Tail bounds for Matrix De-noising application

Kevin Kim[1], Dong Hu[1], Georgios Mavroudeas[1]

[1]Department of Computer Science
Rensselaer Polytechnic Institute

December, 2019

Motivation I

- **Overview:** Matrix de-noising is one of the popular topics in Machine Learning. Many existing papers have focused on the assumption that the data matrix has low rank. In this presentation we want to focus on one specific de-noising algorithm on low rank matrices and analyze it from the probability perspective.

Motivation II

One important application of matrix de-noising is image de-noising

(a) Ground Truth (b) Noisy Input($\sigma = 30$) (c) De-noised Image

Figure: Example of image de-noising

Problem Formulation

Oracle: We can sample any matrix with noise where the noise has distribution $\mathcal{N}(0, \sigma^2)$ for some σ.

Problem: For any low-rank matrix $\boldsymbol{A}^{m\times n}$ with $rk(\boldsymbol{A}) = r \ll \min\{m, n\}$, following the oracle we get an noisy matrix $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{G}$ where $[\boldsymbol{G}^{m\times n}]_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Instead of using $\boldsymbol{C}$ as an approximation of $\boldsymbol{A}$, we want a better approximation. Denote estimator $\hat{\boldsymbol{A}}$ where $\hat{\boldsymbol{A}} = \boldsymbol{C}\boldsymbol{X}$ and the linear operator matrix $\boldsymbol{X}^{n\times n}$. In order to get the best approximation we are trying to solve the optimization problem

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \mathbb{E}\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_F^2 \\ \text{s.t.} \quad & \hat{\boldsymbol{A}} = \boldsymbol{C}\boldsymbol{X} \\ & rk(\hat{\boldsymbol{A}}) \leq r \end{aligned}$$

Problem Solution

It can be shown that

$$\boldsymbol{X}^* = \arg\min_{\boldsymbol{X}} \mathbb{E} \left\| \hat{\boldsymbol{A}} - \boldsymbol{A} \right\|_F^2 \tag{1}$$

$$= (\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I})^{-\frac{1}{2}} \left[(\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{A}^\top \boldsymbol{A} \right]_r \tag{2}$$

where $[\boldsymbol{M}]_r = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top$ is the rank-r truncation of $\boldsymbol{M}$ if $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is its SVD.

Proposed Solution

$$\boldsymbol{X}^* = (\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I})^{-\frac{1}{2}} \left[(\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{A}^\top \boldsymbol{A} \right]_r$$

WARNING: We are not allowed to use $\boldsymbol{A}$ in our expression!!!

Proposed Solution

Observation:

$$\mathbb{E}\left[\boldsymbol{C}^\top \boldsymbol{C}\right] = \boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I}$$

Our estimated linear operator:

$$\tilde{\boldsymbol{X}} = (\boldsymbol{C}^\top \boldsymbol{C})^{-\frac{1}{2}} \left[(\boldsymbol{C}^\top \boldsymbol{C})^{-\frac{1}{2}} (\boldsymbol{C}^\top \boldsymbol{C} - \sigma^2 m \boldsymbol{I})\right]_r$$

We use $\tilde{\boldsymbol{A}} = \boldsymbol{C}\tilde{\boldsymbol{X}}$ to approximate $\hat{\boldsymbol{A}} = \boldsymbol{C}\boldsymbol{X}^*$.

Challenge: Is this estimation accurate?

Probability to the Rescue

Bounding our Solution

- The main challenge is to find probability tail bounds that shows our solution concentrates around and is unlikely to deviate far from the optimal solution.
- Namely, we want to show

$$\mathbb{E}\left\|\boldsymbol{C}\tilde{\boldsymbol{X}} - \boldsymbol{C}\hat{\boldsymbol{X}}\right\|_2 < \varepsilon$$

  w.h.p for some small ε.
- Notice

$$\mathbb{E}\left\|\boldsymbol{C}\tilde{\boldsymbol{X}} - \boldsymbol{C}\hat{\boldsymbol{X}}\right\|_2 \leq \mathbb{E}\left[\|\boldsymbol{C}\|_2 \left\|\tilde{\boldsymbol{X}} - \hat{\boldsymbol{X}}\right\|_2\right]$$

Main Theory I

Probability Tools Overview

Here some of the existing probability tools will be presented for scalar random variables.

