1. Comparison of works

Table 1: Related regularization work. Let dimension of Hankel be n, system order be R and number of samples be T. Error in suboptimal Hankel spectral norm comes from naive matrix inequality of Frobenius norm.

Paper	This work	Cai et al. (2016)
Sample complexity for optimal Hankel spectral norm	$T = O(R^2 \log^2 n)$	
Sample complexity for suboptimal Hankel spectral norm	$T = O(R \log^2 n)$	$T = O(R \log^2 n)$
Input	Multiple rollout	Multiple rollout
Error in impulse response Frobenius norm		$O(\sigma\sqrt{n/T})$
Error in optimal Hankel spectral norm	$O(\sigma\sqrt{n/T})$	
Error in suboptimal Hankel spectral norm	$O(\sigma\sqrt{n^2/T})$	$O(\sigma\sqrt{n^2/T})$

Table 2: Related least square work. Let dimension of Hankel be n, system order be R and number of samples be T.

Paper	This work	Oymak and Ozay (2018)	Sarkar et al. (2019)
Sample complexity	$T = O(n)$	$T = O(n)$	$T = O(n^2)$
Input	Single rollout	Single rollout	Single rollout
Error in impulse response Frobenius norm		$O(\sigma\sqrt{n/T})$	$O(\sigma\sqrt{n/T})$
Error in Hankel spectral norm	$O(\sigma\sqrt{n/T})$	$O(\sigma\sqrt{n^2/T})$	$O(\sigma\sqrt{n/T})$

2. More experiments

First we generate synthetic data and compare the performance of Oymak and Ozay (2018) and Sarkar et al. (2019) in Figure 1. We can see that, due to the constant overhead $O(\frac{1}{1-\rho}\sqrt{n/T})$ in Sarkar et al. (2019) algorithm, the resulted error is larger than Oymak and Ozay (2018). Figure 2 compares them in the setting when output noise exists and Oymak and Ozay (2018) has smaller error as well.

In this subsection, we check Theorem 7 via synthetic experiments, and compare with least square estimator. In the following experiment, we have a fixed strictly stable SISO linear system with order 9, the Hankel size n is initiated as 20 which exceeds the order. The input is multiple rollout, scaled i.i.d Gaussian, which means that we send in the input up to time $2n-1$, and observe the output at the end as an observation, and restart the system. The input satisfies that, after scaling by K^{-1}, $\mathbf{E}(\boldsymbol{U}^T\boldsymbol{U}) = I$, which is the assumption in Theorem 7. The observed output can be noiseless and noisy, and the numbers of observations are 30 (undetermined for least square) and 60 (determined for least square).

We tune the regularized model by training with different weight λ of regularization. To tune the least square model, there are two ways: (1) fix the size of Hankel matrix, and run Ho-Kalman algorithm with different rank truncation, or (2) change the size of Hankel matrix. We pick the model associated with the smallest validation error at the end, and run it on test set. The size of training, validation and test set is $1:3:6$.

2.1. Noiseless, enough observations (Fig 3 and 4)

When the output is noiseless and $T = 60$, we can see that both regularized and least square algorithms do well. When $\lambda \to 0$ in regularization or the size and rank tends to 20 in least square method, it almost perfectly fit the model. The singular values of the estimated Hankel is the same since it is perfect recovery.

2.2. Noisy, enough observations (Fig 5 and 6)

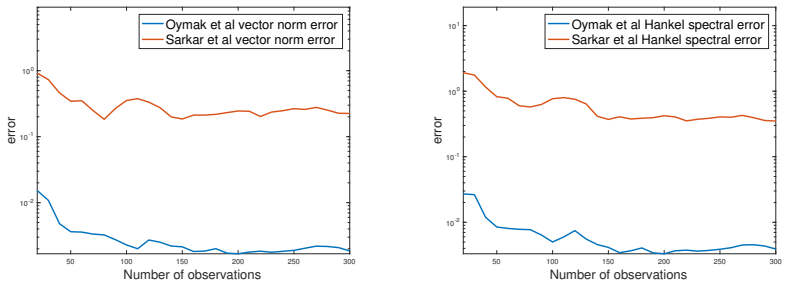
With enough data, when the output is noisy, both regularization and least square do the job well. In Figure 5, we can see that in terms of validation error, there is a best weight λ and Hankel size n, below and above which the validation error both grow. Then we can pick the optimizer associated with those weight, size or rank as our estimation of the system.

2.3. Noiseless, not enough observations (Fig 7 and 8)

Without enough data for least square, even if the output is noiseless, least square is underdetermined, even if we take the solution with the smallest 2 norm in impulse response, it suffers big error on validation and test set. However, the error of regularization remains small and as λ getting small, the error still tends 0. It indicates that, the solution with the least Hankel nuclear norm behaves better than least impulse Frobenius norm in low sample complexity case.

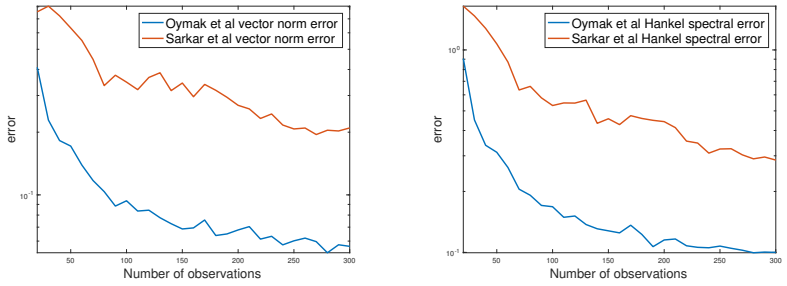
2.4. Noisy, not enough observations (Fig 9 and 10)

Finally not enough data and noisy. We can see that regularized algorithm is robust to noise, where as least square algorithms remain bad.



