Intersecting singularities for multi-structured estimation

Emile Richard EMILE.RICHARD@MINES-PARISTECH.FR
CBIO Mines ParisTech, INSERM U900, Institut Curie

Francis Bach FRANCIS.BACH@INRIA.FR
SIERRA project-team, INRIA - Département d'Informatique de l'Ecole Normale Supérieure, Paris

Jean-Philippe Vert JEAN-PHILIPPE.VERT@MINES-PARISTECH.FR
CBIO Mines ParisTech, INSERM U900, Institut Curie

1. Theoretical guarantees for lifted trace norm regularized estimation

We reformulated the "trace + 1" penalty using a linear mapping Π and introduced a new penalty, the block norm, using Φ. Using the general formalism of *lifted trace norms* we state theoretical results that help us better understand the behaviour of each of the two norms and compare them more easily. Due to space constraints, all proofs are postponed to appendices available as supplementary materials.

1.1. Lifted trace norms

We call *lifting* a linear mapping $\Lambda : \mathbb{R}^{n\times m} \to \mathbb{R}^{n'\times m'}$ and call the penalty induced by $\|\Lambda(X)\|_*$ on the matrix X the Λ-trace or *lifted trace norm.* Such penalties have been used in compressed sensing (Hosseini Kamal & Vandergheynst, 2013), in statistics (Grave et al., 2011), and have similarities with fused sparsity inducing type of penalties $\|\Lambda(X)\|_1$ studied for instance by Dalalyan & Chen (2012); Vert & Bleakley (2010); Vaiter et al. (2012). Note that a lifted trace norm is not necessarily a norm. It verifies triangle inequality and positive homogeneity, but only separates points so becomes a norm if Λ is injective (*i.e.*, $\Lambda(X) = 0 \Rightarrow X = 0$). We denote by $\|\Lambda\| = \max_{\|X\|_F\leq 1} \|\Lambda(X)\|_F$ the operator norm of the linear map Λ. The mapping Λ^* denotes the adjoint operator of Λ. If $\Lambda(X) = U_{\Lambda(X)}\Sigma_{\Lambda(X)}V_{\Lambda(X)}^\top$ is the singular value decomposition of $\Lambda(X)$, the subgradient of the Λ-trace at X is given by

$$\partial\|\Lambda(X)\|_* = \left\{ \Lambda^*\left(U_{\Lambda(X)}V_{\Lambda(X)}^\top + \mathcal{P}_{\Lambda(X)}^\perp(Z)\right) \quad \text{where} \quad Z\in\mathbb{R}^{N\times M} \quad \text{and} \quad \|Z\|_{op}\leq 1 \right\} .$$

From this expression one can see that when $\Lambda(X)$ is rank deficient then $\|\Lambda(X)\|_*$ is nondifferentiable, in cases where the image of Λ^* is the whole space $\mathbb{R}^{n\times m}$. This makes the rank of $\Lambda(X)$ a particularly interesting quantity in this context.

In the following $X^\star$ denotes the target matrix to be estimated and $\omega : \mathbb{R}^{n\times m} \to \mathbb{R}^d$ a set of linear measurements:

$$\omega(X) = \Big(\langle\Omega_1, X\rangle, \cdots, \langle\Omega_d, X\rangle\Big)^\top \quad .$$

We call the Ω_is design matrices and we will be interested in the estimation procedures (i) minimizing the least squares loss $\ell(X) = \frac{1}{d}\|\omega(X)-y\|_2^2$ penalized with lifted trace norm and (ii) minimizing the Λ-trace subject to $\omega(X) = \omega(X^\star)$.

1.2. Least squares regression with lifted trace-norm penalty

We consider linear regression and prove oracle inequalities for the estimation procedure using techniques introduced by Koltchinskii et al. (2011). That is, we consider the model

$$y = \omega(X^\star) + \epsilon \in \mathbb{R}^d$$

where $\epsilon\in\mathbb{R}^d$ having i.i.d zero mean entries.

Assumption 1 *We assume that the lifting Λ is orthogonal, that is $\Lambda^*\Lambda = \|\Lambda\|^2 Id$, which is for instance the case of Φ and Π.*

For the two orthogonal liftings of interest Π and Φ, the operator norms respectively are given by $\|\Pi\|^2 = (1-\beta)^2+\beta^2$ and $\|\Phi\|^2 = (n+m)(1-\beta)^2+\beta^2$.

Definition 1 *The* cone of restriction $\mathcal{C}(X, \kappa, \Lambda)$ *is the* *set of matrices* $B \in \mathbb{R}^{n\times m}$ *satisfying*

$$\|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(B))\|_* \leq \kappa \|\mathcal{P}_{\Lambda(X)}(\Lambda(B))\|_* \quad . \qquad (1)$$

The restricted eigenvalue *of* ω *at* X *is*

$$\mu_{\kappa,\Lambda}(X) = \inf\Bigg\{\mu > 0 \quad \textit{such that}$$
$$\|\mathcal{P}_{\Lambda(X)}(\Lambda(B))\|_F \leq \frac{\mu}{\sqrt{d}}\|\omega(B)\|_2 \quad , \quad \forall B \in \mathcal{C}(X,\kappa,\Lambda)\Bigg\} \quad .$$

Define the objective

$$\mathcal{L}(X) = \frac{1}{d}\|\omega(X) - y\|_2^2 + \lambda\|\Lambda(X)\|_* \quad , \qquad (2)$$

and consider the following estimation procedure

$$\widehat{X} = \underset{X\in\mathcal{S}}{\arg\min}\,\mathcal{L}(X) \quad , \qquad (3)$$

where $\mathcal{S} \subset \mathbb{R}^{n\times m}$ is the convex cone of admissible solutions. We can state the following oracle inequality on the estimate $\hat{X}$.

