HAL
open science

Beyond Moore-Penrose Part I: Generalized Inverses that Minimize Matrix Norms

Ivan Dokmanić, Rémi Gribonval

▶ **To cite this version:**

Ivan Dokmanić, Rémi Gribonval. Beyond Moore-Penrose Part I: Generalized Inverses that Minimize Matrix Norms. 2017. hal-01547183v2

HAL Id: hal-01547183

https://inria.hal.science/hal-01547183v2

Preprint submitted on 13 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Beyond Moore-Penrose
Part I: Generalized Inverses that Minimize Matrix Norms

Ivan Dokmanić and Rémi Gribonval

Abstract

This is the first paper of a two-long series in which we study linear generalized inverses that minimize matrix norms. Such generalized inverses are famously represented by the Moore-Penrose pseudoinverse (MPP) which happens to minimize the Frobenius norm. Freeing up the degrees of freedom associated with Frobenius optimality enables us to promote other interesting properties. In this Part I, we look at the basic properties of norm-minimizing generalized inverses, especially in terms of uniqueness and relation to the MPP.

We first show that the MPP minimizes many norms beyond those unitarily invariant, thus further bolstering its role as a robust choice in many situations. We then concentrate on some norms which are generally *not* minimized by the MPP, but whose minimization is relevant for linear inverse problems and sparse representations. In particular, we look at mixed norms and the induced $\ell^p \to \ell^q$ norms. An interesting representative is the *sparse pseudoinverse* which we study in much more detail in Part II.

Next, we shift attention from norms to matrices with interesting behaviors. We exhibit a class whose generalized inverse is always the MPP—even for norms that normally result in different inverses—and a class for which many generalized inverses coincide, but not with the MPP. Finally, we discuss efficient computation of norm-minimizing generalized inverses.

1 Introduction

Generalized inverses arise in applications ranging from over- and underdetermined linear inverse problems to sparse representations with redundant signal dictionaries. The most famous generalized matrix inverse is the Moore-Penrose pseudoinverse, which happens to minimize a slew of matrix norms. In this paper we study various alternatives. We start by introducing some of the main applications and discussing motivations.

Linear inverse problems. In discrete linear inverse problems, we seek to estimate a signal $\mathbf{x}$ from measurements $\mathbf{y}$, when they are related by a linear system, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$, $\mathbf{A} \in \mathbb{C}^{m\times n}$ up to a noise term $\mathbf{n}$. Such problems come in two rather different flavors: overdetermined ($m > n$) and underdetermined ($m < n$). Both cases may occur in the same application, depending on how we tune the modeling parameters. For example, in computed tomography the entries of the system matrix quantify how the ith ray affects the jth voxel. If we target a coarse resolution (fewer voxels than rays), $\mathbf{A}$ is tall and we deal with an overdetermined system. In this case, we may estimate $\mathbf{x}$ from $\mathbf{y}$ by applying a generalized (left) inverse, very often the Moore-Penrose pseudoinverse (MPP).[1] When the system is underdetermined ($m < n$), we need a suitable signal model to get a meaningful solution. As we discuss in Section 2.4, for most common models (e.g.

[1] The Moore-Penrose pseudoinverse was discovered by Moore in 1920 [Moo20], and later independently by Penrose in 1955 [Pen08].

sparsity), the reconstruction of $\mathbf{x}$ from $\mathbf{y}$ in the underdetermined case is no longer achievable by a linear operator in the style of the MPP.

Redundant representations. In redundant representations, we represent lower-dimensional vectors through higher-dimensional frame and dictionary expansions. The frame expansion coefficients are computed as $\boldsymbol{\alpha} = \mathbf{A}^*\mathbf{x}$, where the columns of a fat $\mathbf{A}$ represent the frame vectors, and $\mathbf{A}^*$ denotes its conjugate transpose. The original signal is then reconstructed as $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$, where $\mathbf{D}$ is a *dual frame* of $\mathbf{A}$, such that $\mathbf{D}\mathbf{A}^* = \mathbf{I}$. There is a unique correspondence between dual frames and generalized inverses of full rank matrices. Different duals lead to different reconstruction properties in terms of resilience to noise, resilience to erasures, computational complexity, and other figures of merit [KC08]. It is therefore interesting to study various duals, in particular those optimal according to various criteria; equivalently, it is interesting to study various generalized inverses.

