Intersecting singularities for multi-structured estimation

Emile Richard, Francis Bach, Jean-Philippe Vert

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 dInformatique de l'Ecole Normale Supérieure, Paris, France

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

Abstract

We address the problem of designing a convex nonsmooth regularizer encouraging multiple structural effects simultaneously. Focusing on the inference of sparse and low-rank matrices we suggest a new complexity index and a convex penalty approximating it. The new penalty term can be written as the trace norm of a linear function of the matrix. By analyzing theoretical properties of this family of regularizers we come up with oracle inequalities and compressed sensing results ensuring the quality of our regularized estimator. We also provide algorithms and supporting numerical experiments.

1. Introduction

Estimating high-dimensional *simple* objects from a few measurements has been widely investigated in statistics and optimization due to the large body of potential applications. In biology, where measurements are expensive to obtain and the data very complex, building reliable predictors from the limited amount of available data plays an increasing role in medicine. In web applications building recommender systems presents the same kind of challenges with huge economic impacts. In some applications such as breakpoint detection (Vert & Bleakley, 2010), clique detection (Alon et al., 1998; Doan & Vavasis, 2010) that is closely related to recommender systems, compressed sensing (Golbabaee & Vandergheynst, 2012) and sparse principal component analysis (d'Aspremont et al., 2007), the goal is to capture rich objects on which *multiple* structural informations are available. In such cases, at the first glance, one would expect the inference to be simpler when more structural information is known, as we search for the solution in a smaller space: the intersections of the multiple low-complexity sets. However, to the best of our knowledge, no general methodology exists for combining multiple priors to recover objects having *simultaneously* the different structures.

A popular methodology for incorporating particular effects in the solution is to use convex regularizers $\mathcal{R}(w)$ that are nondifferentiable at points w that fulfill the structural constraints, such as sparse vectors or low-rank matrices (Bach et al., 2011; Chandrasekaran et al., 2012). Assume we are given nonsmooth regularizers $\mathcal{R}_1$ and $\mathcal{R}_2$, each inducing a particular desired behavior, and we want to build a new regularizer inducing both properties. A natural approach is to add together both regularizers to form a joint regularizer $\mathcal{R}_+ = \mathcal{R}_1 + \mathcal{R}_2$, to enforce both constraints (Richard et al., 2012). However adding regularizers encourages objects having one *or* the other property, not necessarily both at the same time. The infimal convolution $\mathcal{R}_\star(w) = \inf_{w_1+w_2=w} \mathcal{R}_1(w_1)+\mathcal{R}_2(w_2)$ corresponds to modelling objects as the sum of two terms, each of them respectively penalized by one of the original regularizer (Candès et al., 2009; Chandrasekaran et al., 2011). This is again different from finding objects presenting the two properties simultaneously. Taking the maximum $\mathcal{R}_\vee = \max(\mathcal{R}_1, \mathcal{R}_2)$ as suggested by Oymak et al. (2012) has the obvious drawback of promoting points on which regularizers take equal value, which is not the goal.

In this paper, we propose a new approach to combine structure-inducing penalties, focusing on the problem of inferring sparse and low-rank matrices. Our approach is based on analyzing the geometry of the space

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

around singular points. To combine two regularizers, we suggest to intersect singularities and relax the new index measuring simultaneously both properties rather than adding regularizers. The subdifferential of a convex function at a singular point is a convex set not reduced to a singleton. We consider the cases where this convex set lies in a well characterized subspace, and by intersecting singularities we mean intersecting the subspaces corresponding to each penalty term. We aim at expanding the intersection or equivalently reducing the complement. We study the case of sparse low-rank matrices where by intersecting the subspaces in which subdifferentials of each of the regularizers lie we build a new measure for matrices which we call *ranksity*: the dimension of the complement to the intersection space (Section 2).

We consider a convex relaxation of ranksity as the trace norm of a linear function of the unknown, and show that the standard sum of regularizers can be written in a similar way. In Section 3 we provide a theoretical analysis of this family of regularizers from statistical and compressed sensing point of views. In Section 4 we provide algorithmic schemes to solve problems of interest and finally show in Section 5 numerical experiments showing the applicability of our regularizer and comparison with baselines.

In the sequel, n and m are integers and w.l.o.g. $n \geq m$. For any matrix $X \in \mathbb{R}^{n\times m}$ the notations $\|X\|_F$, $\|X\|_1$, $\|X\|_\infty$, $\|X\|_0$, $\|X\|_*$ and $\|X\|_{op}$, $\mathrm{rank}(X)$ stand for the Frobenius norm, the entry-wise ℓ_1 and ℓ_∞ norms, the number of nonzero elements, the trace-norm (or nuclear norm, the sum of the singular values), the operator norm (the largest singular value) and the rank of X. The letters r and s denote the rank and the sparsity index of X. Given matrices A and B, we denote by $\langle A, B\rangle = \mathrm{tr}(A^\top B)$, $A \circ B$ and $A \otimes B$ the inner product, the Hadamard and the Kronecker products of matrices. A vector in $\mathbb{R}^d$ is always understood as a $d\times 1$ matrix and $\mathbf{vec}(X)$ denotes the vectorized version of X, $\mathrm{Diag}(x)$ and $\mathrm{diag}(X)$ are respectively the matrix having the vector x at its diagonal and 0s elsewhere and the vector formed by $X_{i,i}$s. The matrix $|X|$ contains the absolute values of entries of X and $\mathrm{sgn}(X)$ is the sign matrix associated with X with the convention $\mathrm{sgn}(0) = 0$. We denote by $U_X\Sigma_X V_X$ the singular value decomposition of X, and we define the subspaces of $\mathbb{R}^{n\times m}$, $\mathbf{span}(X)$ and $\mathbf{supp}(X)$ as the ranges of linear applications $(A, B) \mapsto AX + XB \in \mathbb{R}^{n\times m}$ and $C \mapsto C \circ X$ respectively. We denote by $\mathcal{P}_X$, $\mathcal{P}_X^\perp$, $\mathcal{Q}_X$ and $\mathcal{Q}_X^\perp$ the orthogonal projectors onto $\mathbf{span}(X)$, $\mathbf{span}^\perp(X)$, $\mathbf{supp}(X)$ and $\mathbf{supp}^\perp(X)$ respectively.