Proposition 1 *Under Assumption 1, for* $\lambda \geq \frac{3}{d}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$*, where* $M = \sum_{i=1}^d \epsilon_i\Omega_i$*, the following holds:*

$$\|\omega(\widehat{X} - X^\star)\|_2^2 \leq$$
$$\inf_{X\in\mathcal{S}}\left\{\|\omega(X - X^\star)\|_2^2 + \lambda^2\mu_{5,\Lambda}(X)^2\,\mathrm{rank}(\Lambda(X))\right\} \quad .$$

Note that as (see the proof) $\widehat{X} - X^\star \in \mathcal{C}(X^\star, 5, \Lambda)$ and by orthogonality of Λ, we bound the estimation error by the prediction error $\|\widehat{X} - X^\star\|_F^2 \leq \frac{36\mu_{5,\Lambda}(X)^2\,\mathrm{rank}(\Lambda(X^\star))}{\|\Lambda\|^2 d}\|\omega(\widehat{X} - X^\star)\|_2^2$ and hence the oracle inequality of Proposition 1 provides a abound on the estimation error.

We point out that using similar techniques, and under the stronger assumption called *Restricted Isometry Property* that assumes there exists $\mu > 0$ such that for any $X_1, X_2 \in \mathcal{S}$

$$\frac{1}{d}\|\omega(X_1 - X_2)\|_2^2 \geq \mu^{-2}\|X_1 - X_2\|_F^2 \ ,$$

one can state that for $\lambda \geq \frac{2}{d}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$, we have

$$\mu^{-2}\|\widehat{X} - X^\star\|_F^2 \leq \|\omega(\widehat{X} - X^\star)\|_2^2 \leq$$
$$\inf_{X\in\mathcal{S}}\left\{\|\omega(X - X^\star)\|_2^2 + \mu^2 c_0^2\lambda^2\,\mathrm{rank}(\Lambda(X))\right\}$$

where $c_0 = \frac{\sqrt{2}+1}{2}$ and the first inequality being true if $X^\star \in \mathcal{S}$. In particular in the case of denoising $\omega = id, y = X^\star + M$ considered for instance by Chandrasekaran & Jordan (2012), this proves that if $\lambda \geq \frac{2}{nm}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$

$$\frac{1}{\sqrt{nm}}\|\widehat{X} - X^\star\|_F \leq c_0\lambda\sqrt{\mathrm{rank}(\Lambda(X^\star))} \ .$$

1.3. Probabilistic results

The theoretical analysis of penalized estimation procedures by a norm highlights that when the dual norm of the noise is low the result is more attractive. This motivates us to understand the behavior of $\|\Lambda(G)\|_{op}$ where G denotes the noise which we assume to Gaussian in this work. To this end let us first define the variance of a lifting using canonical matrices $E_{i,j}$ having 1 at the (i,j) entry and 0 everywhere else as

$$v_\Lambda^2 = \|\sum_{i,j}\Lambda(E_{i,j})\Lambda(E_{i,j})^\top\|_{op} \vee \|\sum_{i,j}\Lambda(E_{i,j})^\top\Lambda(E_{i,j})\|_{op} \ .$$

Using results stated in (Tropp, 2010), we know that for a matrix G having i.i.d. centered normal entries

$$\mathbb{E}\left[\|\Lambda(G)\|_{op}\right] \leq \sqrt{2v_\Lambda^2\log(N+M)} \ ,$$

and we can control the deviation for $t > 0$ as

$$\mathbb{P}\left[\|\Lambda(G)\|_{op} \geq \sqrt{2v_\Lambda^2(\log(N+M)+t)}\right] \leq e^{-t} \ .$$

We can bound the Πs variance $v_\Pi^2(\beta) \leq \beta^2 \vee n(1-\beta)^2$ and observe that by setting $\beta = \frac{\sqrt{n}}{1+\sqrt{n}}$ we get the upper bound on the expectation over standard normal matrices G

$$\mathbb{E}\|\Pi(G)\|_{op} \leq \sqrt{\frac{2n}{(1+\sqrt{n})^2}\log(n+m+2nm)} \ .$$

The variance of Φ can be controlled by $v_\Phi^2(\beta) \leq (1+n)(1-\beta)^2 + \beta^2$, which suggests to set $\beta = \frac{n+1}{n+2}$ in order to obtain

$$\mathbb{E}\|\Phi(G)\|_{op} \leq 2\sqrt{\frac{n+1}{n+2}\log(n+m)} \ .$$

We also define the observable variance under the linear map ω as

$$v_{\omega,\Lambda}^2 = \frac{1}{d}\left\|\sum_{i=1}^d\Lambda(\Omega_i)\Lambda(\Omega_i)^\top\right\|_{op} \vee \left\|\sum_{i=1}^d\Lambda(\Omega_i)^\top\Lambda(\Omega_i)\right\|_{op} ,$$

which is a function of β for Π and Φ and equal to $\frac{1}{nm}v_\Lambda^2$ in case of denoising $\omega = id$. We finally assume the noise vector elements ϵ_i are independently drawn from $\mathcal{N}(0,\sigma^2)$.

