High Dimensional Statistics RIT: *Chapter 9 & 10* Quick Sketch

Zhirui Li

University of Maryland

November 20, 2024

Matrix Regression

For space of $d_1 \times d_2$ matrices, the one possible inner product is defined as $\langle\langle A, B \rangle\rangle = \mathrm{Trace}(A^T B) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} A_{jk} B_{jk}$. The norm induced by the inner product is $\|A\|_F = \sqrt{\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} (A_{jk})^2}$.

Consider the *Matrix Regression*: We observe $Z_i = (X_i, y_i), i = [n] = \{1, 2, \cdots, n\}$ where $X_i \in \mathbb{R}^{d_1 \times d_2}$ are covariates and $y_i \in \mathbb{R}$ are response variables.

For simplicity, assume we have linear link: $y_i = \langle\langle X_i, \Theta^* \rangle\rangle + w_i$, where w_i are noise variables.

For simplicity define the observation operator $\mathcal{X}_n : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}^n$ given by $[\mathcal{X}_n(\Theta)]_i = \langle\langle X_i, \Theta \rangle\rangle$, then the full regression can be written as

$$\mathbf{y} = \mathcal{X}_n(\Theta^*) + \mathbf{W}$$

Matrix Regression w/ Rank Constraints

Matrix regression: $\mathbf{y} = \mathcal{X}_n(\Theta^*) + \mathbf{W}$.

In application, Θ^* could be low-rank or approximated by a low rank matrix. We could apply rank penalty, which will make the regression problem non-convex.

Instead, we use a nuclear norm penalty and have

$$\hat{\boldsymbol{\Theta}} \in \arg \min_{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \| y - \mathcal{X}_n(\boldsymbol{\Theta}) \|_2^2 + \lambda_n \|\boldsymbol{\Theta}\|_{\text{nuc}} \right\} \tag{1}$$

where $\|\boldsymbol{\Theta}\|_{\mathrm{nuc}} = \sum_{j=1}^{\min(d_1, d_2)} \sigma_j(\boldsymbol{\Theta})$, i.e., the sum of singular values of $\boldsymbol{\Theta}$.

We take a detour to Chapter 9, to understand the nature of Eq. 1.

General Regularized M-estimator

Given an indexed family of probability distributions $\{P_\theta : \theta \in \Omega\}$ where θ is the parameter to be estimated and Ω is the parameter space.

Consider an observed sample $\mathbf{Z}^n = (Z_1, Z_2, \ldots, Z_n)$, each of $Z_i \in \mathcal{Z}$ where $\mathcal{Z}$ is the sample space. Suppose $Z_i \sim P = P_{\theta^*}$, our goal is to estimate θ^*.

Wainwright defines the *cost function*, which I'll refer to as the *loss function* later, as $\mathcal{L}_n : \Omega \times \mathcal{Z}^{\otimes n} \mapsto \mathbb{R}$.

The risk (called population cost function by wainwright) is defined as $\mathcal{L}(\theta) = \mathbb{E}\left(\mathcal{L}_n(\theta; \mathbf{Z}^n)\right)$.

The target parameter θ^* is then $\theta^* = \arg\min_{\theta \in \Omega} \mathcal{L}(\theta)$.

Remark (language "stole" from Dr. Slud): in many settings, θ^* lies in the interior of Ω, and it is the calculus minimizer in the sense that $\nabla \mathcal{L}(\theta^*) = 0$.

General Regularized M-estimator

To ensure certain imposed structure of θ^* (e.g., sparsity), we introduce appropriate penalty and have

$$\widehat{\theta} \in \arg\min_{\theta \in \Omega} \{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \Phi(\theta)\} \tag{2}$$

where λ_n is a user defined weight parameter, Φ is a proper chosen function of θ, for example, the L_p norm.

Back to Matrix Regression

Eq. 2 $\widehat{\theta} \in \arg\min_{\theta \in \Omega} \{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \Phi(\theta)\}$.

Eq. 1 $\hat{\mathbf{\Theta}} \in \arg \min_{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathcal{X}_n(\mathbf{\Theta})\|_2^2 + \lambda_n \|\mathbf{\Theta}\|_{\text{nuc}} \right\}$

The nuclear norm provides a natural relaxation of rank of the matrix in the following sense: given $\Theta \in \mathbb{R}^{d_1 \times d_2}$, perform the SVD: $\Theta = UDV^T$, where D is a diagonal matrix with entries

$$\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \cdots \geq \sigma_{\min(d_1,d_2)}(\Theta) \geq 0$$

Note that the rank of Θ is the number of non-zero singular values: $\text{rank}(\Theta) = |\{j : \sigma_j(\Theta) > 0\}|$.