2. Add or intersect singularities?

Nonsmooth convex regularizers (Bach et al., 2011; Chandrasekaran et al., 2012) have recently received tremendous interest for estimating objects having particular structural properties. Indeed, their convexity makes them computationally attractive: they lead to polynomially converging algorithmic schemes that are easy to implement. From a statistical point of view their analysis benefits from a relatively good understanding of the behaviors and a series of theoretical results ensure the quality of the provided estimators. The nondifferentiable points of such penalties attract the minimizers of optimization procedures and this is the key to their success. The location and the strength of these promoted points can be read in the penalty's subgradients expressions. The subgradients of the trace-norm and the ℓ_1 norm, which are widely used to infer respectively low-rank and sparse matrices, are the sets

$$U_X V_X^\top + \mathcal{P}_X^\perp(\mathcal{B}_{op}) \quad \text{and} \quad \mathrm{sgn}(X) + \mathcal{Q}_X^\perp(\mathcal{B}_\infty),$$

$\mathcal{B}_{op}$ and $\mathcal{B}_\infty$ being the unit balls of the operator and ℓ_∞ norms respectively. A point is nondifferentiable when the subgradient at this point is not reduced to a singleton. To understand the strength of the nondifferentiability we recall that the normal cone of a convex function at a given point is the set of points obtained by multiplying an element of the subgradient by a nonnegative real number. Let us define the dimension of a cone as the dimension of its affine hull. The dimension of the normal cone is a fair measure of the singularity at a given point. From the subgradient expressions one can see that the singularity of the ℓ_1 norm at X is reflected onto the dimension of $\mathbf{supp}(X)^\perp$ through the range of $\mathcal{Q}_X^\perp$, and similarly, the nondifferentiability of the trace norm at X is reflected onto $\dim(\mathbf{span}(X)^\perp)$ through $\mathcal{P}_X^\perp$. This makes the subspaces $\mathbf{span}(X)$ and $\mathbf{supp}(X)$ privileged subspaces for trace-norm and ℓ_1 norm penalized estimation procedures. In fact they are respectively the tangent spaces to the manifolds of rank $r = \mathrm{rank}(X)$ matrices and $s = \|X\|_0$ sparse matrices at X. In the following we discuss two alternative possibilities for building a regularizer for sparse low-rank estimation.

2.1. Summing: the "Trace + 1" penalty

For estimating sparse low-rank matrices, previous approaches (Richard et al., 2012; Oymak et al., 2012; Doan & Vavasis, 2010) have suggested to add regularizers resulting in the "*trace + 1*" penalty $X \mapsto (1-\beta)\|X\|_* + \beta\|X\|_1$. The subgradient of this penalty is not reduced to a singleton as soon as the subgradient of the ℓ_1 or the trace norm component is

not a singleton, *i.e.*, when the matrix is either sparse *or* low-rank. By letting

$$\Pi(X) = \begin{pmatrix} (1-\beta)X & \mathbf{0}_{n\times nm} \\ \mathbf{0}_{nm\times m} & \beta\ \mathrm{Diag}(\mathbf{vec}(X)) \end{pmatrix} \ ,$$

the trace + 1 penalty can be reformulated as $\|\Pi(X)\|_*$, the trace norm of $\Pi(X)$. It can be thought of as a convex relaxation of the index $\operatorname{rank}\Pi(X) = \operatorname{rank}(X) + \|X\|_0$. This index measures the sparsity and the rank by taking its maximal value $nm + n \wedge m$ when X is full rank *and* dense. Note that this index does not penalize density and high-rank in a disjunctive manner, in the sense that it does not have large values when X is dense *or* full-rank.

2.2. Intersecting: the ranksity index

A drawback of adding nonsmooth penalties is that the singularities of the sum are located on the union of singularities of each of the components and not at the intersection of them. We argue that if the goal is to measure matrices having both the sparse *and* low-rank properties, the penalty to use should be nondifferentiable at points which are both sparse and low-rank, namely at the intersection of the singularities of the ℓ_1 and singularities of the trace norm. For building such a penalty we will build a norm such that the dimension of its normal cone at a given point X is given by the dimension of $\mathbf{span}(X)^\perp \cap \mathbf{supp}(X)^\perp$, the intersection of the normal cones of both individual penalties. To this end let us first define the *ranksity* index as the dimension of the orthogonal space to $\mathbf{span}(X)^\perp \cap \mathbf{supp}(X)^\perp$ that is precisely $\mathbf{span}(X) + \mathbf{supp}(X)$:

$$\mathbf{ranksity}(X) = \dim\left(\mathbf{span}(X) + \mathbf{supp}(X)\right) \ . \quad (1)$$