Corollary 1 (Block norm) *Consider the* Φ*-trace penalty and calibrate for* $t > 0$

$$\lambda = \frac{6\sigma v_{\Phi,\omega}}{\beta^2 + (n+m)(1-\beta)^2}\sqrt{\frac{\log(n+m)+t}{d}} \ ,$$

then with probability at least $1 - e^{-t}$*,*

$$\|\omega(\widehat{X} - X^\star)\|_2^2 \leq \inf_{X \in \mathcal{S}} \left\{ \|\omega(X - X^\star)\|_2^2 \right. \\ \left. + c^2 \frac{\log(n+m)+t}{d} \mathbf{ranksity}(X) \right\} \ ,$$

where $c = \frac{6\sigma\sigma v_{\Phi,\omega}\mu_{5,\Phi}(X)}{\beta^2+(n+m)(1-\beta)^2}$ *depends on* β*.*

Corollary 2 (Trace + 1) *Consider the* Π*-trace penalty and calibrate for* $t > 0$

$$\lambda = \frac{3\sigma v_{\Pi,\omega}}{\beta^2 + (1-\beta)^2}\sqrt{\frac{2\log(n+m+2nm)+t}{d}} \ ,$$

then with probability at least $1 - e^{-t}$*,*

$$\|\omega(\widehat{X} - X^\star)\|_2^2 \leq \inf_{X \in \mathcal{S}} \left\{ \|\omega(X - X^\star)\|_2^2 \right. \\ \left. + c^2 \frac{\log(n+m+2nm)+t}{d} \bigg(\mathrm{rank}(X) + \|X\|_0 \bigg) \right\} \ ,$$

where $c = \frac{3\sqrt{2}\sigma v_{\Pi,\omega}\mu_{5,\Pi}(X)}{\beta^2+(1-\beta)^2}$*.*

In both cases it is the minimizer of respectively $\beta \mapsto \frac{v_{\Phi,\omega}(\beta)}{\beta^2+(n+m)(1-\beta)^2}$ and $\frac{v_{\Pi,\omega}(\beta)}{\beta^2+(1-\beta)^2}$ that calibrates β. The two corollaries are interesting because they show that after a natural calibration of the tuning parameter λ, the convex estimation procedure (3) outputs the optimal estimators for the nonconvex penalties rank $+\ell_0$ and **ranksity**, respectively. In addition the multiplicative factor behind these estimators sharply reminds us of known optimal rates, such as $(\log n)/p$ for the Lasso.

1.4. Compressed sensing and exact recovery

Consider the constrained convex optimization problem

$$\min_X \|\Lambda(X)\|_* \ \text{ s.t. } \ \omega(X) = \omega(X^\star) \ , \qquad (4)$$

where the design matrices Ω_i are i.i.d. Gaussians. We have the following bound on the minimum required such observations for perfect recovery of $X^\star$.

Proposition 2 *The minimum required number of Gaussian i.i.d. observations for achieving perfect recovery of* $X^\star$ *with overwhelming probability by solving (4) where* Λ *is an orthogonal lifting is at most*

$$d_\Lambda = \mathbb{E}\left[\|\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\|^2_{op}\right] \mathrm{rank}(\Lambda(X^\star)) + 1 \ ,$$

the expectation being taken over the set of i.i.d. standard normal matrices G*.*

In the case of the orthogonal lifting Φ, the quantity $\|\mathcal{P}^\perp_{\Phi(X^\star)}(\Phi(G))\|_{op}$ can be naively bounded by $\|\Phi(G)\|^2_{op}$ for which we already have an upper bound.

Corollary 3 (Block norm) *For the* Φ*-trace penalty, by taking* $\beta = (n+1)/(n+2)$*,* $d_\Phi \leq 1 + 4$ **ranksity**$(X^\star)$ $\log(n+m)$ *i.i.d. Gaussian observations are enough to achieve with overwhelming probability perfect recovery of* $X^\star$ *by solving (4).*

For Π the situation is simpler as we have a better understanding of the behavior of $\mathcal{P}^\perp_{\Pi(X^\star)}(\Pi(G))$. In fact

$$\|\mathcal{P}^\perp_{\Pi(X^\star)}(\Pi(G))\|_{op} = \\ \left\| \begin{pmatrix} (1-\beta)\mathcal{P}^\perp_{X^\star}(G) & 0 \\ 0 & \beta \ \mathrm{Diag}(\mathbf{vec}(\mathcal{Q}^\perp_{X^\star}(G))) \end{pmatrix} \right\|_{op} \ .$$

allows us to analyze the terms separately and state

Corollary 4 (Trace + 1) *In the case of* Π*-trace penalty, take* $\beta = 1 - \frac{1}{\sqrt{n+m-2r}}$*, and assume* $r < m-2$*, we have*

$$d_\Pi \leq 1 + c_1(r+s)\log\left(c_2 + \frac{nm-s}{2} \ \right)$$

where $c_1 = \frac{8}{3}$ *and* $c_2 = 1 + e^{\frac{3}{4\beta^2}} \leq 2.3$*.*

On a bi-clique of size (k,l) we get $d_\Pi \leq c_1 kl \log(nm-s)$ and $d_\Phi \leq 4\{(n+m-1)+(k-1)(l-1)\}\log(n+m)$.

References

Chandrasekaran, V. and Jordan, M. I. Computational and statistical tradeoffs via convex relaxation. *Preprint*, 2012.

Chandrasekaran, V., Recht, B., Parrilo, P.A., and Willsky, A.S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12, 2012.

Dalalyan, A. S. and Chen, Y. Fused sparsity and robust estimation for linear models with unknown variance. In *NIPS 2012*, 2012.

Grave, E., Obozinski, G., and Bach, F. Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems 24*, pp. 2187–2195, 2011.

Hosseini Kamal, M. and Vandergheynst, P. Joint low-rank and sparse light field modeling for dense multiview data compression. *ICASSP*, 2013.