Generalized inverses beyond the MPP. In general, for a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, there are many different generalized inverses. If $\mathbf{A}$ is invertible, they all match. The MPP $\mathbf{A}^\dagger$ is special because it optimizes a number of interesting properties. Much of this optimality comes from geometry: for $m < n$, $\mathbf{A}^\dagger\mathbf{A}$ is an orthogonal projection onto the range of $\mathbf{A}^*$, and this fact turns out to play a key role over and over again. Nevertheless, the MPP is only one of infinitely many generalized inverses, and it is interesting to investigate the properties of others. As the MPP minimizes a particular matrix norm—the Frobenius norm[2]—it seems natural to study alternative generalized inverses that minimize different matrix norms, leading to different optimality properties. Our initial motivation for studying alternative generalized inverses is twofold:

(i) *Efficient computation:* Applying a sparse pseudoinverse requires less computation than applying a full one [DKV13, LLM13, CHKK, KKL12]. We could take advantage of this fact if we knew how to compute a sparse pseudoinverse that is in some sense stable to noise. This sparsest pseudoinverse may be formulated as

$$\mathsf{ginv}_0(\mathbf{A}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{X}} \ \|\mathrm{vec}(\mathbf{X})\|_0 \ \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{I} \tag{1}$$

where $\|\cdot\|_0$ counts the total number of non-zero entries in a vector and $\mathrm{vec}(\cdot)$ transforms a matrix into a vector by stacking its columns. The non-zero count gives the naive complexity of applying $\mathbf{X}$ or its adjoint to a vector. Solving the optimization problem (1) is in general NP-hard [Nat95, DMA97], although we will see that for most matrices $\mathbf{A}$ finding *a* solution is trivial and not very useful: just invert any full-rank $m \times m$ submatrix and zero the rest. This strategy is not useful in the sense that the resulting matrix is poorly conditioned. On the other hand, the vast literature establishing equivalence between ℓ^0 and ℓ^1 minimization suggests to replace (1) by the minimization of the entrywise ℓ^1 norm

$$\mathsf{ginv}_1(\mathbf{A}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{X}} \ \|\mathrm{vec}(\mathbf{X})\|_1 \ \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{I}. \tag{2}$$

Not only is (2) computationally tractable, but we will show in Part II that unlike inverting a submatrix, it also leads to well-conditioned matrices which are indeed sparse.

(ii) *Poor man's ℓ^p minimization:* Further motivation for alternative generalized inverses comes from an idea to construct a linear poor man's version of the ℓ^p-minimal solution to an

[2] We will see later that it actually minimizes many norms.

underdetermined set of linear equations $\mathbf{y} = \mathbf{A}\mathbf{x}$. For a general p, the solution to

$$\widehat{\mathbf{x}} \stackrel{\text{def}}{=} \arg\min \|\mathbf{x}\|_p \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{y}, \tag{3}$$

cannot be obtained by any linear operator $\mathbf{B}$ (see Section 2.4). That is, there is no $\mathbf{B}$ such that $\mathbf{z} \stackrel{\text{def}}{=} \mathbf{B}\mathbf{y}$ satisfies $\widehat{\mathbf{x}} = \mathbf{z}$ for every choice of $\mathbf{y}$. The exception is $p = 2$ for which the MPP does provide the minimum ℓ^2 norm representation $\mathbf{A}^\dagger\mathbf{y}$ of $\mathbf{y}$; Proposition 2.1 and comments thereafter show that this is indeed the only exception. On the other hand, we can obtain the following bound, valid for any $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{y}$, and in particular for $\widehat{\mathbf{x}}$:

$$\|\mathbf{z}\|_p = \|\mathbf{B}\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{B}\mathbf{A}\|_{\ell^p \to \ell^p} \|\mathbf{x}\|_p, \tag{4}$$

where $\|\cdot\|_{\ell^p \to \ell^p}$ is the operator norm on matrices induced by the ℓ^p norm on vectors. If $\mathbf{A}\mathbf{B} = \mathbf{I}$, then $\mathbf{z} = \mathbf{B}\mathbf{y}$ provides an admissible representation $\mathbf{A}\mathbf{z} = \mathbf{y}$, and

$$\|\mathbf{z}\|_p \leq \|\mathbf{B}\mathbf{A}\|_{\ell^p \to \ell^p} \|\widehat{\mathbf{x}}\|_p. \tag{5}$$