$\mathbf{ranksity}(X)$ takes its maximum value nm on matrices X that are either dense (possibly low-rank) or full rank (possibly sparse). In fact if X is dense, then $\mathbf{supp}(X) = \mathbb{R}^{n\times m}$, and if X is full rank, then $\mathbf{span}(X) = \mathbb{R}^{n\times m}$. Providing a closed form expression for ranksity is not straightforward in general, but for instance in the case of block-diagonal adjacency (binary-valued) matrices X having r nonzero non-overlapping blocks of size $k_i \times l_i$ we can obtain by recursion that

$$\mathbf{ranksity}(X) = r(m+n-r) + \sum_{i=1}^{r}(k_i-1)(l_i-1) \ .$$

It is convenient to have in mind the relationships with the rank and the sparsity index:

$$\dim(\mathbf{supp}(X)) = s \ \ , \ \ \dim(\mathbf{span}(X)) = (n+m-r)r$$

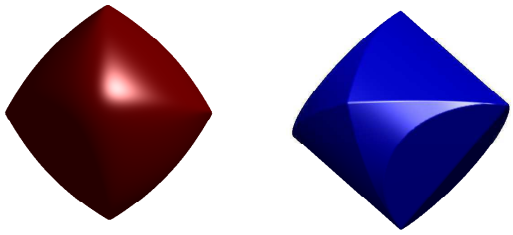

Figure 1. Unit balls of the block norm for $\beta = .7$, $\{X\ ,\ \|\Phi(X)\|_* \le 1\}$ (red) and of the "trace +1" with $\beta = .5$ $\{X\ ,\ \|\Pi(X)\|_* \le 1\}$ (blue) where $X = \begin{pmatrix} x & y \\ y & z \end{pmatrix}$.

based on which we can easily derive the bounds

$$s \vee (m+n-r)r \le \mathbf{ranksity} \le s + (m+n-r)r$$

which show that this nonconvex discontinuous function is sandwiched by two non-decreasing functions of the rank and the sparsity index.

2.3. A convex regularizer for low ranksity estimation

Recall that $\mathbf{vec}(AX) = (X^\top \otimes I_n)\mathbf{vec}(A)$ and $\mathbf{vec}(XB) = (I_m \otimes X)\mathbf{vec}(B)$. It follows that for any $(A,B,C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{m\times m} \times \mathbb{R}^{n\times m}$, we have

$$\mathbf{vec}(AX + XB + X \circ C) =$$
$$[X^\top \otimes I_n\ ,\ I_m \otimes X\ ,\ \mathrm{Diag}(\mathbf{vec}(X))] \begin{pmatrix} \mathbf{vec}(A) \\ \mathbf{vec}(B) \\ \mathbf{vec}(C) \end{pmatrix} \ ;$$

the term inside the **vec** on the left hand side describes precisely the sum of the subspaces $\mathbf{span}(X)$ and $\mathbf{supp}(X)$ used to define ranksity in (1). This is why, after weighting the terms to control the tradeoffs, we define

$$\Phi(X) =$$
$$[(1-\beta)X^\top \otimes I_n\ ,\ (1-\beta)I_m \otimes X\ ,\ \beta\ \mathrm{Diag}(\mathbf{vec}(X))] \ .$$

This lifting is built so that for any $\beta \in \ ]0,1[$, the range of the matrix $\Phi(X)$ is isomorphic to $\mathbf{span}(X) + \mathbf{supp}(X)$. Using this fundamental property we can state the closed-form expression (valid for $\beta \in \ ]0,1[$):

$$\operatorname{rank}\Phi(X) = \mathbf{ranksity}(X) \ .$$

This property suggests in turn to consider $\|\Phi(X)\|_*$ as a convex surrogate of $\mathbf{ranksity}(X)$, which we call

block norm and which can be used as a regularizer to infer low-ranksity matrices. Notice that for $\beta = 0$, $\|\Phi(X)\|_* = \|\Pi(X)\|_* = \|X\|_1$ and for $\beta = 1$, $\|\Phi(X)\|_* \leq (n+m)\|X\|_*$ and $\|\Pi(X)\|_* = \|X\|_*$. For $\beta \in]0,1[$, $\|\Phi(\cdot)\|_*$ has no singularities on matrices that are full-rank and sparse or low-rank and dense as opposed to $\|\Pi(\cdot)\|_*$ which has this undesirable property. In Figure 1 one can clearly see that in the case of the "trace + 1" penalty on $X = \begin{pmatrix} x & y \\ y & z \end{pmatrix}$, the singularities mimic the shapes of a cylinder that represents the unit ball of the trace norm and that of the unit ball of the ℓ_1 norm. As opposed, the block norm ball has only 4 nonsmooth points located at $\begin{pmatrix} \pm 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 \\ 0 & \pm 1 \end{pmatrix}$ that exactly correspond to the intersections of the singularities of the ℓ_1 and trace norm balls.

3. 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.

3.1. 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

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

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:

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

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 $\omega(X) = \omega(X^\star)$.