(a) (b)

Figure 1: Comparison of (a) impulse Frobenius norm (b) Hankel spectral norm error when output is noiseless between Oymak and Ozay (2018) and Sarkar et al. (2019) with synthetic data. System is randomly generated with order 9 and Hankel $H \in \mathbb{R}^{9\times 9}$. Single trajectory and input is i.i.d. Gaussian.



(a) (b)

Figure 2: Comparison of (a) impulse Frobenius norm (b) Hankel spectral norm error when output SNR is 10 between Oymak and Ozay (2018) and Sarkar et al. (2019) with synthetic data. System is randomly generated with order 9 and Hankel $H \in \mathbb{R}^{9\times 9}$. Single trajectory and input is i.i.d. Gaussian.

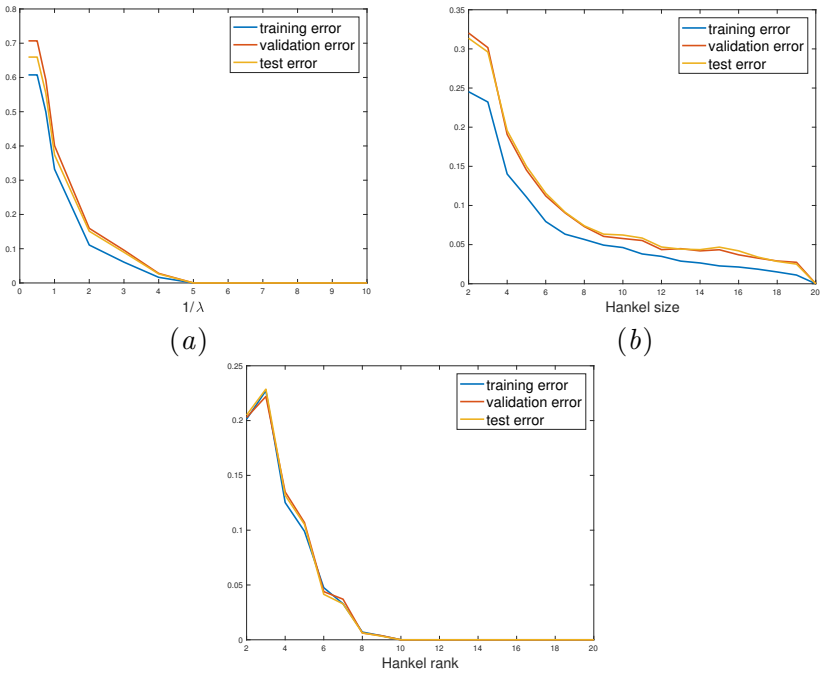


Figure 3: System estimation for synthetic data, noiseless, assuming $n = 20$. Training data size $= 60$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

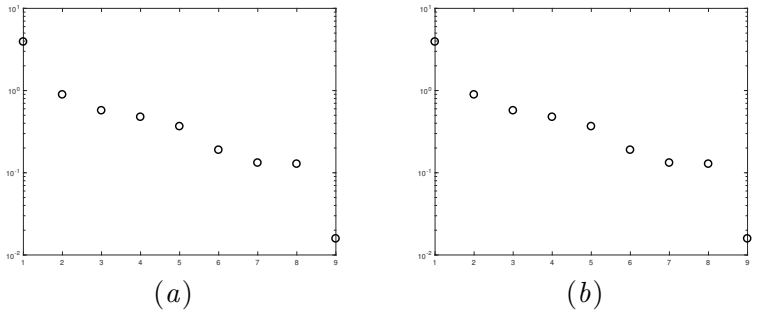


Figure 4: Synthetic, $SNR = 10$, training size is 60, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

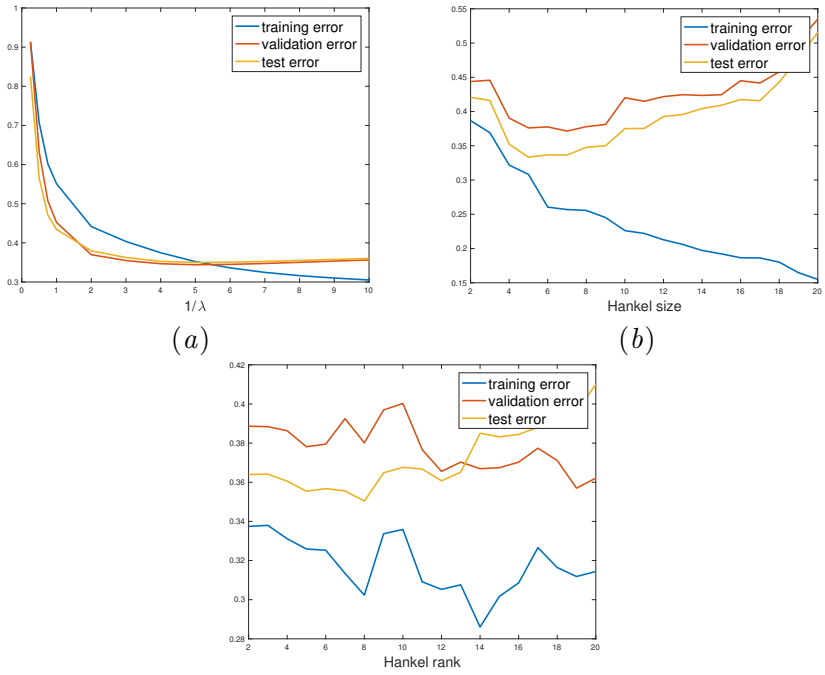


Figure 5: System estimation for synthetic data, SNR = 10, assuming $n = 20$. Training data size = 60. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