Koltchinskii, V., Lounici, K., and Tsybakov, A. Nuclear norm penalization and optimal rates for noisy matrix completion. *Annals of Statistics*, 2011.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *ArXiv e-prints*, April 2010.

Vaiter, S., Peyré, G., Dossal, C., and Fadili, J. Robust sparse analysis regularization. *to appear in IEEE Transactions on Information Theory*, 2012.

Vert, J.-P and Bleakley, K. Fast detection of multiple change-points shared by many signals using group lars. *Advances in Neural Information Processing Systems 23 (NIPS)*, pp. 2343–2351, 2010.

Proof of Proposition 1. Pick $X \in \mathcal{S}$, in the convex cone of admissible solutions. Let $\mathcal{P}_{\Lambda(X)}$ denote the projector onto $\mathbf{span}(\Lambda(X))$. We start by setting some technical lemmas.

Lemma 1 *For all $M \in \mathbb{R}^{n\times n}$, we have*

$$\begin{aligned}&\langle M, \hat{X} - X\rangle \leq \\ &\|\mathcal{P}_{\Lambda(X)}(\Lambda(M))\|_* \|\mathcal{P}_{\Lambda(X)}(\Lambda(\hat{X} - X))\|_{op}/\|\Lambda\|^2 \\ &+ \|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(M))\|_{op} \|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\hat{X} - X))\|_*/\|\Lambda\|^2\end{aligned}$$

and

$$\begin{aligned}&\langle M, \hat{X} - X\rangle \leq && (5)\\ &\sqrt{2 \ \operatorname{rank}(\Lambda(M))}\|\mathcal{P}_{\Lambda(X)}(\Lambda(M))\|_{op} && (6)\\ &\|\mathcal{P}_{\Lambda(X)}(\Lambda(\hat{X} - X))\|_F/\|\Lambda\|^2 && (7)\\ &+ \|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(M))\|_{op} && (8)\\ &\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\hat{X} - X))\|_*/\|\Lambda\|^2 && (9)\end{aligned}$$

Lemma 2 *There exists $Z \in \partial\|\Lambda(X)\|_*$ such that*

$$\begin{aligned}&-\langle Z, \widehat{X} - X\rangle \leq \\ &\qquad \sqrt{\operatorname{rank}(\Lambda(X))}\|\widehat{X} - X\|_F\|\Lambda\| - \|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\end{aligned}$$

and

$$-\langle Z, \widehat{X} - X\rangle \leq \|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X} - X))\|_* - \|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_* \ . \tag{10}$$

Lemma 3 *Let $M = \sum_{i=1}^d \epsilon_i\Omega_i$, we have*

$$\nabla\|\omega(\widehat{X}) - y\|_2^2 = 2\langle\omega(\widehat{X} - X^\star), \omega(\widehat{X} - X)\rangle - 2\langle M, \widehat{X} - X\rangle \ . \tag{11}$$

By optimality, an element of the subgradient of $\mathcal{L}$ at $\widehat{X}$ belongs to the normal cone of $\mathcal{S}$ at $\widehat{X}$. We have $\langle\partial\mathcal{L}(\widehat{X}), \widehat{X} - X\rangle \leq 0$. On the other hand, by the monotonicity of the subgradient of the convex function $\|\Lambda(\cdot)\|_*$ we have $\langle\widehat{X} - X, \widehat{Z} - Z\rangle \geq 0$. Therefore we can deduce by using Lemma 3, that for $M = \sum_{i=1}^d \epsilon_i\Omega_i$,

$$\begin{aligned}&\langle\partial\mathcal{L}(\widehat{X}), \widehat{X} - X\rangle - \lambda\langle\widehat{Z} - Z, \widehat{X} - X\rangle \leq 0 && (12)\\ \Leftrightarrow\ &\langle\frac{1}{d}\nabla\|\omega(\widehat{X}) - y\|_2^2 + \lambda Z, \widehat{X} - X\rangle \leq 0 && (13)\\ \Leftrightarrow\ &\frac{2}{d}\langle\omega(\widehat{X} - X^\star), \omega(\widehat{X} - X)\rangle \leq && (14)\\ &\frac{2}{d}\langle M, \widehat{X} - X\rangle - \lambda\langle Z, \widehat{X} - X\rangle \ . && (15)\end{aligned}$$

We recall the identity

$$\begin{aligned}&2\langle\omega(\widehat{X} - X^\star), \omega(\widehat{X} - X)\rangle = \\ &\|\omega(\widehat{X} - X^\star)\|_2^2 + \|\omega(\widehat{X} - X)\|_2^2 - \|\omega(X - X^\star)\|_2^2 \ .\end{aligned}$$

It shows that if $\langle\omega(\widehat{X} - X^\star), \omega(\widehat{X} - X)\rangle \leq 0$, then the bound trivially holds. So lets assume $\langle\omega(\widehat{X} - X^\star), \omega(\widehat{X} - X)\rangle > 0$.