This expression suggests that the best linear generalized inverse $\mathbf{B}$ in the sense of minimal worst case ℓ^p norm blow-up is the one that minimizes $\|\mathbf{B}\mathbf{A}\|_{\ell^p \to \ell^p}$, motivating the definition of

$$\mathsf{pginv}_{\ell^p \to \ell^p}(\mathbf{A}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{X}} \|\mathbf{X}\mathbf{A}\|_{\ell^p \to \ell^p} \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{I}. \tag{6}$$

Objectives. Both (i) and (ii) above are achieved by minimization of some matrix norm. The purpose of this paper is to investigate the properties of generalized inverses $\mathsf{ginv}(\cdot)$ and $\mathsf{pginv}(\cdot)$ defined using various norms by addressing the following questions:

1. Are there norm families that all lead to the same generalized inverse, thus facilitating computation?
2. Are there specific classes of matrices for which different norms lead to the same generalized inverse, potentially different from the MPP?
3. Can we quantify the stability of matrices that result from these optimizations? In particular, can we control the Frobenius norm of the sparse pseudoinverse $\mathsf{ginv}_1(\mathbf{A})$, and more generally of any $\mathsf{ginv}_p(\mathbf{A})$, $p \geq 1$, for some random class of $\mathbf{A}$? This is the topic of Part II.

1.1 Prior Art

Several recent papers in frame theory study alternative dual frames, or equivalently, generalized inverses.[3] These works concentrate on existence results and explicit constructions of sparse frames and sparse dual frames with prescribed spectra [CHKK, KKL12]. Krahmer, Kutyniok, and Lemvig [KKL12] establish sharp bounds on the sparsity of dual frames, showing that generically, for $\mathbf{A} \in \mathbb{C}^{m \times n}$, the sparsest dual has $mn - m^2$ zeros. Li, Liu, and Mi [LLM13] provide bounds on the sparsity of duals of Gabor frames which are better than generic bounds. They also introduce the idea of using ℓ^p minimization to compute these dual frames, and they show that under certain circumstances, the ℓ^p minimization yields the sparsest possible dual Gabor frame. Further examples of non-canonical dual Gabor frames are given by Perraudin *et al.*, who use convex optimization to derive dual Gabor frames with more favorable properties than the canonical one [PHSB14], particularly in terms of time-frequency localization.

[3] Generalized inverses of full-rank matrices.

Another use of generalized inverses other than the MPP is when we have some idea about the subspace we want the solution of the original inverse problem to live in. We can then apply the restricted inverse of Bott and Duffin [BD53], or its generalizations [MN70]. The authors in [SG10] show how to compute *approximate* MPP-like inverses with an additional constraint that the minimizer lives in a particular matrix subspace, and how to use such matrices to precondition linear systems.

An important use of frames is in channel coding where the need for robust reconstruction in presence of channel errors leads to the design of optimal dual frames [LHH, LH10]. Similarly, one can try to compute the best generalized inverse for reconstruction from quantized measurements [LPY08]. This is related to the concept of Sobolev dual frames which minimize matrix versions of ℓ^2-type Sobolev norms [BLPY09] and which admit closed-form solutions.[4] The authors in [BLPY09] show that these alternative duals give a linear reconstruction scheme for $\Sigma\Delta$ quantization with a lower asymptotic reconstruction error than the cannonical dual frame (the MPP). Sobolev dual frames have also been used in compressed sensing with quantized measurements [GLP+12]. Some related ideas go back to the Wexler-Raz identity and its role in norm-minimizing dual functions [DLL94].

A major role in the theory of generalized inverses and matrix norms is played by unitarily invariant norms, studied in depth by Mirsky [Mir60]. Many results on the connection between these norms and the MPP are given by Ziętak [Zie97]; we comment on these connections in detail in Section 4. In their detailed account of generalized inverses [BIG03], Ben-Israel and Greville use the expression *minimal properties of generalized inverses*, but they primarily concentrate on variations of the square-norm minimality. Additionally, they define a class of non-linear generalized inverses corresponding to various metric projections. We are primarily concerned with generalized inverses that are themselves matrices, but one can imagine various *decoding rules* that search for a vector satisfying a model, and being consistent with the measurements [BDP+14]. In general, such decoding rules are not linear.