The convex relaxation (particularly popular in SDP) of the rank constraints tells us a proper Φ in Eq. 2 would be the nuclear norm – the ℓ_1 norm of the vector of singular values of Θ.

Analysis of the nuclear norm regularization

Earlier this semester, Chugang talked about the Lasso regression and a general framework relates to "decomposable" regularizers.

We now quickly state relevant definitions and results from Ch. 9 of Wainwright book, with the proof for **none** of them.

To start, we mention that our goal is to **bound** $\hat{\theta} - \theta^*$**.**

Decomposable Regularizers

Assume the parameter space $\Omega \subseteq \mathbb{R}^d$ is equipped with an inner product $\langle \cdot, \cdot \rangle$, and $\| \cdot \|$ is a norm induced by this inner product. (Note again that for space of $d_1 \times d_2$ matrices, the inner product is the trace and the norm is the matrix Frobenius norm.)

Take a pair of subspace $\mathbb{M} \subseteq \bar{\mathbb{M}} \subseteq \mathbb{R}^d$, recall the *orthogonal complement* of $\bar{\mathbb{M}}$ is $\bar{\mathbb{M}}^\perp := \{v \in \mathbb{R}^d : \langle u, v \rangle = 0, \forall u \in \bar{\mathbb{M}}\}$.

Decomposable Regularizers

Given a pair of subspaces $\mathbb{M} \subseteq \bar{\mathbb{M}}$, a norm-based regularizer Φ is decomposable with respect to $(\mathbb{M}, \bar{\mathbb{M}}^\perp)$ if

$$\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta) \quad \text{for all } \alpha \in \mathbb{M} \text{ and } \beta \in \overline{\mathbb{M}}^\perp.$$

Decomposable Regularizers: Why?

Prop 9.13, Wainwright 2019

Let $\mathcal{L}_n : \Omega \mapsto \mathbb{R}$ be a convex function, let the regularizer $\Phi : \Omega \to [0, \infty)$ be a norm, and let $\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right)$ be a pair of subspace of $\mathbb{R}^d$ such that Φ is decomposable on this pair. Given $\Phi^* \left(\nabla \mathcal{L}_n(\theta^*)\right) \leq \frac{\lambda_n}{2}$, where Φ^* is the dual norm of Φ, we have

$$\hat{\Delta} = \hat{\theta} - \theta^* \in \left\{\Delta \in \Omega : \Phi\left(\Delta_{\overline{\mathbb{M}}^{\perp}}\right) \leq 3\Phi\left(\Delta_{\overline{\mathbb{M}}}\right) + 4\Phi\left(\theta^*_{\mathbb{M}^{\perp}}\right)\right\}$$

Note: $\nabla \mathcal{L}_n(\theta^*)$ is frequently referred to as the score function.
The dual norm of Φ, Φ^*, is defined such that $\Phi^*(\mathbf{v}) = \sup_{\Phi(\mathbf{u}) \leq 1} \langle u, v \rangle$.

Curvature

In classical mathematical statistics, the curvature of the loss function is captured by the Fisher's information, and is used to quantify the variance of MLE via Rao-Cramer Lower Bound.

In high dimensional settings, strict convexity in all directions are often prohibited.

We follow Section 9.3 and discuss two restricted curvature conditions, and corresponding results.

Restricted Strong Convexity

Given any differentiable loss function, we look at the 1st order Taylor Expansion error

$$\mathcal{E}_n(\Delta) := \mathcal{L}_n\left(\theta^* + \Delta\right) - \mathcal{L}_n\left(\theta^*\right) - \left\langle \nabla \mathcal{L}_n\left(\theta^*\right), \Delta \right\rangle$$

Restricted Strong Convexity

For a given norm $\|\cdot\|$ and regularizer $\Phi(\cdot)$, the loss function satisfies a restricted strong convexity condition with radius $R > 0$, curvature $\kappa > 0$ and tolerance τ_n^2 if

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta) \quad \text{for all } \Delta \in \mathbf{B}_R(\mathbf{0})$$

Combined with the decomposability of the regularizers, the following theorem achieves our goal (bounding $\hat{\theta} - \theta^*$)

Restricted Strong Convexity

Theorem 9.19, Wainwright 2019

Assume that the loss function is convex, satisfies the restricted strong convexity condition with parameters above, and Φ is decomposable over $(\mathbb{M}, \bar{\mathbb{M}}^{\perp})$, then