- Sub-Gaussian Bounds:
  - A random variable X is sub-Gaussian if it satisfies: $\forall \lambda \in R$

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}$$

and the concentration bound is:

$$\mathbb{P}\left[|X-\mu| \geq t\right] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

- Sub-exponential Bounds:
  - A random variable X is sub-exponential if it satisfies: $\exists v, \alpha > 0$, such that $\forall |\lambda| < \frac{1}{\alpha}$

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{v^2\lambda^2}{2}}$$

and it's a generalization of the sub-Gaussian bound

Main Theory II

Probability Tools Overview

- Bernstein Bounds:
  - Can be applied to sum of series of centered bounded i.i.d random variables. Let $S_1, ..., S_n$ be random variables with $\mathbb{E}[S_i] = 0$ and $|S_i| \leq L$ and let

$$Z = \sum_{i=1}^{n} S_i, \;\; Var(Z) = \sum_{i=1}^{n} Var(S_i)$$

then we have

$$\mathbb{P}(|Z| \geq t) \leq 2e^{-\frac{t^2/2}{Var(Z)+Lt/3}}$$

Matrix Concentration Inequalities I

From Scalar Random Variables to Random Matrices

- In general, it is hard to extend the bounds derived for scalar random variables to random matrices. In our case if we can represent our random matrices by a sum of bounded i.i.d random matrices, we can apply the matrix Bernstein bound.

Matrix Concentration Inequalities II

From Scalar Random Variables to Random Matrices

Matrix Bernstein bound If $\boldsymbol{S}_1, ..., \boldsymbol{S}_n$ are centered i.i.d random matrices such that $\forall i, \boldsymbol{S}_i \in R^{d_1 x d_2}$ and $||\boldsymbol{S}_i|| \leq L$ and

$$\boldsymbol{Z} = \sum_{i=1}^{n} \boldsymbol{S}_i$$

If we denote $V(\boldsymbol{Z})$ as the co-variance matrix where:

$$V(\boldsymbol{Z}) = \max\{||\mathbb{E}[\boldsymbol{Z}^* \boldsymbol{Z}]||, ||\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^*]||\}$$

Then the concentration bound is:

$$\mathbb{P}(||\boldsymbol{Z}|| \geq t) \leq (d_1 + d_2) e^{-\frac{t^2/2}{V(\boldsymbol{Z}) + Lt/3}}$$

Also we can get a bound for the expectation of $\boldsymbol{Z}$:

$$\mathbb{E}[||\boldsymbol{Z}||] \leq \sqrt{2V(\boldsymbol{Z}) \log(d_1 + d_2)} + \frac{1}{3} L \log(d_1 + d_2)$$

Matrix Concentration Inequalities III

From Scalar Random Variables to Random Matrices

Another class of useful bounds ideal for our case, are the bounds derived from Wishart random matrices. These are random matrices where each row is multivariate gaussian random variable $N(0, \boldsymbol{\Sigma})$. This is exactly what the noise matrix $\boldsymbol{G}$ is. If we denote $\boldsymbol{X}$ a Wishart random variable with co-variance matrix $\boldsymbol{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$ then some existing bounds are:

$$\mathbb{P}\left[\frac{\sigma_{max}(\boldsymbol{X})}{\sqrt{n}} \geq \gamma_{max}(\sqrt{\boldsymbol{\Sigma}})(1+\delta) + \sqrt{(\frac{tr(\boldsymbol{\Sigma})}{n})}\right] \leq e^{-\frac{n\delta^2}{2}} \qquad (3)$$

Not sure if this slide is necessary, maybe useful for bounding $\boldsymbol{G}$

Using Random Matrix Bounds I

Our Attempt to Bound the Solution

We can simplify the optimal and the proposed solution to

$$\boldsymbol{X}^* = \left[\boldsymbol{I} - \sigma^2 m(\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I})^{-1}\right]_r$$
$$\tilde{\boldsymbol{X}} = \left[\boldsymbol{I} - \sigma^2 m(\boldsymbol{C}^\top \boldsymbol{C})^{-1}\right]_r$$

So for $\tilde{\boldsymbol{X}}$ to be concentrating around $\boldsymbol{X}^*$ it has to that $\boldsymbol{C}^\top \boldsymbol{C}$ concentrates around $\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I}$