Finally, sparse pseudoinverse was previously studied in [DKV13], where it was shown empirically that the minimizer is indeed a sparse matrix, and that it can be used to speed up the resolution of certain inverse problems.

1.2 Our Contributions and Paper Outline

We study the properties of generalized inverses corresponding to norms[5] listed in Table 1, placed either on the candidate inverse $\mathbf{X}$ itself ($\mathsf{ginv}(\cdot)$) or on the projection $\mathbf{XA}$ ($\mathsf{pginv}(\cdot)$).

A number of relevant definitions and theoretical results are laid out in Section 2. In Section 3 we put forward some preliminary results on norm equivalences with respect to norm-minimizing generalized inverses. We also talk about poor man's ℓ^p minimization, by discussing generalized inverses that minimize the worst case and the average case ℓ^p blowup. These inverses generally do not coincide with the MPP. We then consider a property of some random matrix ensembles with respect to norms that do not lead to the MPP, and show that they satisfy what we call the *unbiasedness property*.

Section 4 discusses classes of norms that lead to the MPP. We extend the results of Ziętak on unitarily invariant norms to left-unitarily invariant norms which is relevant when minimizing the norm of the projection operator $\mathbf{XA}$, which is in turn relevant for *poor man's ℓ^p minimization* (Section 3.5). We conclude Section 4 by a discussion of norms that almost never yield the MPP. We prove that most mixed norms (column-wise and row-wise) almost never lead to the MPP. A particular representative of these norms is the entrywise ℓ^1 norm giving the *sparse pseudoinverse*.

[4] Our *poor man's ℓ^p minimization* in Section 3 has a similar formulation but without a closed-form solution.

[5] Strictly speaking, we also consider *quasi*-norms (typically for $0 < p, q < 1$)

Two fundamental questions about sparse pseudoinverses—those of uniqueness and stability—are discussed in Part II.

While Section 4 discusses norms, in Section 5 we concentrate on matrices. We have seen that many norms yield the MPP for all possible input matrices and that some norms generically do not yield the MPP. In Section 5 we first discuss a class of matrices for which some of those latter norms in fact *do* yield the MPP. It turns out that for certain matrices whose MPP has "flat" columns (*cf.* Theorem 5.1), $\mathsf{ginv}_\nu(\mathbf{A})$ contains the MPP for a large class of mixed norms ν, also those that normally do not yield the MPP. This holds in particular for partial Fourier and Hadamard matrices. Next, we exhibit a class of matrices for which many generalized inverses coincide, but *not* with the MPP.

Finally, in Section 6 we discuss how to efficiently compute many of the mentioned pseudoinverses. We observe that in some cases the computation simplifies to a vector problem, while in other cases it is indeed a full matrix problem. We use the alternating-direction method of multipliers (ADMM) [PB14] to compute the generalized inverse, as it can conveniently address both the norms on $\mathbf{X}$ and on $\mathbf{XA}$.

1.3 Summary and Visualization of Matrix Norms

We conclude the introduction by summarizing some of the results and norms in Table 1 and using the *matrix norm cube* in Figure 1—a visualization gadget we came up with for this paper. We place special emphasis on MPP-related results.

The norm cube is an effort to capture the various equivalences between matrix norms for particular choices of parameters. Each point on the displayed planes corresponds to a matrix norm. More details on these equivalences are given in Section 3. For example, using notations in Table 1, the Schatten 2-norm equals the Frobenius norm as well as the entrywise 2-norm. The induced $\ell^2 \to \ell^2$ norm equals the Schatten ∞-norm, while the induced $\ell^1 \to \ell^p$ norm equals the largest column p-norm, that is to say the $|p, \infty|$ columnwise mixed norm.