3.2. 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*

$$\|\mathcal{P}^\perp_{\Lambda(X)}(\Lambda(B))\|_* \leq \kappa \|\mathcal{P}_{\Lambda(X)}(\Lambda(B))\|_* \ . \qquad (2)$$

The restricted eigenvalue *of ω at X is*

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

Define the objective

$$\mathcal{L}(X) = \frac{1}{d}\|\omega(X) - y\|_2^2 + \lambda\|\Lambda(X)\|_* \ , \qquad (3)$$

and consider the following estimation procedure

$$\widehat{X} = \arg\min_{X\in\mathcal{S}} \mathcal{L}(X) \ , \qquad (4)$$

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:*

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

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 $\lambda \geq \frac{2}{nm}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$

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

3.3. 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 $1 - e^{-t}$,

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

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 $1 - e^{-t}$,

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

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 (4) outputs the optimal estimators for the nonconvex penalties $\operatorname{rank} + \ell_0$ 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.

3.4. Compressed sensing and exact recovery

Consider the constrained convex optimization problem

$$\min_X \|\Lambda(X)\|_* \ \text{ s.t. } \ \omega(X) = \omega(X^\star) \ , \tag{5}$$

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 (5) where* Λ *is an orthogonal lifting is at most*

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

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 $\|\Phi(G)\|_{op}^2$ 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 (5).*

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

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

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 $d_\Phi \leq 4\{(n+m-1)+(k-1)(l-1)\}\log(n+m)$.

4. Algorithms for minimizing Λ-trace penalized objectives

4.1. A Chambolle-Pock algorithm for general losses

The Λ-trace penalties have similarities with total variation minimization as in both cases a *simple* (having an easy to compute proximal operator) norm of a linear function of the variable is being minimized. In our case the unconstrained optimization problem

$$\min_X \ell(X) + \lambda\|\Lambda(X)\|_*$$

can be re-written as a primal-dual problem

$$\min_X \max_Z \langle \lambda\Lambda(X), Z\rangle - \delta_{\mathcal{B}_{op}}(Z) + \ell(X) \ ,$$

where $\delta_{\mathcal{B}_{op}}$ is the indicator of the unit ball of the operator norm. The Chambolle-Pock framework (Chambolle & Pock, 2011) applies and we can derive the Algorithm 1 for cases where ℓ is convex and simple. The accelerated algorithm applies for cases where ℓ in addition to being convex has a Lipschitz continuous gradient, for instance in least squares regression. In these settings, the second prox step is replaced by a gradient descent step and the tuning parameters α, ξ, θ are updated at each step. We refer to Chambolle & Pock (2011) for technical details such as the choices of tuning parameters that were set according to the paper's remarks in our experiments. In numerical experiments following the lifting Λ it can be convenient to use other algorithms, for instance in case of Π we used ADMM (see Boyd et al., 2011, for a survey) in our experiments.

Algorithm 1 Chambolle-Pock algorithm for Λ-trace penalized optimization

Initialize $X, Z, \bar{X}, \alpha, \xi, \theta$
Repeat until convergence

$$\begin{aligned} Z &\leftarrow \mathrm{proj}_{\mathcal{B}_{op}}(Z + \alpha\lambda\Lambda(\bar{X})) \\ X^{new} &\leftarrow \mathrm{prox}_{\xi\ell}(X - \xi\lambda\Lambda^*(Z)) \\ \bar{X} &\leftarrow X^{new} + \theta(X^{new} - X) \\ X &\leftarrow X^{new} \end{aligned}$$

4.2. A Frank-Wolfe algorithm for smoothly differentiable losses and orthogonal liftings

In cases where the loss function ℓ to minimize has a Lipschitz continuous gradient (least squares regression is a standard example) rather than penalizing the loss by the lifted trace norm one can solve the following constrained optimization problem:

$$\min_X \ell(X) \quad \text{s.t.} \quad \|\Lambda(X)\|_* \leq C \ ,$$

where C is a constant replacing the tuning parameter λ in this setting. The advantage of this equivalent formulation is in the possibility to use algorithmic schemes offered by Frank-Wolfe or conditional gradient algorithm (see Jaggi, 2013, for a recent survey). We argue that these algorithms allow to save both computational and memory resources as they require only the top singular vectors of a $n' \times m'$ matrix rather than the full SVD of it at each iteration. Frank-Wolfe algorithm in this situation requires at each iteration k to solve the following linear subproblem

$$\min_S \langle \nabla\ell(X_k), S\rangle \quad \text{s.t.} \quad \|\Lambda(S)\|_* \leq C \ ,$$

which in the case of orthogonal liftings (*i.e.* $\Lambda^*\Lambda = \|\Lambda\|^2 id$ such as Π and Φ) can be written using the variable $\Sigma = \Lambda(S)$ as

$$\min_\Sigma \langle \nabla\ell(X_k), \Lambda^*(\Sigma)\rangle \quad \text{s.t.} \quad \|\Sigma\|_* \leq C$$

and therefore reduces to

$$\min_\Sigma \langle \Lambda\left(\nabla\ell(X_k)\right), \Sigma\rangle \quad \text{s.t.} \quad \|\Sigma\|_* \leq C \ .$$

The latter optimization problem is a linear problem to solve over an atomic set. We know that the top singular vectors of $\Lambda\left(\nabla\ell(X_k)\right)$ provide a fairly good approximate solution to the problem and in addition they present the advantage of being extremely fast to obtain thanks to the Lanczos method. The pseudo code can be found in Algorithm 2 and we refer to Jaggi (2013) for further technical details such as variants using refined step-sizes and the stopping criterion.

