q) I_k$$

(SDR) $\max \langle A, X \rangle$

s.t. $X \mathbb{1}_n = m \mathbb{1}_n$

$\operatorname{diag}(X) = \mathbb{1}_n$

$X \succeq 0$ $X \geq 0$

$$\mathcal{L}(X) = (\langle X, \Phi_1 \rangle, \cdots, \langle X, \Phi_n \rangle)$$

$$\Phi_i = e_i \mathbb{1}^T + \mathbb{1}_n e_i^T$$

dual:
(SDD)

$$\min\ 2m\mathbb{1}_n^T\mu + \mathbb{1}_n^T\nu$$

$$\text{s.t. } \Lambda \triangleq \mathcal{L}^*(\mu) + \operatorname{diag}^*(\nu) - A - \Gamma \succeq 0$$

$$\Gamma \geq 0$$

WLOG: $X_0 = Z_0 Z_0^T = I_k \otimes E_m$ is the ground truth

$$Z_0 = \begin{bmatrix} \vdots & & & \\ & \vdots & & \\ & & \ddots & \\ & & & \vdots \end{bmatrix} \qquad X_0 = \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & \cdots & S_k \end{array} \\ \begin{bmatrix} E_m & & & \\ & E_m & & \\ & & \ddots & \\ & & & E_m \end{bmatrix} \end{array} \begin{array}{c} S_1 \\ S_2 \\ \vdots n \\ S_k \end{array}$$

← 定义 S_i

S_i: set of indices of nodes in community i

希望 X_0 是 unique solution

Idea: Find a dual solution (μ, ν, Γ) s.t.

$(X_0, (\mu, \nu, \Gamma))$ form a primal-dual opt. pair.

(PF:) $\mathcal{L}(X) = 2m\mathbb{1}$, $\operatorname{diag}(X) = \mathbb{1}_n$, $X \succeq 0$, $X \geq 0$

(DF) $\Lambda \succeq 0$, $\Gamma \geq 0$

(CS) $\langle \Lambda, X \rangle = 0$, $\Gamma_{ij} X_{ij} = 0 \quad \forall i, j$

观察: (SDR) 中, region is compact, opt a contious function on a compact set, it has soln.

下面来解这个 KKT:

From (CS), ① $(X_0)_{S_k,S_k} = E_m \Rightarrow \Gamma_{S_k S_k} = \Lambda$

② $\Lambda X_0 = 0 \iff \langle \Lambda, X_0 \rangle = 0$

(由 spectral decomposition)

$\Rightarrow \mathrm{Ker}(\Lambda) \supseteq \mathrm{range}(X_0) = \{X_0 u\}\ \forall u$

Observe: $\mathbb{I}_{S_k} \in \mathbb{R}^n$ is an eigenvector of X_0 $\forall k$, where

$$\mathbb{I}_{S_k} = \begin{bmatrix} 0 \\ \hline \vdots \\ \hline \mathbb{1}_m \\ \hline \vdots \\ \hline 0 \end{bmatrix} \begin{array}{l} S_1 \\ \\ S_k \\ \\ S_K \end{array}$$

$\Rightarrow \mathrm{range}(X_0) = \mathrm{span}\{\mathbb{I}_{S_1}, \cdots, \mathbb{I}_{S_K}\}$

$\therefore \mathrm{span}\{\mathbb{I}_{S_1}, \cdots \mathbb{I}_{S_K}\} \subseteq \mathrm{Ker}(\Lambda)$

Prop 1: Let μ, ν, Γ be s.t. $\Gamma \geq 0$, Suppose

$\left.\begin{array}{l} \text{(A1)}\ \mathrm{Ker}(\Lambda) = \mathrm{span}\{\mathbb{I}_{S_1}, \cdots, \mathbb{I}_{S_K}\},\ \Lambda \succeq 0 \\ \text{(A2)}\ \Gamma_{S_k,S_k} = 0\ \ \forall k, \end{array}\right\}$ 保证 opt.