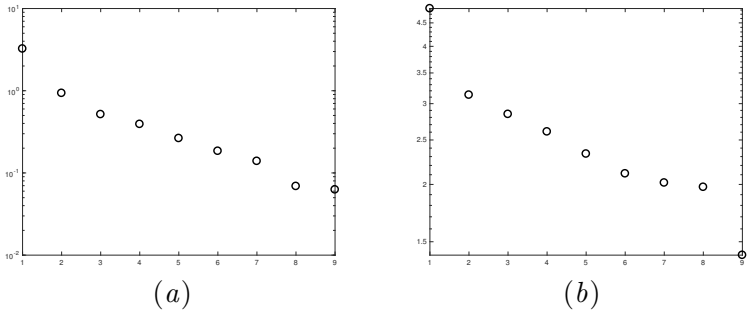


Figure 6: Synthetic, $SNR = 10$, training size is 60, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

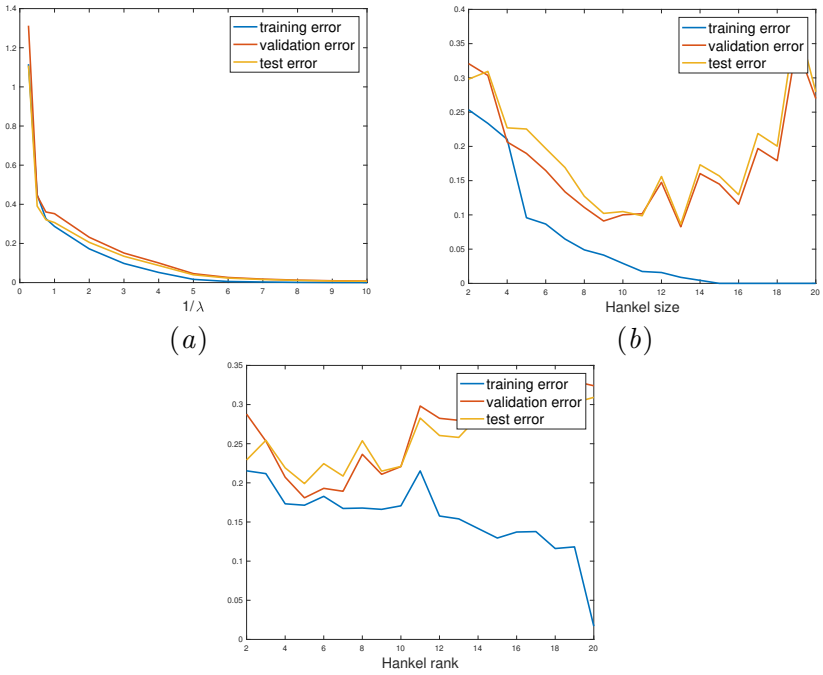


Figure 7: System estimation for synthetic data, noiseless, assuming $n = 20$. Training data size $= 30$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

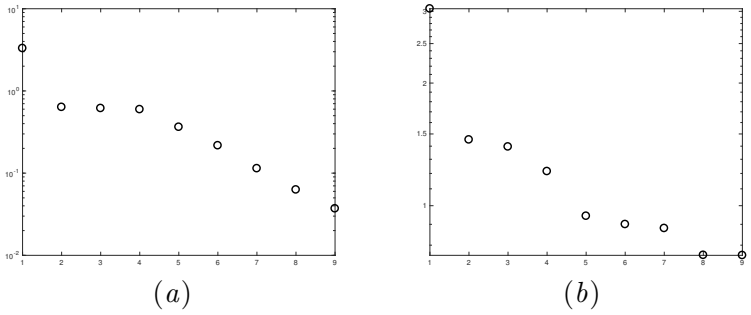


Figure 8: Synthetic, noiseless, training size is 30, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

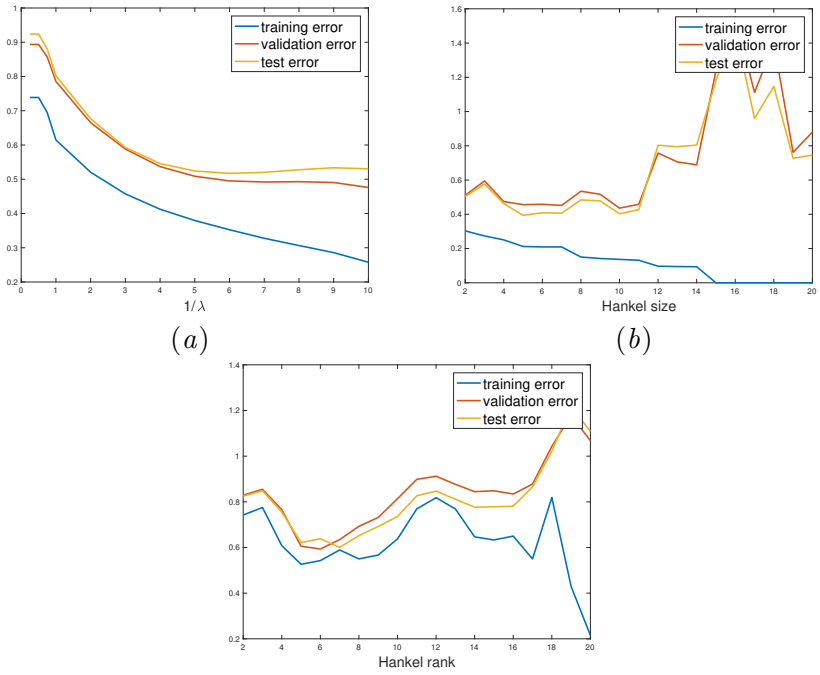


Figure 9: System estimation for synthetic data, SNR = 10, assuming $n = 20$. Training data size = 30. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. (c) Training and validation error of different Hankel rank with same size $n = 20$.

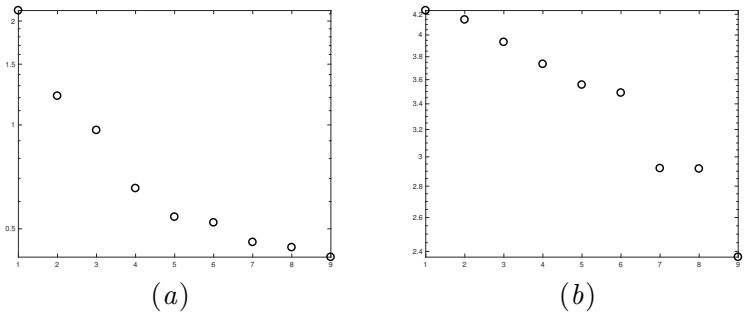


Figure 10: Synthetic, $SNR = 10$, training size is 30, singular value of (a) unregularized Hankel before rank truncation (b) regularized Hankel.