Algorithm 2 Frank-Wolfe algorithm for Λ-trace penalized optimization

Initialize $X_0 = 0$
for $k = 0 \cdots K$ **do**
Compute top singular vectors of $\Lambda(\nabla\ell(X_k))$: u, v
Update $X_{k+1} = (1-\gamma)X_k - \gamma\frac{C}{\|\Lambda\|^2}\Lambda^*(uv^\top)$ where $\gamma = \frac{2}{k+2}$
end for

5. Numerical experiments

5.1. Nonsmooth ℓ: decomposing simply structured noise and doubly structured signal

We know from Chandrasekaran et al. (2011) that a matrix that is built by adding a sparse matrix to a low-rank matrix can be decomposed onto its components by solving the so-called robust PCA problem. It is known however that robust PCA fails in recovering the additive components when the two types of structure are present in one of the components *e.g.* when the low rank component is itself sparse. In this work we specifically focus on the two following challenging tasks:

1. **[SL +S]** The observation is the sum of a sparse component representing the noise, and a sparse low-rank component which is the signal. The problem is closely related to the *planted clique* (Alon et al., 1998) problem that is popular in theoretical computer science: the simply sparse component can be seen as the adjacency matrix of the random graph and the adjacency matrix of the clique is sparse and has rank 1. In this setup, if Y denotes the observation, the objective function to minimize is $\mathcal{L}(X) = \|X-Y\|_1 + \lambda\|\Lambda(X)\|_*$ where we refer to robust PCA, trace + 1 and block norm to choices of Λ being the identity, Π and Φ.

2. **[SL + L]** In this setup the observation is the sum of a sparse-low rank (signal) and a low-rank and dense (noise) matrix. The goal is to separate the two. Such a situation is met in high dimensional underdetermined settings when estimating a covariance matrix based on the sample covariance, and we know that some groups of highly related variables are present forming blocks in the covariance matrix. In this context the objective function is $\mathcal{L}(X) = \|X - Y\|_* + \lambda\|\Lambda(X)\|_*$ and we deserve the lifting $\mathrm{Diag}(\mathbf{vec}(X))$ for robust PCA.

In each case we compared our algorithms to elementwise thresholding of the observation and to the singu-

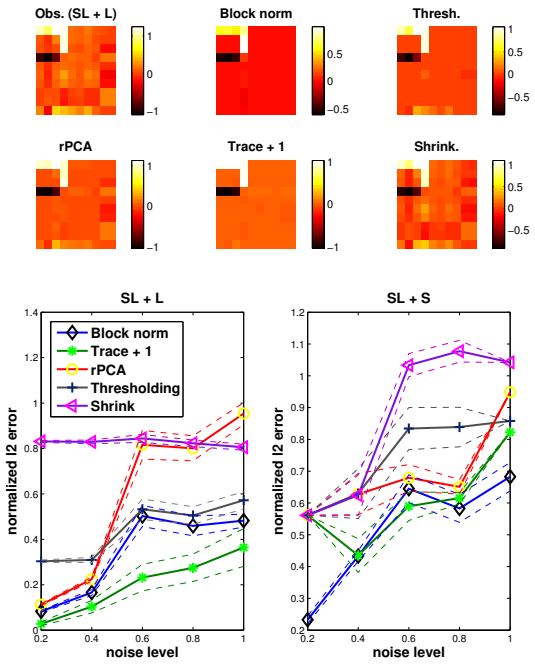

Figure 2. Up: illustration of the **[SL +L]** experimental setup. Bottom: comparison of the performance of various algorithms on decomposing a signal onto sparse-low rank and other component.

lar value thresholding of the matrix sometimes called shrinkage. Since minimizing an objective containing $\|\Phi(X)\|_*$ as regularizer requires singular value decompositions of matrices of size $nm \times (n^2 + m^2 + mn)$ at each iteration, the $O(n^6)$ cost of each iteration pushed us to compare the algorithms on relatively small data sets in our numerical experiments. For simulations we took $n = m = 10$ and the ground truth $X^\star$ was a matrix of rank two having 9 nonzero elements. In the first case the noise is generated as i.i.d. gaussian entries at sparse (15%) positions with various variances given by the noise level, and in the second case **[SL + L]** the noise is a rank 2 dense matrix generated proportional to the noise level times the highest singular value of $X^\star$. We selected the tuning parameters using cross-validation and emphasize that 0 and 1 where included in the potential values for β but were not the favorite values following the cross-validation step. The results over 10 runs of the experiment can be found in Figure 2. The two algorithms penalizing the trace + 1 and the block norm are superior to the competitors.

	Denoising		Multitask	
σ^2	Trace + 1	Block	Trace + 1	Block
.01	0.57 ± 0.01	**0.48 ± 0.01**	0.89 ± 0.00	**0.87 ± 0.00**
.02	0.48 ± 0.01	**0.47 ± 0.01**	0.93 ± 0.01	**0.86 ± 0.01**
.05	0.55 ± 0.01	**0.45 ± 0.01**	0.92 ± 0.01	**0.89 ± 0.03**
.10	0.57 ± 0.01	**0.46 ± 0.01**	0.92 ± 0.00	**0.88 ± 0.00**
.15	0.52 ± 0.01	**0.50 ± 0.01**	0.87 ± 0.01	0.87 ± 0.01
.25	0.56 ± 0.01	**0.54 ± 0.01**	0.98 ± 0.04	**0.94 ± 0.03**
.50	0.75 ± 0.02	**0.72 ± 0.02**	0.90 ± 0.02	**0.86 ± 0.01**
1.0	1.00 ± 0.02	1.00 ± 0.02	0.93 ± 0.00	0.93 ± 0.01

Table 1. Relative estimation ℓ_2 errors for multitask learning and denoising. The columns correspond to the regularizers and the type of problem, each row is a noise level denoted by σ^2.