For many matrix norms ν, we prove (*cf.* Corollary 4.2) that $\mathsf{ginv}_\nu(\mathbf{A})$ and $\mathsf{pginv}_\nu(\mathbf{A})$ always contain the MPP. For other norms (*cf.* Theorem 4.3) we prove the existence of matrices $\mathbf{A}$ such that $\mathsf{ginv}_\nu(\mathbf{A})$ (resp. $\mathsf{pginv}_\nu(\mathbf{A})$) does not contain the MPP. This is the case for poor man's ℓ^1 minimization, both in its worst case flavor $\mathsf{pginv}_{\ell^1\to\ell^1}(\mathbf{A}) = \mathsf{pginv}_{|1,\infty|}(\mathbf{A})$ and in an average case version $\mathsf{pginv}_{\overline{2,1}}(\mathbf{A})$, *cf.* Section 3.5 and Proposition 3.1.

A number of questions remain open. Perhaps the main group is to induced norms for $q_\mathrm{o} \neq 2$ and their relation to the MPP.

Norm name	**Symbol**	**Definition**	$\mathbf{A}^\dagger \in \mathsf{ginv}_\nu(\mathbf{A})$	$\mathbf{A}^\dagger \in \mathsf{pginv}_\nu(\mathbf{A})$	[Proof]
Schatten	$\lVert\mathbf{M}\rVert_{S_p}$	$\lVert\mathrm{spec}(\mathbf{M})\rVert_p$	✓ $1 \le p \le \infty$	✓ $1 \le p \le \infty$	[Cor 4.2-1]
Columnwise	$\lVert\mathbf{M}\rVert_{\lvert p,q\rvert}$	$\left\lVert\{\lVert\mathbf{m}_j\rVert_p\}_{j=1}^n\right\rVert_q$	✓ $p = 2$, $1 \le q \le \infty$	✓ $p = 2$, $1 \le q \le \infty$	[Cor 4.2-2]
mixed norm			✗ $p \neq 2$, $0 < p, q < \infty$	✗ $p \neq 2$, $1 \le p < \infty$, $0 < q < \infty$	[Thm 4.3-1&3]
Entrywise	$\lVert\mathbf{M}\rVert_p$	$\lVert\mathrm{vec}(\mathbf{M})\rVert_p$	✓ $p = 2$	✓ $p = 2$	[Cor 4.2-2]
			✗ $p \neq 2$, $0 < p < \infty$	✗ $p \neq 2$, $0 < p < \infty$	[Thm 4.3-1&3]
Rowwise	$\lVert\mathbf{M}\rVert_{\overline{p,q}}$	$\left\lVert\{\lVert\mathbf{m}^i\rVert_p\}_{i=1}^m\right\rVert_q$	✓ $(p, q) = (2, 2)$	✓ $(p, q) = (2, 2)$	[Cor 4.2-2]
mixed norm			✗ $(p, q) \neq (2, 2)$, $0 < p, q < \infty$	✗ $(p, q) \neq (2, 2)$, $1 \le p < \infty$, $0 < q < \infty$	[Thm 4.3-2&4]
Induced	$\lVert\mathbf{M}\rVert_{\ell^p\to\ell^q}$	$\sup_{\mathbf{x}\neq 0} \frac{\lVert\mathbf{Mx}\rVert_q}{\lVert\mathbf{x}\rVert_p}$	✓ $1 \le p \le \infty$, $q = 2$	✓ $1 \le p \le \infty$, $q = 2$	[Cor 4.2-3]

Table 1: Summary of matrix norms considered in this paper. Notations are mostly introduced in Section 2. Checkmarks (✓) indicate parameter values for which the property is true for any full-rank *real or complex* matrix $\mathbf{A}$ Crosses (✗) indicate parameter values for which for any $m < n$, $m \ge 3$, there exists a full-rank $\mathbf{A} \in \mathbb{R}^{m\times n}$ such that $\mathbf{A}^\dagger \notin \mathsf{ginv}_\nu(\mathbf{A})$ (resp. $\mathbf{A}^\dagger \notin \mathsf{pginv}_\nu(\mathbf{A})$).