In this case the bound (10) in Lemma 2 and equation (15) imply

$$\lambda\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_* \leq \frac{2}{d}\langle M, \widehat{X} - X\rangle + \lambda\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X} - X))\|_* \ . \tag{16}$$

By using Lemma 1 , first inequality (5), we have

$$\begin{aligned}&(\lambda - \frac{2}{d}\frac{\|\Lambda(M)\|_{op}}{\|\Lambda\|^2})\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X} - X))\|_* \\ &\leq (\lambda + \frac{2}{d}\frac{\|\Lambda(M)\|_{op}}{\|\Lambda\|^2})\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X} - X))\|_* \ .\end{aligned}$$

This shows that for $\lambda \geq \frac{3}{d}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$, by using the fact that for $x \geq 3$, $\frac{x-2}{x+2} \geq \frac{1}{5}$, the following holds true

$$\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X} - X))\|_* \leq 5\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X} - X))\|_* \ .$$

As a consequence, $\widehat{X} - X \in \mathcal{C}(X, 5, \Lambda)$. On the other hand, by using Lemma 1, second inequality (9) and (15) we have

$$
\begin{aligned}
&\frac{1}{d}\left(\|\omega(\widehat{X}-X^\star)\|_2^2+\|\omega(\widehat{X}-X)\|_2^2-\|\omega(X-X^\star)\|_2^2\right)\\
&\leq \frac{2}{d}\left(\sqrt{2\,\mathrm{rank}(\Lambda(X))}\frac{\|\Lambda(M)\|_{op}}{\|\Lambda\|^2}\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_F\right.\\
&\qquad\left.+\ \frac{\|\Lambda(M)\|_{op}}{\|\Lambda\|^2}\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\right)\\
&+\lambda\sqrt{\mathrm{rank}(\Lambda(X))}\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_F-\lambda\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\ .
\end{aligned}
\tag{17}
$$

By using the definition of the restricted eigenvalue $\mu(X)=\mu_{5,\Lambda}(X)$, given that $\widehat{X}-X\in\mathcal{C}(X,5,\Lambda)$,

$$
\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_F\leq\frac{\mu(X)}{\sqrt{d}}\|\omega(\hat{X}-X)\|_2
$$

so we can write, again thanks to $\lambda\geq\frac{3}{d}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$,

$$
\begin{aligned}
&\frac{1}{d}\left(\|\omega(\widehat{X}-X^\star)\|_2^2+\|\omega(\widehat{X}-X)\|_2^2-\|\omega(X-X^\star)\|_2^2\right)\\
&\leq\frac{\mu(X)}{\sqrt{d}}\lambda\sqrt{\mathrm{rank}(\Lambda(X))}\left(1+\frac{2\sqrt{2}}{3}\right)\|\omega(\widehat{X}-X)\|_F\quad .
\end{aligned}
$$

So by $bx-x^2\leq\left(\frac{b}{2}\right)^2$ we finally get

$$
\begin{aligned}
&\frac{1}{d}\|\omega(\widehat{X}-X^\star)\|_2^2\leq\\
&\frac{1}{d}\|\omega(X-X^\star)\|_2^2+\lambda^2\frac{\mu(X)^2}{4d}(1+\frac{2\sqrt{2}}{3})^2\,\mathrm{rank}(\Lambda(X))\ .\ \square
\end{aligned}
$$

Proof of Lemma 1 Let us decompose $\Lambda(M)$ onto the direct sum formed by the span of $\Lambda(X)$ and the orthogonal space:

$$
\Lambda(M)=\mathcal{P}_{\Lambda(X)}(M)+\mathcal{P}^\perp_{\Lambda(X)}(M)\quad .
$$

By using assumption 1 and Holder's inequality twice

$$
\begin{aligned}
&\langle M,\hat{X}-X\rangle=\langle\Lambda(M),\Lambda(\hat{X}-X)\rangle/\|\Lambda\|^2\leq\\
&\|\mathcal{P}_{\Lambda(X)}(\Lambda(M))\|_*\|\mathcal{P}_{\Lambda(X)}(\Lambda(\hat{X}-X))\|_{op}/\|\Lambda\|^2\\
&\quad+\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(M))\|_{op}\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\hat{X}))\|_*/\|\Lambda\|^2
\end{aligned}
$$

The other bound is obtained in a similar fashion by using Cauchy-Schwarz on the first term and also the fact that $\|\mathcal{P}_{\Lambda(X)}(M)\|_F\leq\sqrt{2\ \mathrm{rank}(\Lambda(X))}\|M\|_F$ since we can write $\mathcal{P}_{\Lambda(X)}(M)=(I-UU^\top)MVV^\top+UU^\top M$ for U and V singular vectors of $\Lambda(X)$.

Proof of Lemma 2. Let

$$
Z=\Lambda^*\left(U_{\Lambda(X)}V^\top_{\Lambda(X)}+\mathcal{P}^\perp_{\Lambda(X)}(W)\right)
$$

denote an element of the subgradient of $\|\Lambda(\cdot)\|_*$, where $\|W\|_{op}\leq 1$. Take $W=-UV^\top$ where $U\Sigma V^\top=\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\hat{X}))$ is a singular value decomposition, then $\|W\|_{op}=1$ and

$$
\begin{aligned}
&\langle\Lambda^*(\mathcal{P}^\perp_{\Lambda(X)}(W)),\hat{X}-X\rangle=\\
&\langle\mathcal{P}^\perp_{\Lambda(X)}(W),\Lambda(\hat{X}-X)\rangle=\\
&-\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*
\end{aligned}
$$

so we can write

$$
\begin{aligned}
&-\langle Z,\widehat{X}-X\rangle=\\
&-\langle\Lambda^*\left(U_{\Lambda(X)}V^\top_{\Lambda(X)}\right),\hat{X}-X\rangle\\
&\qquad+\langle\Lambda^*(\mathcal{P}^\perp_{\Lambda(X)}(W)),\hat{X}-X\rangle=\\
&-\langle U_{\Lambda(X)}V^\top_{\Lambda(X)},\Lambda(\hat{X}-X)\rangle\\
&\qquad+\langle\mathcal{P}^\perp_{\Lambda(X)}(W),\Lambda(\hat{X}-X)\rangle=\\
&-\langle U_{\Lambda(X)}V^\top_{\Lambda(X)},\Lambda(\widehat{X}-X)\rangle\\
&\qquad-\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\ .
\end{aligned}
$$