3. Proof of least square spectral norm error

Theorem 1 *Denote the discrete Fourier transform matrix by F. Denote $\xi_{(i)} \in \mathbb{R}^T, i = 1, ..., m$ as the noise that corresponds to each dimension of output. The solution $\hat{h}$ of*

$$\hat{h} := h + \boldsymbol{U}^\dagger \xi = \min_{h'} \frac{1}{2}\|\boldsymbol{U}h' - y\|_F^2. \tag{1}$$

obeys

$$\|\hat{h} - h\|_F \leq \|\xi\|_F / \sigma_{\min}(\boldsymbol{U})$$
$$\|\mathcal{H}(\hat{h} - h)\|_2 \leq \left\| \left[\|F\boldsymbol{U}^\dagger \xi_{(1)}\|_\infty, ..., \|F\boldsymbol{U}^\dagger \xi_{(m)}\|_\infty \right] \right\|_2.$$

Proof (1) has close form solution and we have $\|\hat{h} - h\| = \|\boldsymbol{U}^\dagger \xi\| \leq \|\xi\|/\sigma_{\min}(\boldsymbol{U})$. To get the error bound in Hankel matrix, we can denote $\bar{\xi} = \boldsymbol{U}^\dagger \xi = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\xi$, and

$$H_{\bar{\xi}} = \begin{bmatrix} \bar{\xi}_1 & \bar{\xi}_2 & \cdots & \bar{\xi}_{2n-1} \\ \bar{\xi}_2 & \bar{\xi}_3 & \cdots & \bar{\xi}_1 \\ \cdots & & & \\ \bar{\xi}_{2n-1} & \bar{\xi}_1 & \cdots & \bar{\xi}_{2n-2} \end{bmatrix}.$$

If $m = 1$, $\bar{\xi} \in \mathbb{R}^{(2n-1)p}$ is a vector (Krahmer et al., 2014, Section 4) proves that

$$H_{\bar{\xi}} = F^{-1}\mathrm{diag}(F\bar{\xi})F.$$

So the spectral norm error is bounded by $\|\mathrm{diag}(F\bar{\xi})\|_2 = \|F\bar{\xi}\|_\infty$.

If $m > 1$, all columns of ξ are independent, so $H_{\bar{\xi}}$ can be seen as concatenation of m independent noise matrices where each satisfies the previous derivation. ■

Theorem 2 *Denote the solution to (1) as $\hat{h}$. Let $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$ is multiple rollout input, where every entry is i.i.d. Gaussian random variable, y be the corresponding output and ξ is i.i.d. Gaussian matrix with each entry has mean 0 and variance σ_ξ, then the spectral norm error is $\|\mathcal{H}(\hat{h} - h)\| \sim O(\sigma_\xi \sqrt{\frac{mnp}{T}} \log(np))$.*

Proof We use Theorem 1. First let $m = 1$. The covariance of $F\bar{\xi} = F\boldsymbol{U}^\dagger\xi$ is $F(\boldsymbol{U}^T\boldsymbol{U})^{-1}F^T$. If $T = \tilde{\Omega}(n)$, it's proven Vershynin (2018) that $\frac{TI}{2} \preceq \boldsymbol{U}^T\boldsymbol{U} \preceq \frac{3TI}{2}$ then $\frac{n}{2T}I \preceq F(\boldsymbol{U}^T\boldsymbol{U})^{-1}F^T \preceq \frac{3n}{2T}I$. So $\|F\bar{\xi}\|_\infty$ should scale as $O(\sigma_\xi\sqrt{\frac{n}{T}}\log n)$. So $\|\mathcal{H}(\bar{\xi})\|_2 \leq \|H_{\bar{\xi}}\|_2 \leq \|F\bar{\xi}\|_\infty = O(\sigma_\xi\sqrt{\frac{n}{T}}\log n)$. If $m > 1$, then by concatenation we simply bound the specral norm by m times MISO case. When $m > 1$, with previous discussion of concatenation, and each submatrix to be concatenated has the same distribution, so the spectral norm error is at most $\sqrt{m}$ times larger. ■

4. Gaussian width of nuclear norm tangent cone in MISO.

We consider recovering a MISO system impulse response. We first calculate the minimum number of observations needed to recover the system regardless of noise rate, which is a simple extension from SISO case in Cai et al. (2016). This can be seen as the sample complexity requirement in noiseless case. For multi-rollout case, we only observe the output at time $2N-1$, we have

$$y_{2N-1} = \sum_{i=1}^{2N-2} CA^{2N-2-i}Bu_i + Du_{2N-1}. \tag{2}$$

Denote the impulse response by $h \in \mathbb{R}^{p(2N-1)}$, which is a block vector

$$h = \begin{bmatrix} h^{(1)} \\ h^{(2)} \\ \dots \\ h^{(2N-1)} \end{bmatrix}$$

where each block $h^{(i)} \in \mathbb{R}^p$. $\beta \in \mathbb{R}^{p(2N-1)}$ is a weighted version of h, with weights

$$K_j = \begin{cases} \sqrt{j}, & 1 \le j \le N \\ \sqrt{2N-j}, & N < j \le 2N-1 \end{cases}$$

and

$$x^{(i)} = K_i h^{(i)}$$

Define the reweighted Hankel map for the same h by

$$\mathcal{G}(\beta) = \begin{bmatrix} \beta^{(1)}/K_1 & \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & \dots \\ \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & \beta^{(4)}/K_2 & \dots \\ \dots & & & \end{bmatrix}^T \in \mathbb{R}^{N \times pN}$$

and $\mathcal{G}^*$ is the adjoint of $\mathcal{G}$. We define each rollout input $u_1, \dots, u_{2N-1}$ as independent Gaussian vectors with

$$u_i \sim \mathcal{N}(0, K_i^2 \mathbf{I})$$

Now let $\boldsymbol{U} \in \mathbb{R}^{T \times p(2N-1)}$, each entry is iid standard Gaussian. We consider the question

$$\begin{aligned} \min_{\beta} \quad & \|\mathcal{G}(\beta)\|_* \\ \text{s.t.,} \quad & \|\boldsymbol{U}\beta - y\|_2 \le \delta \end{aligned} \tag{3}$$

where the norm of overall (state and output) noise is bounded by δ.