(A3) For $k \neq l$, Γ_{S_k,S_l} has at least 1 non-zero element

← 保证 uniqueness

Then, X_0 is the unique opt soln to (SDR)

(μ, v, Γ) is opt for (SDD)

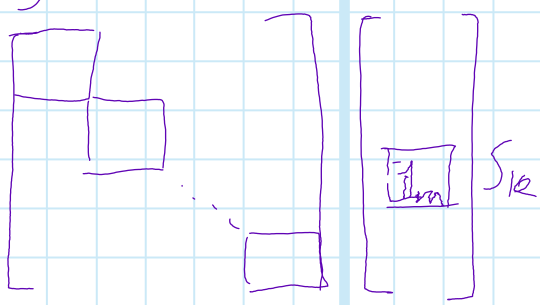
For (A1): It holds if

(A1-1) $\Lambda \mathbb{1}_{S_k} = 0 \quad \forall k$

(A1-2) $u^T \Lambda u \geq \varepsilon \|u\|_2^2 \quad \forall u \in \text{span}\{\mathbb{1}_{S_1}, \cdots \mathbb{1}_{S_K}\}^\perp$

Consider (A1-1):

$0 = \Lambda \mathbb{1}_{S_k} =$



$$\Lambda_{S_k, S_k} \mathbb{1}_m = 0 \qquad \Lambda_{S_k^c, S_k} \mathbb{1}_m = 0$$

Use the def'n of Λ:

$$\mathcal{L}^*(\mu) = \sum_i \mu_i \Phi_i = \mu \mathbb{1}_n^T + \mathbb{1}_n \mu^T$$

$$\text{diag}^*(\nu) = \text{Diag}(\nu)$$

$$\text{diag}(X) = (\langle X, e_1 e_1^T \rangle, \cdots, \langle X, e_n e_n^T \rangle)$$

$$\text{diag}^*(\nu) = \sum_i \nu_i e_i e_i^T = \text{Diag}(\nu)$$

diag: take the diag of matrix as a vector

Diag: take vector and form a matrix with it as diagonal

$$(\mu \mathbb{1}_n^T)_{S_k,S_k} = \mu_{S_k} \mathbb{1}_m^T$$

$$(\mathbb{1}_n \mu^T)_{S_k,S_k} = \mathbb{1}_m \mu_{S_k}^T$$

$$0 = \Lambda_{S_k,S_k} \mathbb{1}_m$$

$$= \Big(\mu_{S_k}\mathbb{1}_m^T + \mathbb{1}_m \mu_{S_k}^T + \mathrm{Diag}(\nu_{S_k}) - A_{S_k,S_k} - \overset{0}{\overset{\|}{\Gamma_{S_k,S_k}}}\Big)\mathbb{1}_m$$

$$= \underbrace{m\mu_{S_k} + (\mu_{S_k}^T \mathbb{1}_m)\mathbb{1}_m}_{\text{如果 } \mu_{S_k} = \frac{1}{2}\phi_k \mathbb{1}_m \text{ 试可以全消了}} + \nu_{S_k} - \underbrace{A_{S_k,S_k}\mathbb{1}_m}_{= [d(S_k)]_{S_k}}$$

可以取 $\nu_{S_k} = [d(S_k)]_{S_k} - m\phi_k \mathbb{1}_m$

$d(S_k) = A\,\mathbb{I}_{S_k}$

下面来看 $\Lambda_{S_\ell,S_k}\mathbb{1}_m = 0$ 那个条件.

$\Lambda_{S_\ell,S_k}\mathbb{1}_m = 0 \quad \forall k \neq \ell \qquad \mathcal{L}^*(\mu) = \mu\mathbb{1}_n^T + \mathbb{1}_n\mu^T$

$\Rightarrow \left[\mu_{S_\ell}\mathbb{1}_m^T + \mathbb{1}_m\mu_{S_k}^T - (A+\Gamma)_{S_\ell,S_k}\right]\mathbb{1}_m = 0$

代入 μ_{S_k}

$\Rightarrow \quad m\left(\frac{1}{2}\phi_\ell + \frac{1}{2}\phi_k\right)\mathbb{1}_m = (A+\Gamma)_{S_\ell,S_k}\mathbb{1}_m.$

$\Rightarrow \mathbb{1}_m$ is an eigenvector of $(A+\Gamma)_{S_\ell,S_k}$

$$(A+\Gamma)_{S_k,S_\ell} = \frac{1}{2}(\phi_k+\phi_\ell)E_m + B_{S_k,S_\ell}$$

where B_{S_k,S_ℓ} acts on $\text{span}\{\mathbb{1}_m\}^\perp$

下面来看 (A1-2)

Note: $\text{span}\{\mathbb{I}_{S_1},\cdots,\mathbb{I}_{S_K}\}^\perp$

$$= \left\{ u = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix} = \sum_k e_k \otimes u_k : \mathbb{1}_m^T u_k = 0 \ \forall k \right\}$$