5.2. Smooth ℓ: dense denoising and multitask learning

We performed numerical tests using the Franck-Wolfe algorithms on two different problems. In these experiments $n = 15, m = 10$. The matrix $X^\star$ is generated using sparse factors having $r = 3$ columns.

1. **Denoising.** In this case the observation $Y = X^\star + M$ where M is a matrix having i.i.d. entries drawn from $\mathcal{N}(0, \sigma^2)$. The values of σ^2 are the noise level and we ran experiments for various values (see Table 1). The loss we used is simply $\ell(X) = \frac{1}{2}\|X - Y\|_F^2$.
2. **Multitask learning.** The observation $Y = \Omega X^\star + \epsilon$ is obtained using the design matrix $\Omega \in \mathbb{R}^{q \times n}$ where $q = 8$, the noise ϵ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and $\ell(X) = \frac{1}{2}\|\Omega X - Y\|_F^2$.

In all the simulation the parameters β and λ are chosen using a cross-validation step and again we noticed the algorithms choose values of $\beta \neq 0, 1$ that correspond to basic regularizers. So without explicitely testing them, superiority to the Lasso and trace norm penalized regression is empirically observed. See Table 1 for relative estimation errors $\|X^\star - \hat{X}\|_F / \|X^\star\|_F$ in these experiments over 100 runs.

6. Discussion and perspectives

We built a regularizer that has singularities exactly at the desired points by combining the two priors in a more appropriate way than just adding the penalties.

Generalization. Our main methodological point can be applied to a complexity index that can be written as the dimension of a linear subspace, as it is the case of the ℓ_0 and the rank (recall $\text{rank}(n + m - \text{rank}) = \dim(\textbf{span})$ is an increasing function of the rank). Other penalties have the same flavor. By letting $\|X\|_{2,1} = \sum_{i=1}^n \|X_{i,.}\|_2$ we can concatenate a matrix corresponding to the

linear space spanned by the columns, or namely the range of $(c_1, \cdots, c_m) \mapsto \sum_{i=1}^m X_{.,i} c_i^\top$ which is $\begin{pmatrix} I_m \otimes X_{.,1} & \cdots & I_m \otimes X_{.,m} \end{pmatrix}$ to other blocks. For instance, our rational on the block norm would suggest, instead of using $\|X\|_{2,1} + \theta\|X\|_*$ (Golbabaee & Vandergheynst, 2012) to use the lifted trace norm defined through the lifting

$$[(1-\beta)X^\top \otimes I_n \ , \ (1-\beta) I_m \otimes X \ , \\ \beta I_m \otimes X_{.,1}, \cdots, \beta I_m \otimes X_{.,m}] \ ,$$

and a similar lifting can be written for estimating sparse and row-sparse matrices, instead of using $(1-\beta)\|X\|_{2,1} + \beta\|X\|_1$.

It is important to point out the applicability limits of our theoretical results as well. For instance in the case of the trace-Lasso penalty $w \mapsto \|P \operatorname{Diag}(w)\|_*$ (Grave et al., 2011) that can be seen as a lifted trace norm, the orthogonality assumption on Λ is not verified, making none of the theoretical results applicable. In our preliminary work we tried lifted max norms which have a nice motivation: they relax the rank of the lifted object on the ℓ_∞ ball rather than on the operator norm, which is meaningful when dealing with adjacency matrices. Somehow the empirical results did not change drastically and the max norm makes the analysis harder.

Acknowledgement

This work was supported by the European Research Council (SIERRA-ERC- 239993 and SMAC-ERC-280032). E.R. benefited from the support of the FMJH Program Gaspard Monge in optimization and operation research (EDF).

References

Alon, N., Krivelevich, M., and Sudakov, B. Finding a large hidden clique in a random graph. *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, 1998.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Convex optimization with sparsity-inducing norms. *S. Sra, S. Nowozin, S. J. Wright. editors, Optimization for Machine Learning*, 2011.

Boyd, S., Parikh, N., Chu, E. Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3, 2011.

Candès, E. J., Li, X., Ma, Y., and W., John. Robust principal component analysis? *Journal of ACM*, 8:1–37, 2009.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 2011.

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

Chandrasekaran, V., Sanghavi, S., Parrilo, P.A., and Willsky, A.S. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Opt.*, 21:572–596, 2011.

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.

d'Aspremont, A., El Ghaoui, L., Jordan, M.I., and Lanckriet, G.R.G. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

Doan, X.V. and Vavasis, S.A. Finding approximately rank-one submatrices with the nuclear norm and l1 norm. *arXiv preprint arXiv:1011.1839*, 2010.

Golbabaee, M. and Vandergheynst, P. Compressed sensing of simultaneous low-rank and joint-sparse matrices. *submitted to IEEE transaction in Information Theory*, 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.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th Annual International Conference on Machine Learning*, 2013.