(a) Any optimal solution satisfies the bound

$$\Phi\left(\widehat{\theta}-\theta^*\right) \leq 4\left[\Psi(\overline{\mathbb{M}})\left\|\widehat{\theta}-\theta^*\right\| + \Phi\left(\theta^*_{\mathbb{M}^{\perp}}\right)\right]$$

(b) If $\left(\overline{\mathbb{M}}, \mathbb{M}^{\perp}\right)$ satisfies $\tau_n^2\Psi^2(\overline{\mathbb{M}}) \leq \frac{\kappa}{64}$ and $\varepsilon_n\left(\overline{\mathbb{M}}, \mathbb{M}^{\perp}\right) \leq R$, we have

$$\left\|\widehat{\theta}-\theta^*\right\|^2 \leq \varepsilon_n^2\left(\overline{\mathbb{M}}, \mathbb{M}^{\perp}\right)$$

where $\Psi(\mathbb{S}) = \sup_{\mathbf{u}\neq\mathbf{0}, \mathbf{u}\in\mathbb{S}} \frac{\Phi(\mathbf{u})}{\|\mathbf{u}\|}$ and

$$\varepsilon_n^2\left(\overline{\mathbb{M}}, \mathbb{M}^{\perp}\right) := 9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\overline{\mathbb{M}}) + \frac{8}{\kappa}\left\{\lambda_n\Phi\left(\theta^*_{\mathbb{M}^{\perp}}\right) + 16\tau_n^2\Phi^2\left(\theta^*_{\mathbb{M}^{\perp}}\right)\right\}$$

Φ^*-Norm Curvature Condition

An alternative way to look at the curvature of the loss function:

Φ^*-Norm Curvature Condition

The loss function satisfies Φ^* curvature condition with curvature $\kappa > 0$, tolerance $\tau_n \geq 0$ and radius R if

$$\Phi^*(\nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*)) \geq \kappa \Phi^*(\Delta) - \tau_n \Phi(\Delta)$$

for all $\Delta \in \{\theta \in \Omega : \Phi^*(\theta) \leq R\}$.

With this we have

Theorem 9.24, Wainwright 2019

Suppose Φ is decomposable over $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$, $\tau_n \Psi^2(\mathbb{M}) < \frac{\kappa}{32}$ and the event $\{\Phi^*(\nabla \mathcal{L}_n(\theta^*)) \leq \frac{\lambda_n}{2}\} \cap \{\Phi^*(\hat{\theta} - \theta^*) \leq R\}$:

$$\Phi^*(\hat{\theta} - \theta^*) \leq \frac{3\lambda_n}{\kappa}$$

Finding subspaces that $\|\cdot\|_{\text{nuc}}$ is decomposable

To apply the results above, the we need to find subspaces $\mathbb{M} \subset \overline{\mathbb{M}}$ of $\mathbb{R}^{d_1 \times d_2}$ such that $\|\cdot\|_{\text{nuc}}$ is decomposable over this pair.

Given $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, let $\text{rowspan}(\boldsymbol{\Theta}) \subset \mathbb{R}^{d_2}$ and $\text{colspan}(\boldsymbol{\Theta}) \subset \mathbb{R}^{d_1}$ be the row space and column space of $\boldsymbol{\Theta}$, respectively. For low-rank purpose, let $r \leq \min(d_1, d_2)$ be a positive integer, which will be the rank of our estimator $\hat{\boldsymbol{\Theta}}$.

Let $\mathbb{U}, \mathbb{V}$ be r-dimensional subspace of vectors of appropriate dimensions. Define

$$\mathbb{M}(\mathbb{U}, \mathbb{V}) := \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} \middle| \, \text{rowspan}(\boldsymbol{\Theta}) \subseteq \mathbb{V}, \text{colspan}(\boldsymbol{\Theta}) \subseteq \mathbb{U} \right\}$$

$$\overline{\mathbb{M}}^{\perp}(\mathbb{U}, \mathbb{V}) := \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} \middle| \, \text{rowspan}(\boldsymbol{\Theta}) \subseteq \mathbb{V}^{\perp}, \text{colspan}(\boldsymbol{\Theta}) \subseteq \mathbb{U}^{\perp} \right\}$$

We will omit $(\mathbb{U}, \mathbb{V})$ when the context is clear. Also note that $\overline{\mathbb{M}} = (\overline{\mathbb{M}}^{\perp})^{\perp}$.