Theorem 3 *Let $\hat{\beta}$ be the true impulse response. If $T = \Omega((\sqrt{pR}\log(N) + \epsilon)^2)$, C is some constant, the solution β^* to (3) satisfies $\|\hat{\beta} - \beta^*\|_2 \le 2\delta/\epsilon$ with probability*

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log(N) + \epsilon) - \epsilon)^2\right).$$

Let $\mathcal{I}(\beta)$ be the descent cone of $\|\mathcal{G}(\beta)\|_*$ at β, we have the following lemma:

Lemma 4 *Assume*

$$\min_{z\in\mathcal{I}(\hat{\beta})} \frac{\|\boldsymbol{U}z\|_2}{\|z\|_2} \geq \epsilon,$$

then $\|\hat{\beta}-\beta^*\|_2 \leq 2\delta/\epsilon$.

(Proof omitted) To prove Theorem 3, we only need lower bound LHS with Lemma 4. The following lemma gives the probability that LHS is lower bounded.

Lemma 5 *Define the Gaussian width*

$$w(S) := E_g(\sup_{\gamma\in S} \gamma^T g) \tag{4}$$

where g is standard Gaussian vector of size p. Let $\Phi = \mathcal{I}(\hat{\beta})\cap\mathbb{S}$ where $\mathbb{S}$ is unit sphere. We have

$$P(\min_{z\in\Phi} \|\boldsymbol{U}z\|_2 < \epsilon) \leq \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right). \tag{5}$$

Now we need to study $w(\Phi)$.

Lemma 6 *(Cai et al. (2016) eq. (17)) Let $\mathcal{I}^*(\beta)$ be the dual cone of $\mathcal{I}(\beta)$, then*

$$w(\Phi) \leq E(\min_{\gamma\in\mathcal{I}^*(\hat{\beta})} \|g-\gamma\|_2). \tag{6}$$

Note that $\mathcal{I}^*(\hat{\beta})$ is just the cone of subgradient of $\mathcal{G}(\hat{\beta})$, so it can be written as

$$\mathcal{I}^*(\hat{\beta}) = \{\mathcal{G}^*(V_1V_2^T + W) | V_1^TW = 0, WV_2 = 0, \|W\|_2 \leq 1\}$$

where $\mathcal{G}(\hat{\beta}) = V_1\Sigma V_2^T$ is the SVD of $\mathcal{G}(\hat{\beta})$[1]. So

$$\min_{\gamma\in\mathcal{I}^*(\hat{x})} \|g-\gamma\|_2 = \min_{\lambda,W} \|\lambda\mathcal{G}^*(V_1V_2^T+W) - g\|_2.$$

For RHS, we have

$$\begin{aligned}\|\lambda\mathcal{G}^*(V_1V_2^T+W) - g\|_2 &= \|\lambda\mathcal{G}\mathcal{G}^*(V_1V_2^T+W) - \mathcal{G}(g)\|_F \\ &= \|\lambda(V_1V_2^T+W) - \mathcal{G}(g)\|_F + \|\lambda(I-\mathcal{G}\mathcal{G}^*)(V_1V_2^T+W)\|_F \\ &\leq \|\lambda(V_1V_2^T+W) - \mathcal{G}(g)\|_F.\end{aligned}$$

Let $\mathcal{P}_W$ be projection operator onto subspace spanned by W, i.e.,

$$\{W | V_1^TW = 0, WV_2 = 0\}$$

1. For simplicity, we only write down real case. Complex case can be seen as a dimension increase by 2 times as in Cai et al. (2016).

and $\mathcal{P}_V$ be projection onto its orthogonal complement. Choose $\lambda = \|\mathcal{P}_W(\mathcal{G}(g))\|_2$ and $W = \mathcal{P}_W(\mathcal{G}(g))/\lambda$.

$$\begin{aligned}\|\lambda(V_1V_2^T + W) - \mathcal{G}(g)\|_F &= \|\mathcal{G}(g) - \mathcal{P}_W(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\|_2 V_1V_2^T\|_F \\ &\le \|\mathcal{P}_V(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\|_2 V_1V_2^T\|_F \\ &\le \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \|\mathcal{P}_W(\mathcal{G}(g))\|_2\|V_1V_2^T\|_F \\ &= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{P}_W(\mathcal{G}(g))\|_2 \\ &= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{G}(g)\|_2.\end{aligned}$$

Bound the first term by (note V_1 and V_2 span R dimensional space, so $V_1 \in \mathbb{R}^{N\times R}$ and $V_2 \in \mathbb{R}^{pN\times R}$)

$$\begin{aligned}\|\mathcal{P}_V(\mathcal{G}(g))\|_F &= \|V_1V_1^T\mathcal{G}(g) + (I - V_1V_1^T)\mathcal{G}(g)V_2V_2^T\|_F \\ &\le \|V_1V_1^T\mathcal{G}(g)\|_F + \|\mathcal{G}(g)V_2V_2^T\|_F \\ &\le 2\sqrt{R}\|\mathcal{G}(g)\|_2.\end{aligned}$$

we get

$$\begin{aligned}w(\Phi) &\le E(\min_{\lambda,W}\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2) \\ &\le E(\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2)\big|_{\lambda=\|\mathcal{P}_W(\mathcal{G}(g))\|_2, W=\mathcal{P}_W(\mathcal{G}(g))/\lambda} \\ &\le 3\sqrt{R}\|\mathcal{G}(g)\|_2.\end{aligned}$$