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

Oymak, S., Jalali, A., Fazel, M., Eldar, Y., and Hassibi, B. Simultaneously structured models with application to sparse and low-rank matrices. *submitted*, 2012. URL `http://arxiv.org/pdf/1212.3753v1.pdf`.

Richard, E., Savalle, P.-A., and Vayatis, N. Estimation of simultaneously sparse and low-rank matrices. In *Proceeding of 29th Annual International Conference on Machine Learning*, 2012.

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.

Appendix

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

$$\begin{aligned}&\langle M, \hat{X} - X\rangle \leq && (6)\\ &\sqrt{2\ \mathrm{rank}(\Lambda(M))}\|\mathcal{P}_{\Lambda(X)}(\Lambda(M))\|_{op} && (7)\\ &\|\mathcal{P}_{\Lambda(X)}(\Lambda(\hat{X}-X))\|_F/\|\Lambda\|^2 && (8)\\ &+\|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(M))\|_{op} && (9)\\ &\|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(\hat{X}-X))\|_*/\|\Lambda\|^2 && (10)\end{aligned}$$

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

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

and

$$-\langle Z, \widehat{X}-X\rangle \leq \|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_* - \|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(\widehat{X}))\|_* \ . \quad (11)$$

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

$$\nabla\|\omega(\widehat{X})-y\|_2^2 = 2\langle\omega(\widehat{X}-X^\star),\omega(\widehat{X}-X)\rangle - 2\langle M, \widehat{X}-X\rangle \ . \quad (12)$$

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 $M = \sum_{i=1}^d \epsilon_i\Omega_i$,

$$\begin{aligned}&\langle\partial\mathcal{L}(\widehat{X}), \widehat{X} - X\rangle - \lambda\langle\widehat{Z}-Z, \widehat{X}-X\rangle \leq 0 && (13)\\ \Leftrightarrow\ &\langle\frac{1}{d}\nabla\|\omega(\widehat{X})-y\|_2^2 + \lambda Z, \hat{X}-X\rangle \leq 0 && (14)\\ \Leftrightarrow\ &\frac{2}{d}\langle\omega(\widehat{X}-X^\star),\omega(\widehat{X}-X)\rangle \leq && (15)\\ &\frac{2}{d}\langle M, \widehat{X}-X\rangle - \lambda\langle Z, \widehat{X}-X\rangle \ . && (16)\end{aligned}$$

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 (11) in Lemma 2 and equation (16) imply

$$\lambda\|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(\widehat{X}))\|_* \leq \frac{2}{d}\langle M, \widehat{X}-X\rangle + \lambda\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_* \ . \quad (17)$$

By using Lemma 1 , first inequality (6), 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 (10) and (16) we have

$$\begin{aligned}&\frac{1}{d}\Big(\|\omega(\widehat{X}-X^\star)\|_2^2 + \|\omega(\widehat{X}-X)\|_2^2 - \|\omega(X-X^\star)\|_2^2\Big) \\ &\leq \frac{2}{d}\Big(\sqrt{2\,\mathrm{rank}(\Lambda(X))}\frac{\|\Lambda(M)\|_{op}}{\|\Lambda\|^2}\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_F \\ &\qquad + \frac{\|\Lambda(M)\|_{op}}{\|\Lambda\|^2}\|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(\widehat{X}))\|_*\Big) \\ &+\lambda\sqrt{\mathrm{rank}(\Lambda(X))}\|\mathcal{P}_{\Lambda(X)}(\Lambda(\widehat{X}-X))\|_F - \lambda\|\mathcal{P}^{\perp}_{\Lambda(X)}(\Lambda(\widehat{X}))\|_* \ .\end{aligned} \quad (18)$$

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 $\lambda \geq \frac{3}{d}\|\Lambda(M)\|_{op}/\|\Lambda\|^2$,

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

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

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

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

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

By using assumption 1 and Holder's inequality twice

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

The other bound is obtained in a similar fashion by using Cauchy-Schwarz on the first term and also the fact that $\|\mathcal{P}_{\Lambda(X)}(M)\|_F \leq \sqrt{2\ \operatorname{rank}(\Lambda(X))}\|M\|_F$ 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

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

so we can write

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

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

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

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

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

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

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

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

$$\|Z(G) - G\|_F^2 = \frac{\|\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\|_{op}^2}{\|\Lambda\|^2}\|\Lambda^* U_{\Lambda(X^\star)}V^\perp_{\Lambda(X^\star)}\|_F^2 \leq \|\mathcal{P}^\perp_{\Lambda(X^\star)}(\Lambda(G))\|_{op}^2 \operatorname{rank}(\Lambda(X^\star)) \quad .$$

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 $r = \operatorname{rank}(X^\star)$. First lets show that for any $G \in \mathbb{R}^{n\times m}$