Finding subspaces that $\|\cdot\|_{\text{nuc}}$ is decomposable

Note that for the choice of the pair of subspaces, we have $\mathbb{M} \subsetneq \overline{\mathbb{M}}$.

To see this, let $d' = \min(d_1, d_2)$, let $\mathbf{U} \in \mathbb{R}^{d_1 \times d'}, \mathbf{V} \in \mathbb{R}^{d_2 \times d'}$ be matrices with orthonormal columns. If we set $\mathbb{U}$ be the span of first r columns of $\mathbf{U}$, $\mathbb{V}$ be the span of first r columns of $\mathbf{V}$.

For matrices $A \in \mathbb{M}, B \in \overline{\mathbb{M}}^{\perp}$, some easy linear algebra shows

$$A = \mathbf{U} \begin{bmatrix} *_A & \mathbf{0}_{r\times(d'-r)} \\ \mathbf{0}_{(d'-r)\times r} & \mathbf{0}_{(d'-r)\times(d'-r)} \end{bmatrix} \mathbf{V}^T; \quad B = \mathbf{U} \begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(d'-r)} \\ \mathbf{0}_{(d'-r)\times r} & *_B \end{bmatrix} \mathbf{V}^T$$

Also, for any $\overline{A} \in \overline{\mathbb{M}}$, $\overline{A} = \mathbf{U} \begin{bmatrix} *_{\overline{A}1} & *_{\overline{A}2} \\ *_{\overline{A}3} & \mathbf{0}_{(d'-r)\times(d'-r)} \end{bmatrix} \mathbf{V}^T$.

Note that $*.$ means a block matrix with arbitrary entries, so indeed we have $\mathbb{M} \subsetneq \overline{\mathbb{M}}$.

Finding subspaces that $\|\cdot\|_{\text{nuc}}$ is decomposable

Finally, note that given $A \in \mathbb{M}, B \in \overline{\mathbb{M}}^{\perp}$, we have (where V^{-T} is short hand for $(V^T)^{-1}$)

$$\begin{aligned}
\|A+B\|_{\text{nuc}} &= \left\|\mathbf{U}^{-1}A\mathbf{V}^{-T} + \mathbf{U}^{-1}B\mathbf{V}^{-T}\right\|_{\text{nuc}} \\
&= \left\| \begin{bmatrix} *_A & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & *_B \end{bmatrix} \right\|_{\text{nuc}} \\
&= \left\| \begin{bmatrix} *_A & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right\|_{\text{nuc}} + \left\| \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & *_B \end{bmatrix} \right\|_{\text{nuc}} \\
&= \left\| \mathbf{U} \begin{bmatrix} *_A & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T \right\|_{\text{nuc}} + \left\| \mathbf{U} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & *_B \end{bmatrix} \mathbf{V}^T \right\|_{\text{nuc}} \\
&= \|A\|_{\text{nuc}} + \|B\|_{\text{nuc}}
\end{aligned}$$

Restricted Strong Convexity and Error Bounds

Our gerneral objective function, Eq 2, is pasted again

$$\hat{\boldsymbol{\Theta}} \in \arg \min_{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \{\mathcal{L}_n(\boldsymbol{\Theta}) + \lambda_n \|\boldsymbol{\Theta}\|_{\text{nuc}}\}$$

To apply Theorem 9.13, the first assumption we need is $\Phi^*(\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)) \leq \frac{\lambda_n}{2}$. Here note that $\|\cdot\|_{\text{nuc}} = \|\cdot\|_2$.

By Prop. 9.13, we have that:
For $\lambda_n \geq 2\|\nabla \mathcal{L}_n(\boldsymbol{\Theta}^*)\|_2$, let $\hat{\Delta} = \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$, and $\hat{\Delta}_{\bar{\mathbb{M}}}$ denote the projection of $\hat{\Delta}$ onto $\overline{\mathbb{M}}$, then

$$\left\|\widehat{\Delta}_{\bar{\mathbb{M}}^\perp}\right\|_{\text{nuc}} \leq 3 \left\|\widehat{\Delta}_{\overline{\mathbb{M}}}\right\|_{\text{nuc}} + 4 \left\|{\boldsymbol{\Theta}^*}_{\mathbb{M}^\perp}\right\|_{\text{nuc}}$$

Restricted Strong Convexity and Error Bounds

When the loss is the standard L_2 loss, the objective becomes Eq 1:

$$\hat{\mathbf{\Theta}} \in \arg\min_{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathcal{X}_n(\mathbf{\Theta})\|_2^2 + \lambda_n \|\mathbf{\Theta}\|_{\text{nuc}} \right\}$$

The the restricted strong convexity condition amounts to lower bounding the term $\frac{\|\mathcal{X}_n(\Delta)\|_2^2}{2n}$.