We know that, if $p = 1$, then $E\|\mathcal{G}(g)\|_2 = O(\log(N))$. For general p, let

$$g^{(i)} = [g_1^{(i)}, ..., g_p^{(i)}]^T,$$

we rearrange the matrix as

$$\begin{aligned}\bar{\mathcal{G}}(g) &= \begin{bmatrix}\begin{bmatrix} g_1^{(1)} & g_1^{(2)}/\sqrt{2} & ... \\ g_1^{(2)}/\sqrt{2} & g_1^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & \begin{bmatrix} g_2^{(1)} & g_2^{(2)}/\sqrt{2} & ... \\ g_2^{(2)}/\sqrt{2} & g_2^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & ...\end{bmatrix} \\ &= [G_1, ..., G_p]\end{aligned}$$

where expectation of operator norm of each block is $\log(N)$. Then (note v below also has a block structure $[v^{(1)}; ...; v^{(N)}]$)

$$\begin{aligned}\|\bar{\mathcal{G}}(g)\| &= \max_{u,v}\frac{u^T\bar{\mathcal{G}}(g)v}{\|u\|\|v\|} \\ &= \max_{u,v^1,...,v^p}\sum_{i=1}^p\frac{u^TG_iv^{(i)}}{\|u\|\|v\|} \\ &\le \max_{v^1,...,v^p} O(\log(N))\frac{\sum_{i=1}^p\|v^{(i)}\|}{\sqrt{\sum_{i=1}^p\|v^{(i)}\|^2}} \\ &\le O(\sqrt{p}\log(N)).\end{aligned}$$

And $\|\bar{\mathcal{G}}(g)\|_2 = \|\mathcal{G}(g)\|_2$. So we have $\|\mathcal{G}(g)\|_2 = \sqrt{p}\log(N)$. So $w(\Phi) = C\sqrt{pR}\log(N)$. Get back to (5), we want the probability be smaller than 1, and we get

$$\sqrt{T-1} - C\sqrt{pR}\log N - \epsilon > 0$$

thus $T = O((\sqrt{pR}\log(n) + \epsilon)^2)$.

5. Proof of main theorem

Theorem 7 *We study the problem*

$$\min_{\hat{\beta}} \quad \frac{1}{2}\|\boldsymbol{U}\hat{\beta} - y\|^2 + \lambda\|\mathcal{G}(\hat{\beta})\|_*, \tag{7}$$

in the MISO (multi-input single-output) setting (m=1, p inputs), where $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$. Let β denote the (weighted) impulse response of the true system which has order R, i.e., $\operatorname{rank}(\mathcal{G}(\beta)) = R$, and let $y = \boldsymbol{U}\beta + \xi$ be the measured output, where ξ is the measurement noise. Finally, denote the minimizer of (7) by $\hat{\beta}$. Define

$$\mathcal{J}(\beta) := \left\{ v \;\middle|\; \langle v, \partial(\frac{1}{2}\|\boldsymbol{U}^T\beta - y\|^2 + \lambda\|\mathcal{G}(\beta)\|_*)\rangle \le 0 \right\},$$
$$\Gamma := \|I - \boldsymbol{U}^T\boldsymbol{U}\|_{2,\mathcal{J}(\beta)},$$

$\mathcal{J}(\beta)$ is the tangent cone at β, and Γ is the spectral RSV. If $\Gamma < 1$, $\hat{\beta}$ satisfies

$$\|\mathcal{G}(\hat{\beta} - \beta)\|_2 \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2 + \lambda}{1 - \Gamma}.$$

Proof Now we bound $\|\mathcal{G}(\hat{\beta} - \beta)\|_2$ by partitioning it to $\|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\|_2$ and $\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\|_2$. We have

$$\begin{aligned}\|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\|_2 &= \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\mathcal{G}(\hat{\beta} - \beta)\|_2 \\ &\le \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}\|\mathcal{G}(\hat{\beta} - \beta)\|_2 \\ &= \Gamma\|\mathcal{G}(\hat{\beta} - \beta)\|_2.\end{aligned} \tag{8}$$

And then we also have

$$\begin{aligned}\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\|_2 &= \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y + \xi)\|_2 \\ &\le \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y)\|_2 + \|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2.\end{aligned}$$

Since $\hat{\beta}$ is the optimizer, we have

$$\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y) + \lambda\mathcal{G}^*(\hat{V}_1\hat{V}_2^T + \hat{W}) = 0,$$

where $\mathcal{G}(\hat{\beta}) = \hat{V}_1\hat{\Sigma}\hat{V}_2^T$ is the SVD of $\mathcal{G}(\hat{\beta})$, $\hat{W} \in \mathbb{R}^{n\times n}$ where $\hat{V}_1^T\hat{W} = 0, \hat{W}\hat{V}_2 = 0, \|\hat{W}\|_2 \le 1$. Then

$$\|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y)\|_2 = \lambda\|\mathcal{G}\mathcal{G}^*(\hat{V}_1\hat{V}_2^T + \hat{W})\|_2 \le \lambda. \tag{9}$$

Combine with (9), we have

$$\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta}-\beta))\|_2 \leq \|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2 + \lambda. \tag{10}$$

Combining (8) and (10), we have

$$\begin{aligned}\|\mathcal{G}(\hat{\beta}-\beta)\|_2 &\leq \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})(\hat{\beta}-\beta)\|_2 + \|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta}-\beta))\|_2\\ &\leq \Gamma\|\mathcal{G}(\hat{\beta}-\beta)\|_2 + \|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2 + \lambda\end{aligned}$$

or equivalently,

$$\|\mathcal{G}(\hat{\beta}-\beta)\|_2 \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2+\lambda}{1-\Gamma}, \quad \Gamma = \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}.$$