We know that $\|U_{\Lambda(X)}V^\top_{\Lambda(X)}\|_F^2\leq\mathrm{rank}(\Lambda(X))$. By Cauchy-Schwarz

$$
-\langle Z,\widehat{X}-X\rangle\leq\sqrt{\mathrm{rank}(\Lambda(X))}\|\widehat{X}-X\|_F\|\Lambda\|-\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\ .
$$

Similarly if we use Holder's instead of Cauchy-Schwarz, and thanks to $\|U_{\Lambda(X)}V^\top_{\Lambda(X)}\|_{op}=1$,

$$
\begin{aligned}
&-\langle Z,\widehat{X}-X\rangle\leq\\
&\qquad\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_*-\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\ .\ \square
\end{aligned}
$$

Proof of Lemma 3.

Given that $\nabla\|\omega(\widehat{X})-y\|_2^2=2\sum_{i=1}^d\Omega_i\langle\Omega_i,\widehat{X}\rangle-y_i\Omega_i$, we obtain

$$
\begin{aligned}
&\langle\nabla\|\omega(\widehat{X})-y\|_2^2,\widehat{X}-X\rangle\\
&\quad=2\sum_{i=1}^d\langle(\langle\Omega_i,\widehat{X}\rangle-y_i)\Omega_i,\widehat{X}-X\rangle\\
&\quad=2\sum_{i=1}^d(\langle\Omega_i,\widehat{X}\rangle-y_i)\langle\Omega_i,\widehat{X}-X\rangle\\
&\quad=2\langle\omega(\widehat{X})-y,\omega(\widehat{X}-X)\rangle\\
&\quad=2\langle\omega(\widehat{X}-X^\star)+\omega(X^\star)-y,\omega(\widehat{X}-X)\rangle\\
&\quad=2\langle\omega(\widehat{X}-X^\star),\omega(\widehat{X}-X)\rangle-2\langle\epsilon,\omega(\widehat{X}-X)\rangle\\
&\quad=2\langle\omega(\widehat{X}-X^\star),\omega(\widehat{X}-X)\rangle-2\langle M,\widehat{X}-X\rangle\ .\ \square
\end{aligned}
$$

Proof of Proposition 2. By orthogonality of Λ we have

$$\|\Lambda\|^2 G = \Lambda^*\Lambda(G) = \Lambda^*\left(\mathcal{P}_{\Lambda(X^\star)}(\Lambda(G)) + \mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\right)$$

Lets build an appropriate element of the normal cone of the Λ-trace at $X^\star$

$$Z(G) = \frac{1}{\|\Lambda\|^2}\Lambda^*(\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))) + \frac{\|\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\|_{op}}{\|\Lambda\|^2}\Lambda^*\left(U_{\Lambda(X^\star)}V^\perp_{\Lambda(X^\star)}\right) \quad ,$$

and get by Cauchy-Schwarz inequality

$$\begin{aligned}\|Z(G) - G\|_F^2 &= \frac{\|\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\|^2_{op}}{\|\Lambda\|^2}\|\Lambda^* U_{\Lambda(X^\star)}V^\perp_{\Lambda(X^\star)}\|_F^2 \\ &\le \|\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\|^2_{op}\,\mathrm{rank}(\Lambda(X^\star)) \quad .\end{aligned}$$

By Lemma 2.7 in (Chandrasekaran et al., 2012) this bounds the squared gaussian width of the tangent cone to $\|\Lambda(\cdot)\|_*$ at $X^\star$ intersected with the unit sphere. We conclude by using Corollary 3.3 from the same paper. □

Proof of Corollary 4

Let $s = \|X^\star\|_0$ and $r = \mathrm{rank}(X^\star)$. First lets show that for any $G \in \mathbb{R}^{n\times m}$

$$\|\mathcal{P}^\perp_{\Pi(X^\star)}(\Pi(G))\|_{op} = \left\|\begin{pmatrix}(1-\beta)\mathcal{P}^\perp_{X^\star}(G) & 0 \\ 0 & \beta\ \mathrm{Diag}(\mathbf{vec}(\mathcal{Q}^\perp_{X^\star}(G)))\end{pmatrix}\right\|_{op} \quad .$$

In fact as the singular value decomposition of $\Pi(X^\star)$ can be written (up to permutations of rows and columns) using the matrices

$$U_{\Pi(X^\star)} = \begin{pmatrix}U_{X^\star} & 0 \\ 0 & \mathrm{Diag}(\mathbf{vec}(\mathrm{sgn}(X^\star)))\end{pmatrix}$$

and

$$V_{\Pi(X^\star)} = \begin{pmatrix}V_{X^\star} & 0 \\ 0 & \mathrm{Diag}(\mathbf{vec}(|\,\mathrm{sgn}(X^\star)|))\end{pmatrix}$$

the formula $\mathcal{P}^\perp(Z) = (I - UU^\top)Z(I - VV^\top)$ implies the result. Since the gaussian distribution is isotropic we know that $\|\mathcal{P}^\perp_{\Pi(X^\star)}(G)\|_{op}$ is distributed as the operator norm of a $(n-r)\times(m-r)$ gaussian matrix and that $\|\mathcal{Q}^\perp_{\Pi(X^\star)}(G)\|_\infty$ is distributed as the ℓ_∞ norm of a vector of length $nm - s$ having iid standard normal entries.