With this, we assume

$$\frac{\|\mathcal{X}_n(\Delta)\|_2^2}{2n} \geq \frac{\kappa}{2} \|\Delta\|_F^2 - c_0 \frac{(d_1 + d_2)}{n} \|\Delta\|_{\text{nuc}}^2, \quad \text{for all } \Delta \in \mathbb{R}^{d_1 \times d_2}$$

Restricted Strong Convexity and Error Bounds

We are ready to state Theorem 9.19 in the context of matrix regression:

Prop. 10.6, Wainwright 2019

Suppose that $\mathcal{X}_n$ satisfies the restricted strong convexity condition with parameter $\kappa > 0$. Then conditioned on the event $\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{X}_i\right\|_2 \leq \frac{\lambda_n}{2}\right\}$, any optimal solution to nuclear norm regularized least squares satisfies the bound

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_F^2 \leq \frac{9}{2}\frac{\lambda_n^2}{\kappa^2} r + \frac{1}{\kappa}\left\{2\lambda_n \sum_{j=r+1}^{d'} \sigma_j(\boldsymbol{\Theta}^*) + \frac{32 c_0 (d_1 + d_2)}{n}\left[\sum_{j=r+1}^{d'} \sigma_j(\boldsymbol{\Theta}^*)\right]^2\right\}$$

for any $r \leq \frac{\kappa n}{128 c_0 (d_1 + d_2)}$.

Φ^*-Norm Curvature Condition

For the Φ^*-Norm Curvature Condition, the assumption in the context of matrix regression with Φ being the nuclear norm becomes

$$\left\| \frac{1}{n} \mathcal{X}_n^* \mathcal{X}_n(\Delta) \right\|_2 \geq \kappa \|\Delta\|_2 - \tau_n \|\Delta\|_{\text{nuc}} \qquad \text{for all } \Delta \in \mathbb{R}^{d_1 \times d_2}$$

And Theorem 9.24 becomes

Prop. 10.7, Wainwright 2019

Assume the Φ^*-Norm Curvature Condition above, consider a matrix $\mathbf{\Theta}^*$ with $\text{rank}(\mathbf{\Theta}^*) < \frac{\kappa}{64\tau_n}$. Then, conditioned on the event $\left\{ \left\| \frac{1}{n} \mathcal{X}_n^* \right\|_2 \leq \frac{\lambda_n}{2} \right\}$, any optimal LS solution of Eq. 1 satisfies the bound

$$\left\| \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right\|_2 \leq 3\sqrt{2} \frac{\lambda_n}{\kappa}$$

Applications: Multivariate Regression

We again take a detour to Ch. 9 and discuss the general setup of multivariate regression.

Suppose we observe $(\mathbf{z}_i, \mathbf{y}_i) \in \mathbb{R}^p \times \mathbb{R}^T$, $i \in [n]$. Then write $Y \in \mathbb{R}^{n\times T}, Z \in \mathbb{R}^{n\times p}$ such that $\mathbf{y}_i, \mathbf{z}_i$ are respectively their i-th row. For simplicity assume the linear model

$$Y = Z\Theta^* + W$$

where $\Theta^* \in \mathbb{R}^{p\times T}$ is the matrix of regression coefficients, and W is the noise matrix.

Mentioned in the book, a naive approach would be to decompose the problem into T sub-problems

$$Y_{\sim,t} = Z\Theta^*_{\sim,t} + W_{\sim,t}, \quad t \in [T]$$

This approach lacks the consideration that columns may have interactions.

Applications: Multivariate Regression

Instead, consider the M-estimator approach. Assume that S is a subset of $[n]$ such that $\Theta[S,:]$ is significant predictor, i.e., Θ is a row sparse matrix. To ensure this row sparsity, we use

$$\Phi(\Theta) = \underbrace{\sum_{j=1}^{p} \|\Theta_{j,\cdot}\|_2}_{\ell_1 \text{ norm of row-wise } \ell_2 \text{ norm}}$$

Group Lasso: Let $\mathcal{G} = \{G_1, G_2, \cdots, G_T\}$ be a disjoint partition of the index set $[d]$, i.e., $\{\cup G_i = [d]\} \wedge \{G_i \cap G_j = \emptyset, i \neq j\}$.
Given $\theta \in \mathbb{R}^d$, let
$\theta_G = \{\theta | i\text{-th component of } \theta \text{ is } 0 \text{ if } i \notin G; \text{ is } \theta_i \text{ if } i \in G\}$.
For any given norm, the group lasso norm is defined as

$$\Phi(\theta) := \sum_{G \in \mathcal{G}} \|\theta_G\|, \quad \text{i.e., the } \ell_1 \text{ norm of the } \ell_2 \text{ norm with-in each group in } \mathcal{G}$$

Applications: Multivariate Regression

The multivariate regression problem with low-rank constraint on Θ can also be solved via the matrix regression.