Bounding Γ. First we prove some side results for later use. From optimality of $\hat{\beta}$, we have

$$\begin{aligned}&\frac{1}{2}\|y-\boldsymbol{U}\hat{\beta}\|^2+\lambda\|\mathcal{G}\hat{\beta}\|_* \leq \frac{1}{2}\|y-\boldsymbol{U}\beta\|^2+\lambda\|\mathcal{G}\beta\|_* = \frac{1}{2}\|\xi\|^2+\lambda\|\mathcal{G}\beta\|_*\\ \Rightarrow\quad &\frac{1}{2}\|\boldsymbol{U}\beta+\xi-\boldsymbol{U}\hat{\beta}\|^2+\lambda\|\mathcal{G}\hat{\beta}\|_* \leq \frac{1}{2}\|\xi\|^2+\lambda\|\mathcal{G}\beta\|_*\\ \Rightarrow\quad &\frac{1}{2}\|\boldsymbol{U}(\beta-\hat{\beta})\|^2+\xi^T\boldsymbol{U}(\beta-\hat{\beta})+\lambda\|\mathcal{G}\hat{\beta}\|_* \leq \lambda\|\mathcal{G}\beta\|_*\\ \Rightarrow\quad &\lambda\|\mathcal{G}\hat{\beta}\|_* \leq \lambda\|\mathcal{G}\beta\|_*+\xi^T\boldsymbol{U}(\hat{\beta}-\beta)\\ \Rightarrow\quad &\|\mathcal{G}\hat{\beta}\|_*-\|\mathcal{G}\beta\|_* \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_*\end{aligned} \tag{11}$$

(11) is an important result to note, and following that,

$$\begin{aligned}&\|\mathcal{G}\hat{\beta}\|_*-\|\mathcal{G}\beta\|_* \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_*\\ \Rightarrow\quad &\langle\mathcal{G}(\hat{x}-x), V_1V_2^T+W\rangle \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}\|\mathcal{G}(\hat{x}-x)\|_*\\ \Rightarrow\quad &\|\mathcal{P}_W\mathcal{G}(\hat{x}-x)\|_* \leq -\langle\mathcal{G}(\hat{x}-x), V_1V_2^T\rangle + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}\|\mathcal{G}(\hat{x}-x)\|_*\\ \Rightarrow\quad &\|\mathcal{P}_W\mathcal{G}(\hat{x}-x)\|_* \leq \|\mathcal{P}_V\mathcal{G}(\hat{x}-x)\|_* + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}(\|\mathcal{P}_V\mathcal{G}(\hat{x}-x)\|_*+\|\mathcal{P}_W\mathcal{G}(\hat{x}-x)\|_*)\\ \Rightarrow\quad &\|\mathcal{P}_W\mathcal{G}(\hat{x}-x)\|_* \leq \frac{1+\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}}{1-\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2}{\lambda}}\|\mathcal{P}_V\mathcal{G}(\hat{x}-x)\|_*\end{aligned} \tag{12}$$

Let $\boldsymbol{U}$ be iid Gaussian matrix with scaling $\boldsymbol{E}(\boldsymbol{U}^T\boldsymbol{U}) = I$. Here we need to study the Gaussian width of the tangent cone $w(\mathcal{J}(\beta))$ of (7). Banerjee et al. (2014) proves that, if (11) is true, and $\lambda \geq 2\|\mathcal{G}(\boldsymbol{U}^T\xi)\|_2$, then the Gaussian width of this set (intersecting with unit ball) is less than 3 times of Gaussian width of $\{\hat{\beta} : \|\mathcal{G}(\hat{\beta})\|_* \leq \|\mathcal{G}(\beta)\|_*\}$, which is $O(\sqrt{R}\log n)$ Cai et al. (2016).
A simple bound is that, let $\delta = \hat{\beta}-\beta$, Γ can be replaced by

$$\max \|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\|_2/\|\mathcal{G}(\delta)\|_2$$

subject to $\hat{\beta} \in \mathcal{J}(\beta)$. With (12), we have $\|\mathcal{P}_W\mathcal{G}(\delta)\|_* \le 3\|\mathcal{P}_V\mathcal{G}(\delta)\|_*$.
Denote $\sigma = \|\mathcal{G}(\delta)\|_2$, we know that $\sigma \ge \max\{\|\mathcal{P}_W\mathcal{G}(\delta)\|_2, \|\mathcal{P}_V\mathcal{G}(\delta)\|_2\}$ and $\|\mathcal{P}_V\mathcal{G}(\delta)\|_2 \ge \|\mathcal{P}_V\mathcal{G}(\delta)\|_*/(2R)$. And simple algebra gives that

$$\max_{0<\sigma_i<\sigma, \sum_i \sigma = S} \sum_i \sigma_i^2 \le S\sigma.$$

So let σ_i be singular values of $\mathcal{P}_V\mathcal{G}(\delta)$ or $\mathcal{P}_W\mathcal{G}(\delta)$, and $S = \|\mathcal{P}_V\mathcal{G}(\delta)\|_*$ or $\|\mathcal{P}_W\mathcal{G}(\delta)\|_*$,

$$\frac{\sigma}{\|\mathcal{P}_V\mathcal{G}(\delta)\|_F} \ge \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}} \ge \sqrt{1/2R}$$
$$\frac{\sigma}{\|\mathcal{P}_W\mathcal{G}(\delta)\|_F} \ge \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_W\mathcal{G}(\delta)\|_*}} \ge \sqrt{1/6R}$$

the second last inequality comes from (12). Thus if $\|(I - \boldsymbol{U}^T\boldsymbol{U})\delta\| = O(1/\sqrt{R})\|\delta\|$, in other words, $\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F = O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F$, whenever δ in tangent cone, we have

$$\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_2 \le \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F \le O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \le \|\mathcal{G}(\delta)\|_2 \tag{13}$$

so $\Gamma < 1$. To get this, we need $\sqrt{T}/w(\mathcal{J}(\beta)) = O(\sqrt{R})$ where $T = O(pR^2\log^2 n)$ (Vershynin, 2018, Thm 9.1.1), still not tight in R, but $O(\min\{n, R^2\log^2 n\})$ is as good as Oymak and Ozay (2018) and better than Sarkar et al. (2019), which are $O(n)$ and $O(n^2)$ correspondingly. (Vershynin, 2018, Thm 9.1.1) is a bound in expectation, but it naively turns into high probability bound since $\Gamma \ge 0$. ■