Using Random Matrix Bounds II

Our Attempt to Bound the Solution

We will try to show that

$$\left\| \boldsymbol{C}^\top \boldsymbol{C} - (\boldsymbol{A}^\top \boldsymbol{A} + \sigma^2 m \boldsymbol{I}) \right\| \tag{4}$$

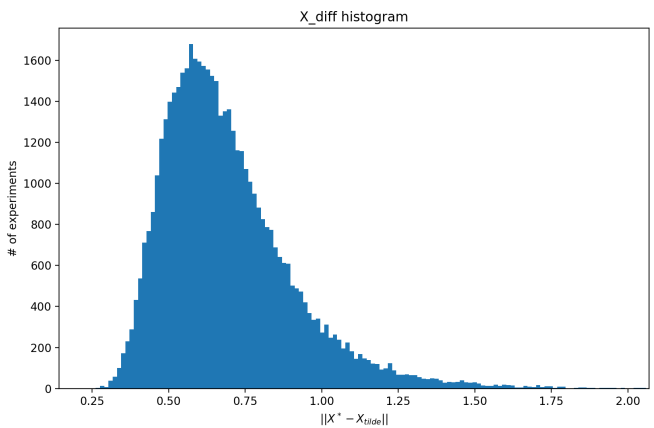
$$= \left\| \boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{G}^\top \boldsymbol{A} + \boldsymbol{G}\boldsymbol{A}^\top + \boldsymbol{G}^\top \boldsymbol{G} - \boldsymbol{A}^\top \boldsymbol{A} - \sigma^2 m \boldsymbol{I} \right\| \tag{5}$$

$$\leq 2 \left\| \boldsymbol{A} \right\| \left\| \boldsymbol{G} \right\| + \left\| \boldsymbol{G}^\top \boldsymbol{G} - \sigma^2 m \boldsymbol{I} \right\| \tag{6}$$

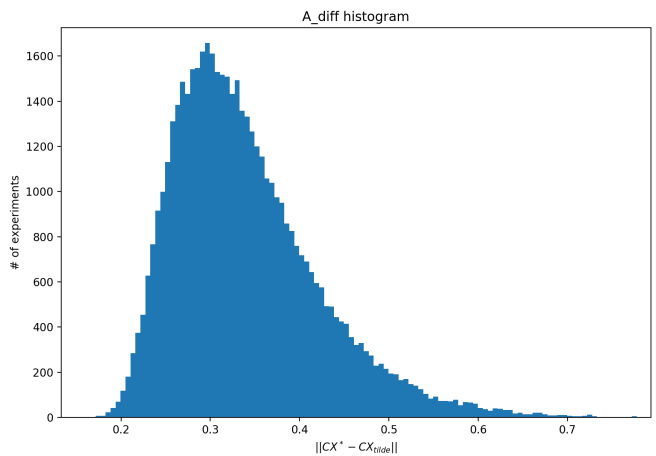
Empirical Results

Experiments

Experiment settings: $\boldsymbol{A} \sim \mathcal{N}(5, 1)$, $\boldsymbol{G} \sim \mathcal{N}(0, 0.01)$. We sample $\boldsymbol{A}$ for 50000 times and plot the histogram for both $\left\|\boldsymbol{X}^* - \tilde{\boldsymbol{X}}\right\|_2$ and $\left\|\boldsymbol{C}\boldsymbol{X}^* - \boldsymbol{C}\tilde{\boldsymbol{X}}\right\|_2$



(a) $\left\|\boldsymbol{X}^* - \tilde{\boldsymbol{X}}\right\|_2$



(b) $\left\|\boldsymbol{C}\boldsymbol{X}^* - \boldsymbol{C}\tilde{\boldsymbol{X}}\right\|_2$

Conclusion

- Both the experiment and the theory shows that if $\sigma_g < \frac{\sigma_r(\boldsymbol{A})}{\sqrt{m}\|\boldsymbol{A}\|_2}$ we have

$$\mathbb{P}\left(\left\|\boldsymbol{C}\boldsymbol{X}^* - \boldsymbol{C}\tilde{\boldsymbol{X}}\right\|_2 \geq \varepsilon\right) \leq \delta$$

ε, δ and σ_g are unsure and need to be determined but there should be some according to theory