To this end, write $\mathbf{y}_i = \langle\langle X_i, \Theta^* \rangle\rangle + W_i$, $i = 1, 2, \ldots, nT$.

Let $E_{jl} \in \mathbb{R}^{n\times T} = [\mathbb{1}_{jl}]$, $X_{jl} = Z^T E_{jl}$ and $y_{jl} = [Y]_{jl}$.

The matrix regression problem is thus

$$y_{jl} = \langle\langle X_{jl}, \Theta^* \rangle\rangle + W_{jl}$$

It's easy to see in this case, the observation operator $\mathcal{X}_n(\Theta^*)$ is simply $Z\Theta^*$.

So the model $Y = Z\Theta^* + W$ is our model for the multivariate regression problem.

Also, the Lease-Square loss has the form $\mathcal{L}_n(\Theta) = \dfrac{1}{2n}\|Y - Z\Theta\|_F$

Applications: Multivariate Regression

Cor. 10.14, Wainwright 2019

Suppose $\Theta^* \in \mathbb{R}^{p\times T}$ has rank at most r, and the noise matrix W has i.i.d. entries that are zero-mean and σ-sub-Gaussian. Let $\widehat{\Sigma} = \frac{Z^T Z}{n}$ be the sample covariance matrix. Then any solution to least square objective with $\lambda_n = 10\sigma\sqrt{\gamma_{\max}(\widehat{\Sigma})}\left(\sqrt{\frac{p+T}{n}} + \delta\right)$ satisfies the bound

$$\left\|\widehat{\Theta} - \Theta^*\right\|_2 \leq 30\sqrt{2}\frac{\sigma\sqrt{\gamma_{\max}(\widehat{\Sigma})}}{\gamma_{\min}(\widehat{\Sigma})}\left(\sqrt{\frac{p+T}{n}} + \delta\right)$$

with probability at least $1 - 2e^{-2n\delta^2}$. Moreover, we have

$$\left\|\widehat{\Theta} - \Theta^*\right\|_F \leq 4\sqrt{2r}\left\|\widehat{\Theta} - \Theta^*\right\|_2 \quad \text{and} \quad \left\|\widehat{\Theta} - \Theta^*\right\|_{\text{nuc}} \leq 32r\|\|\widehat{\Theta} - \Theta^*\|_2.$$

$\gamma(\widehat{\Sigma})$ is the set of eigenvalues of $\widehat{\Sigma}$.

Applications: Multivariate Regression

Proof:

$$\nabla \mathcal{L}_n(\Theta^* + \Delta) - \nabla \mathcal{L}_n(\Theta^*) = \frac{1}{n} Z^T(\mathbf{y} - Z(\Theta^* + \Delta)) - \frac{1}{n} Z^T(\mathbf{y} - Z\Theta^*) = \frac{Z^T Z}{n}\Delta = \widehat{\Sigma}\Delta$$

For any $\mathbf{u} \in \mathbb{R}^T$, we have $\|\widehat{\Sigma}\Delta\mathbf{u}\|_2 \geq \gamma_{\min}(\widehat{\Sigma})\|\Delta\mathbf{u}\|_2$, thus

$$\|\widehat{\Sigma}\Delta\|_2 = \sup_{\|\mathbf{u}\|_2=1} \|\widehat{\Sigma}\Delta\mathbf{u}\|_2 \geq \gamma_{\min}(\widehat{\Sigma}) \sup_{\|\mathbf{u}\|_2=1} \|\Delta\mathbf{u}\|_2 = \gamma_{\min}(\widehat{\Sigma})\|\Delta\|_2$$

So the Φ^* norm curvature condition $[\|\nabla \mathcal{L}_n(\Theta^* + \Delta) - \nabla \mathcal{L}_n(\Theta^*)\|_2 \geq \kappa\|\Delta\|_2 + \tau_n\|\Delta\|_{\text{nuc}}]$ is satisfied with $\kappa = \gamma_{\min}(\widehat{\Sigma})$ and $\tau_n = 0$.