Let $J = \mathcal{Q}^\perp_{X^\star}(G)$ and $H = \mathcal{P}^\perp_{X^\star}(G)$ and

$$z = \max\left\{(1-\beta)^2\|H\|^2_{op}\ ,\ \beta^2\|J\|^2_\infty\right\} \quad ,$$

and notice that by Jensen inequality, for all $t > 0$

$$\begin{aligned}&\exp\left(t\ \mathbb{E}[z]\right) \le \mathbb{E}\exp(tz) \\ &\le \mathbb{E}\exp(t(1-\beta)^2\|H\|^2_{op}) + \sum_{i=1}^{nm-s}\mathbb{E}\exp(t\beta^2 J_i) \\ &= \mathbb{E}\exp(t(1-\beta)^2\|H\|^2_{op}) + \frac{nm-s}{\sqrt{1-2t\beta^2}} \quad ,\end{aligned}$$

where J_is are iid χ^2 variables and the last relation being the moment generating function of χ^2. For bounding the term $\mathbb{E}\exp(t(1-\beta)^2\|H\|^2_{op})$, let us recall

$$\mathbb{P}[\|H\|_{op} > \sqrt{n-r} + \sqrt{m-r} + s] \le \exp(-s^2/2)$$

and introduce $f(x) = \exp(t(1-\beta)^2x^2)$. We have $f^{-1}(z) = \frac{1}{1-\beta}\sqrt{\frac{\log(z)}{t}}$ strictly increasing $[1;\infty) \to \mathbb{R}$. Denoting $R = \sqrt{n-r} + \sqrt{m-r}$ we have the sequence of inequalities

$$\mathbb{E}\exp(t(1-\beta)^2\|H\|^2_{op}) \tag{18}$$
$$= \mathbb{E}f(\|H\|_{op}) \tag{19}$$
$$= \int_1^\infty \mathbb{P}[f(\|H\|_{op}) > h]\ \ dh \tag{20}$$
$$\le \int_1^{1+f(R)} 1\ dh \tag{21}$$
$$+ \int_{1+f(R)}^\infty \mathbb{P}[f(\|H\|_{op}) > h]dh \tag{22}$$
$$= f(R) \tag{23}$$
$$+ \int_0^\infty \mathbb{P}[\|H\|_{op} > f^{-1}(f(R) + 1 + \zeta)]d\zeta \tag{24}$$
$$\le f(R) \tag{25}$$
$$+ \int_0^\infty \mathbb{P}[\|H\|_{op} > R + f^{-1}(1+\zeta)]d\zeta \tag{26}$$
$$\le f(R) \tag{27}$$
$$+ \int_0^\infty 2ts(1-\beta)^2\ \exp\left(-s^2/2 + ts^2(1-\beta)^2\right)ds \tag{28}$$
$$\le f(R) + 1 \tag{29}$$

where (26) is due to the sublinearity of $f^{-1}(z) = \frac{1}{(1-\beta)}\sqrt{\frac{\log(z)}{t}}$:

$$f^{-1}(z+z') \leq f^{-1}(z) + f^{-1}(z')$$

and (29) is true for any $t < \frac{1}{2(1-\beta)^2}$. We have for $t < \frac{1}{2}\min\left(\frac{1}{(1-\beta)^2}, \frac{1}{\beta^2}\right)$,

$$\mathbb{E}[z] \leq \\ \frac{1}{t}\log\left\{1 + \exp[2t(1-\beta)^2(n+m-2r)] + \frac{nm-s}{\sqrt{1-2t\beta^2}}\right\}$$

By taking $t = \frac{3}{8\beta^2}$ and $(1-\beta)^2 = \frac{1}{n+m-2r}$ the latter expression gives

$$\mathbb{E}[z] \leq \\ \frac{8\beta^2}{3}\log\left\{1 + e^{\frac{3}{4\beta^2}} + \frac{nm-s}{2}\right\} .$$

The bound in Proposition 4 (skippin 1+)becomes

$$(r+s)\frac{8\beta^2}{3}\log\left\{1 + e^{\frac{3}{4\beta^2}} + \frac{nm-s}{2}\right\} \\ \leq c_1(r+s)\log\left\{c_2 + \frac{nm-s}{2}\right\}$$

where $c_1 = \frac{8}{3}$ and $c_2 = 1 + e^{\frac{3}{4\beta^2}} \leq 2.3$.

Lemma 4 *The variance (see (Tropp, 2010)) of the set of $\Phi(E_{i,j})$s where $1 \leq i \leq n$, $1 \leq j \leq m$ is bounded by*

$$\sigma^2 = \\ \|\sum_{i=1}^{n}\sum_{j=1}^{m}\Phi(E_{i,j})\Phi(E_{i,j})^\top\|_{op} \\ \vee \|\sum_{i=1}^{n}\sum_{j=1}^{m}\Phi(E_{i,j})^\top\Phi(E_{i,j})\|_{op} \\ \leq (1 + (n \vee m))(1-\beta)^2 + \beta^2$$

Proof of Lemma 4. Lets recall for $E^{n_1,m_1}_{i_1,j_1}$ and $E^{n_2,m_2}_{i_2,j_2}$ denoting canonical elements of size $n_1 \times m_1$ and $n_2 \times m_2$, the Kronecker product expression:

$$E^{n_1,m_1}_{i_1,j_1} \otimes E^{n_2,m_2}_{i_2,j_2} = E^{n_1 n_2, m_1 m_2}_{(i_1-1)n_2+i_2,(j_1-1)m_2+j_2} .$$

Using this and by expressing $I_n = \sum_{i=1}^{n} E^{n,n}_{i,i}$, after some algebra we get

$$\Phi(E_{i,j})\Phi(E_{i,j})^\top = \\ \Big((1-\beta)^2 E^{m,m}_{j,j} \otimes I_n \\ + (1-\beta)^2 I_m \otimes E^{n,n}_{i,i} + \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)}\Big) .$$