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

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

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

and

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

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

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

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

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

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

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

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

$$\begin{aligned} &\mathbb{E}\exp(t(1-\beta)^2\|H\|_{op}^2) && (19)\\ &= \mathbb{E}f(\|H\|_{op}) && (20)\\ &= \int_1^\infty \mathbb{P}[f(\|H\|_{op}) > h]\ \ dh && (21)\\ &\leq \int_1^{1+f(R)} 1\ dh && (22)\\ &\quad + \int_{1+f(R)}^\infty \mathbb{P}[f(\|H\|_{op}) > h]dh && (23)\\ &= f(R) && (24)\\ &\quad + \int_0^\infty \mathbb{P}[\|H\|_{op} > f^{-1}(f(R)+1+\zeta)]d\zeta && (25)\\ &\leq f(R) && (26)\\ &\quad + \int_0^\infty \mathbb{P}[\|H\|_{op} > R + f^{-1}(1+\zeta)]d\zeta && (27)\\ &\leq f(R) && (28)\\ &\quad + \int_0^\infty 2ts(1-\beta)^2\ \exp\left(-s^2/2 + ts^2(1-\beta)^2\right)ds && (29)\\ &\leq f(R)+1 && (30) \end{aligned}$$

where (27) 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 (30) 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 5 (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 $E_{i_1,j_1}^{n_1,m_1}$ and $E_{i_2,j_2}^{n_2,m_2}$ denoting canonical elements of size $n_1\times m_1$ and $n_2\times m_2$, the Kronecker product expression:

$$E_{i_1,j_1}^{n_1,m_1}\otimes E_{i_2,j_2}^{n_2,m_2}=E_{(i_1-1)n_2+i_2,(j_1-1)m_2+j_2}^{n_1n_2,m_1m_2}.$$

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

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

Adding up the terms results in a very simple object:

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

The second term is also quite friendly, in fact

$$
\begin{aligned}
&\Phi(E_{i,j})^\top \Phi(E_{i,j}) = \\
&\begin{pmatrix}
(1-\beta)^2 E^{n,n}_{i,i} \otimes I_n & (1-\beta)^2 E^{n^2,m^2}_{in,jm} & \beta(1-\beta) E^{n^2,nm}_{ni,n(j-1)+i} \\
(1-\beta)^2 E^{m^2,n^2}_{jm,in} & (1-\beta)^2 I_m \otimes E^{m,m}_{j,j} & \beta(1-\beta) E^{m^2,nm}_{mj,n(j-1)+i} \\
\beta(1-\beta) E^{nm,n^2}_{i+n(j-1),ni} & \beta(1-\beta) E^{nm,m^2}_{i+n(j-1),mj} & \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)}
\end{pmatrix} \\
&= \begin{pmatrix}
(1-\beta)^2 \sum_{k=1}^n E^{n^2,n^2}_{n(i-1)+k,n(i-1)+k} & (1-\beta)^2 E^{n^2,m^2}_{in,jm} & \beta(1-\beta) E^{n^2,nm}_{ni,n(j-1)+i} \\
(1-\beta)^2 E^{m^2,n^2}_{jm,in} & (1-\beta)^2 \sum_{k=1}^m E^{m^2,m^2}_{j+(k-1)m,j+(k-1)m} & \beta(1-\beta) E^{m^2,nm}_{mj,n(j-1)+i} \\
\beta(1-\beta) E^{nm,n^2}_{i+n(j-1),ni} & \beta(1-\beta) E^{nm,m^2}_{i+n(j-1),mj} & \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)}
\end{pmatrix} \\
&= \begin{pmatrix}
(1-\beta)^2 E^{n^2,n^2}_{ni,ni} & (1-\beta)^2 E^{n^2,m^2}_{in,jm} & \beta(1-\beta) E^{n^2,nm}_{ni,n(j-1)+i} \\
(1-\beta)^2 E^{m^2,n^2}_{jm,in} & (1-\beta)^2 E^{m^2,m^2}_{mj,mj} & \beta(1-\beta) E^{m^2,nm}_{mj,n(j-1)+i} \\
\beta(1-\beta) E^{nm,n^2}_{i+n(j-1),ni} & \beta(1-\beta) E^{nm,m^2}_{i+n(j-1),mj} & \beta^2 E^{nm,nm}_{i+n(j-1),i+n(j-1)}
\end{pmatrix} \\
&+ \begin{pmatrix}
(1-\beta)^2 \sum_{k\neq i}^n E^{n^2,n^2}_{n(i-1)+k,n(i-1)+k} & 0_{n^2,m^2} & 0_{n^2,nm} \\
0_{m^2,n^2} & (1-\beta)^2 \sum_{k\neq j}^m E^{m^2,m^2}_{j+(k-1)m,j+(k-1)m} & 0_{m^2,nm} \\
0_{nm,n^2} & 0_{nm,m^2} & 0_{nm,nm}
\end{pmatrix}
\end{aligned}
$$

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

$$
\begin{aligned}
&\left\| \begin{pmatrix}
(1-\beta)^2 m & (1-\beta)^2 & \beta(1-\beta) \\
(1-\beta)^2 & (1-\beta)^2 n & \beta(1-\beta) \\
\beta(1-\beta) & \beta(1-\beta) & \beta^2
\end{pmatrix} \right\|_{op} \\
&= \left\| \begin{pmatrix}
1-\beta & 0 & 0 \\
0 & 1-\beta & 0 \\
0 & 0 & \beta
\end{pmatrix} \left\{ \begin{pmatrix}
m-1 & 0 & 0 \\
0 & n-1 & 0 \\
0 & 0 & 0
\end{pmatrix} + \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{pmatrix} \right\} \begin{pmatrix}
1-\beta & 0 & 0 \\
0 & 1-\beta & 0 \\
0 & 0 & \beta
\end{pmatrix} \right\|_{\text{op}} \\
&\leq (1-\beta)^2 (1 + (n \vee m)) + \beta^2 \quad ,
\end{aligned}
$$

so

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