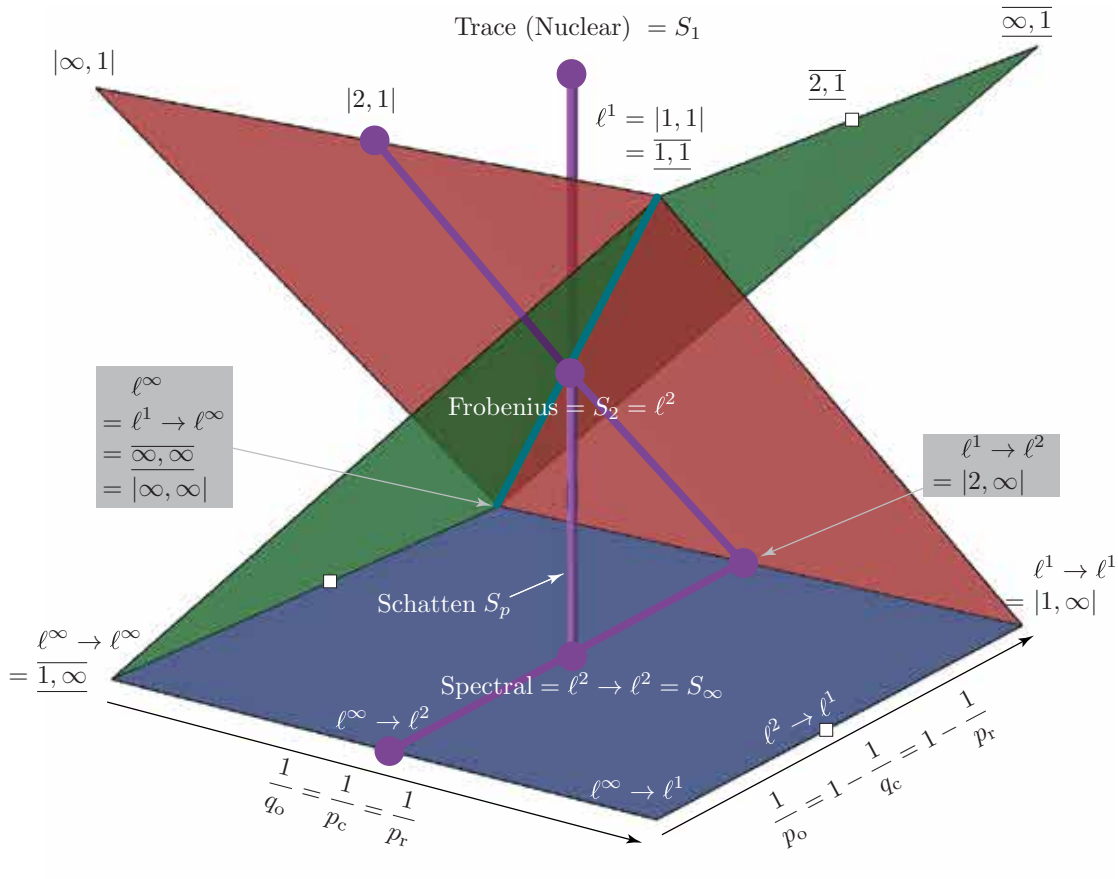

Figure 1: The blue plane is that of operator norms $\ell^{p_o} \to \ell^{q_o}$, the red one of columnwise mixed norms $|p_c, q_c|$, and the green one of rowwise mixed norms $\overline{p_r, q_r}$. The intersection of rowwise and columnwise mixed norms is shown by the thick blue line—these are entrywise ℓ^p norms. The vertical gray line is that of Schatten norms, with the nuclear norm S_1 at the top, the Frobenius norm S_2 in the middle (intersecting entrywise norms), and the spectral norm S_∞ at the bottom (intersecting operator norms). Among columnwise mixed norms $|p_c, q_c|$, all norms with a fixed value of p_c lead to the same minimizer. Purple circles and lines indicate norms ν for which $\mathsf{ginv}_\nu(\mathbf{A})$ and $\mathsf{pginv}_\nu(\mathbf{A})$ contain the MPP.