Now in the theorem, λ_n has been specified, and we can show (detail omitted)

$$P\left[\left\|\frac{1}{n} Z^T W\right\|_2 \geq 5\sigma\sqrt{\gamma_{\max}(\widehat{\Sigma})}\left(\sqrt{\frac{p+T}{n}} + \delta\right)\right] \leq 2e^{-2n\delta^2}$$

Applications: Multivariate Regression

Proof continued: So with probability at least $1 - 2e^{-2n\delta^2}$, we have
$\|\nabla \mathcal{L}_n(\Theta^*)\|_2 = \left\| \frac{1}{n} Z^T W \right\|_2 \leq \frac{\lambda_n}{2}$.

All conditions of Prop. 10.7 have been met, and the 2-norm bound follows.

Now since $\operatorname{rank}(\overline{\mathbb{M}}) \leq 2r$, we know (where $\sigma_i(\cdot)$ denotes the i-the singular value)

$$\|\hat{\Delta}\|_{\text{nuc}} = \sum_{i=1}^{n} \sigma_i(\hat{\Delta}) \leq 4\sqrt{2r} \left(\sum_{i=1}^{n} [\sigma_i(\hat{\Delta})]^2 \right)^{1/2} = 4\sqrt{2r}\|\hat{\Delta}\|_F$$

Therefore,

$$\|\hat{\Delta}\|_F^2 = \langle \hat{\Delta}, \hat{\Delta} \rangle \leq \|\hat{\Delta}\|_{\text{nuc}} \|\hat{\Delta}\|_2 \leq 4\sqrt{2r}\|\hat{\Delta}\|_F \|\hat{\Delta}\|_2$$

Applications: Low Rank Matrix Completion

Another application is Low Rank Matrix Completion. The goal is to estimate $\mathbf{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ based on noisy observations of some of its entries. We need to assume $\mathbf{\Theta}^*$ is low rank, or can be well-approximated by a low rank matrix.

To formulate the goal into a matrix regression problem, assume we observe $\tilde{y}_i = \mathbf{\Theta}^*_{a(i),b(i)} + \frac{w_i}{\sqrt{d_1 d_2}}, i \in [n]$, where $a(i), b(i)$ is the indices in $\mathbf{\Theta}^*$ of the i-th observation and w_i is the noise. The normalizing constant $\sqrt{d_1 d_2}$ ensures $E\|\mathcal{X}_n(\mathbf{\Theta}^*)\|_2^2 = n\|\mathbf{\Theta}^*\|_F^2$.

Let $X_i \in \mathbb{R}^{d_1 \times d_2}$ be the matrix with 0 everywhere except for $X_{a(i),b(i)} = \sqrt{d_1 d_2}$, and let $y_i = \sqrt{d_1 d_2}\tilde{y}_i$, it is clear that

$$y_i = \langle\langle X_i, \mathbf{\Theta}^* \rangle\rangle + w_i$$

Applications: Low Rank Matrix Completion

In high dimensional setting, $n << d_1 d_2$. The first issue arises when, for example, $\boldsymbol{\Theta}^B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$.

For this matrix, we have $\mathcal{X}_n(\boldsymbol{\Theta}^B) = 0$ with high probability.

To mitigate this issue, the book suggested to impose the so-called matrix incoherence condition, which ensures the singular vectors of $\boldsymbol{\Theta}^*$ are relatively spread out (entries have absolute values close to each other).

Applications: Low Rank Matrix Completion

Rigorously speaking, let $\mathbf{\Theta} = UDV^T$ be its SVD, then columns of U, V are normalized. If the entries of such columns are perfectly spread out, then each entry will have absolute value $1/\sqrt{d_1}$ for U and $1/\sqrt{d_2}$ for V.

As a result, rows of U will have Euclidean norm $\sqrt{r/d_1}$. Note that UU^T has diagonal entries corresponding to the norm of rows of U, so the matrix incoherence condition imposes

$$\|UU^T - \frac{r}{d_1}I\|_{\max} \leq \mu\frac{\sqrt{r}}{d_1}$$

where μ is called the incoherence parameter.

Analogous algebra motivates the other condition $\|VV^T - \frac{r}{d_2}I\|_{\max} \leq \mu\frac{\sqrt{r}}{d_2}$.

Applications: Low Rank Matrix Completion

Another issue before we state the formal result is that the matrix incoherence condition is not robust under noise. As an example, let $\mathbf{z} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^d$, consider $Z^* = \frac{\mathbf{z}^T \mathbf{z}}{d}$, it can be shown Z^* is rank 1 (trivially) and satisfy the incoherence condition with properly chosen μ (details omitted).