6. Bounding Γ, where do we lose?

The previous proof is not tight here.

$$\underbrace{\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_2 \le \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F}_{\text{not tight}} \le O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \le \|\mathcal{G}(\delta)\|_2 \tag{14}$$

If we can show that, for all δ in the tangent cone (thus independent of $\boldsymbol{U}$), $\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_2 = O(1/\sqrt{R})\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F$ for $\boldsymbol{U} \in \mathbb{R}^{O(R\log^2 n)\times n}$, then we can get the correct sample complexity. The difficulty is that, we do not know the distribution of $(I - \boldsymbol{U}^T\boldsymbol{U})\delta$. Let $M = I - \boldsymbol{U}^T\boldsymbol{U}$ and $g := M\delta$. Let $\tilde{g}$ be a Gaussian vector with same mean and covariance as g that will be studied later. We know that $g_i = \sum M_{ij}\delta_j$. Let $z_{ij} = U_{:,i}^T U_{:,j}$,

u, v denote standard Gaussian vectors of dimension T, we have (the last equation: $i \neq j$)

$$\begin{aligned}
E((1-z_{ii}^2)^2) &= E((1-\frac{1}{T}u^Tu)^2) \\
&= 1 - \frac{2}{T}\sum_{i=1}^{T} E(u_i^2) + \frac{1}{T^2}(\sum_{i=1}^{T} E(u_i^4) + \sum_{i\neq j}^{T} E(u_i^2u_j^2)) = \frac{2}{T}. \\
E(z_{ij}^2) &= E((\frac{1}{T}u^Tv)^2) \\
&= \frac{1}{T^2}E(\sum u_i^2v_i^2) = \frac{1}{T}. \\
E(g_i) &= 0, \\
E(g_i^2) &= E((\sum M_{ij}\delta_j)^2) \\
&= \delta_i^2 E((1-z_{ii}^2)^2) + \sum_{j\neq i}\delta_j^2 E(z_{ij}^2) + \sum_{j\neq k}\delta_j\delta_k E(M_{ij}M_{ik}) \\
&\leq \frac{1}{T}(\delta_i^2 + \|\delta\|^2). \\
E(g_ig_j) &= E((\sum M_{ik}\delta_k)(\sum M_{jl}\delta_l)) \\
&= \delta_i\delta_j E(M_{ij}M_{ji}) \\
&= \frac{1}{T}\delta_i\delta_j.
\end{aligned}$$

So

$$Cov(g) = \frac{1}{T}(\|\delta\|^2 I + \delta\delta^T).$$

The problem is that g is not Gaussian so even we know mean and variance it's still hard to deal with. Let's study Gaussian first. If $\tilde{g} = \tilde{g}_1 + \check{g}_2\delta$ where $\tilde{g}_1 \sim \mathcal{N}(0, \frac{\|\delta\|^2}{T}I)$ and $\check{g}_2 \sim \mathcal{N}(0, 1/T)$, then we have

$$\begin{aligned}
E(\|\mathcal{G}(\tilde{g})\|_2) &\leq E(\|\mathcal{G}(\tilde{g}_1)\|_2) + E(|\check{g}_2|\|\mathcal{G}(\delta)\|_2) \\
&\leq \frac{1}{\sqrt{T}}(\|\delta\|\frac{\log n}{\sqrt{n}} + \|\mathcal{G}(\delta)\|_2) \\
&\leq \frac{1}{\sqrt{T}}(\underbrace{\frac{\sqrt{R}\log n}{\sqrt{n}}}_{\text{proven in paper}} + 1)\|\mathcal{G}(\delta)\|_2 \\
&\leq \frac{2}{\sqrt{T}}\|\mathcal{G}(\delta)\|_2.
\end{aligned}$$

If we have

$$P(\|\mathcal{G}(\tilde{g})\|_2 > \alpha E(\|\mathcal{G}(\tilde{g})\|_2)) \leq \psi(\alpha),$$

then let $\alpha = \sqrt{T}/2$, we have

$$P(\|\mathcal{G}(\tilde{g})\|_2 > E(\|\mathcal{G}(\delta)\|_2)) \leq \psi(\sqrt{T}/2)$$

We hope that $\psi(\alpha) = \exp(-O(\alpha^2))$ or $\log(\psi(\alpha)) = -O(\alpha^2)$. Then with a set of Gaussian width $\sqrt{R}\log n$, we use a union bound and have (if we ignore the difference between g and $\tilde{g}$)

$$P(\max_{\delta} \|\mathcal{G}(g)\|_2 > \|\mathcal{G}(\delta)\|_2) \leq \psi(\sqrt{T}/2)\exp(O(R\log^2 n)) = \exp(O(R\log^2 n) + \log(\psi(\sqrt{T}/2))).$$

So if the derivation of a Gaussian vector can be applied to a non-Gaussian $g = (I - \boldsymbol{U}^T\boldsymbol{U})\delta$ with the same mean and variance, and $\|\mathcal{G}(g)\|_2$ is a subGaussian random variable, then we can get the tight bound.

References

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564, 2014.

Jian-Feng Cai, Xiaobo Qu, Weiyu Xu, and Gui-Bo Ye. Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and computational harmonic analysis*, 41(2):470–490, 2016.

Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11): 1877–1904, 2014.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.