Take $u=(u_1,\cdots,u_K) \in \text{span}\{\mathbb{I}_{S_1},\cdots,\mathbb{I}_{S_K}\}^\perp$

$$u^T \Lambda u = \sum_{k,\ell} u_k^T \Lambda_{S_k,S_k} u_\ell$$

$$= \sum_k u_k^T \Lambda_{S_k,S_k} u_k + \sum_{k\neq\ell} u_k^T \Lambda_{S_k,S_\ell} u_\ell$$

$$\Lambda_{S_k,S_k} = \mu_{S_k}\mathbb{1}^T + \mathbb{1}_m \mu_{S_k}^T + \text{diag}^*(\nu_{S_\ell}) - A_{S_k,S_k}$$

$$\Rightarrow u_k^T \Lambda_{S_k,S_k} u_k$$

$$= u_k^T\left[\text{diag}^*(\nu_{S_k}) - A_{S_k,S_k}\right] u_k,$$

$$\Delta \triangleq A - \mathbb{E}[A \mid z]$$

$$\Rightarrow \Delta_{S_k,S_k} = A_{S_k,S_k} - pE_m$$

$$= u_k^T\left[\text{diag}^*(\nu_{S_k}) - pE_m - \Delta_{S_k,S_k}\right] u_k$$

can be drop $\because E_m = \mathbb{1}_m\mathbb{1}_m^T$

$\mathbb{1}_m u_k = 0$

$$= u_k^T \left[\mathrm{diag}^*\left([d(S_k)]_{S_k} - \phi_k m I_m - \Delta_{S_k,S_k} \right] u_k$$

$$\Lambda_{S_k,S_l} = \mu_{S_k} \mathbb{1}_m^T + \mathbb{1}_m \mu_{S_l}^T - (A+\Gamma)_{S_k,S_l}$$

$$= \frac{1}{2}(\phi_k + \phi_l) E_m - (A+\Gamma)_{S_k,S_l} = -B_{S_k,S_l}$$

$$\Rightarrow u^T \Lambda u = \sum_k u_k^T \left[\mathrm{diag}^*\left([d(S_k)]_{S_k} \right) - \phi_k m I_m - \Delta_{S_k,S_k} \right] u_k$$

$$- \sum_{k \neq l} u_k^T B_{S_k,S_l} u_l$$

Set $B_{S_k,S_l} = P_{\mathbb{1}_m^\perp} A_{S_k,S_l} P_{\mathbb{1}_m^\perp}$

<u>Obs</u>: Γ_{S_k,S_l} is determined

$$u_k^T B_{S_k,S_l} u_l = u_k^T P_{\mathbb{1}_m^\perp} A_{S_k,S_l} P_{\mathbb{1}_m^\perp} u_l$$

$$\Delta_{S_k,S_l} = A_{S_k,S_l} - q E_m$$

$$= u_k^T P_{\mathbb{1}_m^\perp} (\underbrace{q E_m}_{\text{projection} = 0} + \Delta_{S_k,S_l}) P_{\mathbb{1}_m^\perp} u_l$$

$$= u_k^T \Delta_{S_k,S_l} u_l$$

$$= \sum_k u_k^T \left[\mathrm{diag}^*[d(S_k)]_{S_k} - \Delta_{S_k,S_k} \right] u_k -$$

$$\sum_k m\phi_k \|\mu_k\|_2^2 - \sum_{k\neq l} u_k^T \Delta_{s_k,s_l} u_l$$

If $\operatorname{diag}^*[d(S_k)]_{S_k} \succeq m\rho_k I_m$ for some ρ_k, then

$$u^T \Lambda u \geq \sum_k \left[(\rho_k - \phi_k) m - \|\Delta_{s_k,s_k}\| \right] \|u_k\|_2^2 - \sum_{k\neq l} u_k^T \Delta_{s_k,s_l} u_l$$

$$\geq \underbrace{\min_k \left[(\rho_k - \phi_k)m - \|\Delta_{s_k,s_k}\| \right]}_{\mathrm{I}} \|u\|_2^2 \underbrace{- \|E_{S_0^c} \circ \Delta\| \cdot \|u\|_2^2}_{\mathrm{II}}$$

as long as $\mathrm{I} > \mathrm{II}$, done.

need to use prob model.

use matrice concentration inequality

以上都是 SDR approuch, from practical, you cannot run in large scale, is it possible to develop some light weight, with theory guarteen.

Sum:

How prob & opt. get tied together.

Several algo ideas