Adding up the terms results in a very simple object:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\Phi(E_{i,j})\Phi(E_{i,j})^\top \\ = (1-\beta)^2 I_m \otimes I_n + (1-\beta)^2 I_m \otimes I_n + \beta^2 I_{nm} \\ = \left(2(1-\beta)^2 + \beta^2\right) I_{nm} .$$

The second term is also quite friendly, in fact

$$\Phi(E_{i,j})^\top \Phi(E_{i,j}) =$$

$$\begin{pmatrix} (1-\beta)^2 E^{n,n}_{i,i} \otimes I_n & (1-\beta)^2 E^{n^2,m^2}_{in,jm} & \beta(1-\beta) E^{n^2,nm}_{ni,n(j-1)+i} \\ (1-\beta)^2 E^{m^2,n^2}_{jm,in} & (1-\beta)^2 I_m \otimes E^{m,m}_{j,j} & \beta(1-\beta) E^{m^2,nm}_{mj,n(j-1)+i} \\ \beta(1-\beta) E^{nm,n^2}_{i+n(j-1),ni} & \beta(1-\beta) E^{nm,m^2}_{i+n(j-1),mj} & \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)} \end{pmatrix}$$

$$= \begin{pmatrix} (1-\beta)^2 \sum_{k=1}^n E^{n^2,n^2}_{n(i-1)+k,n(i-1)+k} & (1-\beta)^2 E^{n^2,m^2}_{in,jm} & \beta(1-\beta) E^{n^2,nm}_{ni,n(j-1)+i} \\ (1-\beta)^2 E^{m^2,n^2}_{jm,in} & (1-\beta)^2 \sum_{k=1}^m E^{m^2,m^2}_{j+(k-1)m,j+(k-1)m} & \beta(1-\beta) E^{m^2,nm}_{mj,n(j-1)+i} \\ \beta(1-\beta) E^{nm,n^2}_{i+n(j-1),ni} & \beta(1-\beta) E^{nm,m^2}_{i+n(j-1),mj} & \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)} \end{pmatrix}$$

$$= \begin{pmatrix} (1-\beta)^2 E^{n^2,n^2}_{ni,ni} & (1-\beta)^2 E^{n^2,m^2}_{in,jm} & \beta(1-\beta) E^{n^2,nm}_{ni,n(j-1)+i} \\ (1-\beta)^2 E^{m^2,n^2}_{jm,in} & (1-\beta)^2 E^{m^2,m^2}_{mj,mj} & \beta(1-\beta) E^{m^2,nm}_{mj,n(j-1)+i} \\ \beta(1-\beta) E^{nm,n^2}_{i+n(j-1),ni} & \beta(1-\beta) E^{nm,m^2}_{i+n(j-1),mj} & \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)} \end{pmatrix}$$

$$+ \begin{pmatrix} (1-\beta)^2 \sum_{k\neq i}^n E^{n^2,n^2}_{n(i-1)+k,n(i-1)+k} & 0_{n^2,m^2} & 0_{n^2,nm} \\ 0_{m^2,n^2} & (1-\beta)^2 \sum_{k\neq j}^m E^{m^2,m^2}_{j+(k-1)m,j+(k-1)m} & 0_{m^2,nm} \\ 0_{nm,n^2} & 0_{nm,m^2} & 0_{nm,nm} \end{pmatrix}$$

Adding up the terms we get on the one hand matrices having only diagonal terms (from the second term of the last equality) and on the other hand (first term) pairwise orthogonal matrices which are also orthogonal to the diagonal terms. The second bunch of matrices that can be written, up to row and column permutations, as the following matrix

$$\begin{pmatrix} (1-\beta)^2 m & (1-\beta)^2 & \beta(1-\beta) \\ (1-\beta)^2 & (1-\beta)^2 n & \beta(1-\beta) \\ \beta(1-\beta) & \beta(1-\beta) & \beta^2 \end{pmatrix} = \begin{pmatrix} 1-\beta & 0 & 0 \\ 0 & 1-\beta & 0 \\ 0 & 0 & \beta \end{pmatrix} \begin{pmatrix} m & 1 & 1 \\ 1 & n & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1-\beta & 0 & 0 \\ 0 & 1-\beta & 0 \\ 0 & 0 & \beta \end{pmatrix} \ .$$

Using triangle inequality

$$\left\| \begin{pmatrix} (1-\beta)^2 m & (1-\beta)^2 & \beta(1-\beta) \\ (1-\beta)^2 & (1-\beta)^2 n & \beta(1-\beta) \\ \beta(1-\beta) & \beta(1-\beta) & \beta^2 \end{pmatrix} \right\|_{op}$$

$$= \left\| \begin{pmatrix} 1-\beta & 0 & 0 \\ 0 & 1-\beta & 0 \\ 0 & 0 & \beta \end{pmatrix} \left\{ \begin{pmatrix} m-1 & 0 & 0 \\ 0 & n-1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \right\} \begin{pmatrix} 1-\beta & 0 & 0 \\ 0 & 1-\beta & 0 \\ 0 & 0 & \beta \end{pmatrix} \right\|_{\mathrm{op}}$$

$$\leq (1-\beta)^2 (1 + (n \vee m)) + \beta^2 \quad ,$$

so

$$\sum_{i=1}^n \sum_{j=1}^m \Phi(E_{i,j})^\top \Phi(E_{i,j}) \leq (1-\beta)^2 (1 + (n \vee m)) + \beta^2 \ . \square$$