Let $\Gamma^* = (1-\delta)Z^* + \delta \mathbf{\Theta}^B$ for some $0 < \delta \leq 1$, we can verify $\mathbf{e}_1^T = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}^T$ is always an eigenvector of Γ^* (trivial: $Z^* \mathbf{e}_1 = \mathbf{0}$), so the incoherence condition is violated.

To address this issue, define the spikeness ratio

$$\alpha_{\text{sp}}(\mathbf{\Theta}) = \frac{\sqrt{d_1 d_2} \|\mathbf{\Theta}\|_{\max}}{\|\mathbf{\Theta}\|_F}$$

Applications: Low Rank Matrix Completion

$$\alpha_{\mathsf{sp}}(\mathbf{\Theta}) = \frac{\sqrt{d_1 d_2}\|\mathbf{\Theta}\|_{\max}}{\|\mathbf{\Theta}\|_F}$$

Since $\|\mathbf{\Theta}\|_F^2 = \sum_{j=1}^{d_1}\sum_{k=1}^{d_2} \mathbf{\Theta}_{jk}^2 \le d_1 d_2 \|\mathbf{\Theta}\|_{\max}^2$, we know $\alpha_{\mathsf{sp}}(\mathbf{\Theta}) \ge 1$.

Since $\|\mathbf{\Theta}\|_{\max} \le \|\mathbf{\Theta}\|_F$, we know $\alpha_{\mathsf{sp}}(\mathbf{\Theta}) \le \sqrt{d_1 d_2}$.

As a remark, for the matrix Γ^*, we have

$$\alpha_{\mathsf{sp}}(\Gamma_\delta^*) \le \frac{(1-\delta)+\delta d}{1-2\delta}$$

Applications: Low Rank Matrix Completion

The following theorem gives a form of the restricted strong convexity condition for the matrix completion problem:

Thm. 10.17, Wainwright 2019

Let $\mathcal{X}_n : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ be the random matrix completion operator formed by n i.i.d. samples of rescaled mask matrices X_i. Then there are universal positive constants (c_1, c_2) such that

$$\left| \frac{1}{n} \frac{\|\mathcal{X}_n(\mathbf{\Theta})\|_2^2}{\|\mathbf{\Theta}\|_F^2} - 1 \right| \le c_1 \alpha_{\mathrm{sp}}(\mathbf{\Theta}) \frac{\|\mathbf{\Theta}\|_{\mathrm{nuc}}}{\|\mathbf{\Theta}\|_F} \sqrt{\frac{d \log d}{n}} + c_2 \alpha_{\mathrm{sp}}^2(\mathbf{\Theta}) \left(\sqrt{\frac{d \log d}{n}} + \delta \right)^2$$

for all non-zero $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, with probability at least $1 - 2e^{-\frac{1}{2} d \log d - n\delta}$.

Applications: Low Rank Matrix Completion

With the theorem above, we have the following result that follows from Prop. 10.6

Cor. 10.18, Wainwright 2019

Consider the n observations of $\tilde{y}_i = \mathbf{\Theta}^*_{a(i),b(i)} + \frac{w_i}{\sqrt{d_1 d_2}}$ such that $\mathbf{\Theta}^*$ is with rank at most r, elementwise bounded as $\|\mathbf{\Theta}^*\|_{\max} \leq \alpha/\sqrt{d_1 d_2}$, and i.i.d. additive noise variables $\{w_i\}_{i=1}^n$ satisfy the Bernstein condition with parameters (σ, b), i.e., $\left|E\left[(w_i - E(w_i))^k\right]\right| \leq \frac{k!}{2}\sigma^2 b^{k-2}, k \geq 2$.
Given a sample size $n > \frac{100b^2}{\sigma^2} d \log d$, if we solve the least square objective function with $\lambda_n^2 = 25\frac{\sigma^2 d \log d}{n} + \delta^2$ for some $\delta \in \left(0, \frac{\sigma^2}{2b}\right)$, then any optimal solution $\widehat{\mathbf{\Theta}}$ satisfies the bound

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_F^2 \leq c_1 \max\left(\sigma^2, \alpha^2\right) r \left(\frac{d \log d}{n} + \delta^2\right)$$

with probability at least $1 - e^{-\frac{n\delta^2}{16d}} - 2e^{-\frac{1}{2}d \log d - n\delta}$.

Thank you!