2 Definitions and Known Results

Throughout the paper we assume that all vectors and matrices are over $\mathbb{C}$ and point out when a result is valid only over the reals. Vectors are all column vectors, and they are denoted by bold lowercase letters, like $\mathbf{x}$. Matrices are denoted by bold uppercase letters, such as $\mathbf{M}$. By $\mathbf{M} \in \mathbb{C}^{m \times n}$ we mean that the matrix $\mathbf{M}$ has m rows and n columns of complex entries. The notation $\mathbf{I}_m$ stands for the identity matrix in $\mathbb{C}^{m \times m}$; the subscript m will often be omitted. We write $\mathbf{m}_j$ for the jth column of $\mathbf{M}$, and $\mathbf{m}^i$ for its ith row. The conjugate transpose of $\mathbf{M}$ is denoted $\mathbf{M}^*$ and the transpose is denoted $\mathbf{M}^\top$. The notation $\mathbf{e}_i$ denotes the ith canonical basis vector. Inner products are denoted by $\langle \cdot, \cdot \rangle$. All inner product are over complex spaces, unless otherwise is indicated in the subscript. For example, an inner product over real $m \times n$ matrices

will be written $\langle \cdot, \cdot \rangle_{\mathbb{R}^{m\times n}}$

2.1 Generalized Inverses

A generalized inverse of a rectangular matrix is a matrix that has some, but not all properties of the standard inverse of an invertible square matrix. It can be defined for non-square matrices that are not necessarily of full rank.

Definition 2.1 (Generalized inverse)**.** $\mathbf{X} \in \mathbb{C}^{n\times m}$ *is a generalized inverse of a matrix* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *if it satisfies* $\mathbf{AXA} = \mathbf{A}$.

We denote by $\mathcal{G}(\mathbf{A})$ the set of all generalized inverses of a matrix $\mathbf{A}$.

For the sake of clarity we will primarily concentrate on inverses of underdetermined matrices ($m < n$). As we show in Section 3.6, this choice does not incur a loss of generality. Furthermore, we will assume that the matrix has full rank: $\operatorname{rank}(\mathbf{A}) = m$. In this case, $\mathbf{X}$ is the generalized (right) inverse of $\mathbf{A}$ if and only if $\mathbf{AX} = \mathbf{I}_m$.

2.2 Correspondence Between Generalized Inverses and Dual Frames

Definition 2.2. *A collection of vectors* $(\boldsymbol{\phi}_i)_{i=1}^n$ *is called a (finite)* frame *for* $\mathbb{C}^m$ *if there exist constants* A *and* B, $0 < A \leq B < \infty$, *such that*

$$A \|\mathbf{x}\|_2^2 \leq \sum_{i=1}^n |\langle \mathbf{x}, \boldsymbol{\phi}_i \rangle|^2 \leq B \|\mathbf{x}\|_2^2 , \tag{7}$$

for all $\mathbf{x} \in \mathbb{C}^m$.

Definition 2.3. *A frame* $(\boldsymbol{\psi}_i)_{i=1}^n$ *is a* dual frame *to* $(\boldsymbol{\phi}_i)_{i=1}^n$ *if the following holds for any* $\mathbf{x} \in \mathbb{C}^m$,

$$\mathbf{x} = \sum_{i=1}^n \langle \mathbf{x}, \boldsymbol{\phi}_i \rangle \boldsymbol{\psi}_i. \tag{8}$$

This can be rewritten in matrix form as

$$\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\Phi}^*\mathbf{x}. \tag{9}$$

As this must hold for all $\mathbf{x}$, we can conclude that

$$\boldsymbol{\Psi}\boldsymbol{\Phi}^* = \mathbf{I}_m, \tag{10}$$

and so any dual frame $\boldsymbol{\Psi}$ of $\boldsymbol{\Phi}$ is a generalized left inverse of $\boldsymbol{\Phi}^*$. Thus there is a one-to-one correspondence between dual frames and generalized inverses of full rank matrices.

2.3 Characterization with the Singular Value Decomposition (SVD)

A particularly useful characterization of generalized inverses is through the singular value decomposition (SVD). This characterization has been used extensively to prove theorems in [KKL12, Zie97] and elsewhere. Consider the SVD of the matrix $\mathbf{A}$

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*, \tag{11}$$

where $\mathbf{U} \in \mathbb{C}^{m\times m}$ and $\mathbf{V} \in \mathbb{C}^{n\times n}$ are unitary and $\mathbf{\Sigma} = \left[\mathsf{diag}(\sigma_1(\mathbf{A}), \ldots, \sigma_m(\mathbf{A})), \mathbf{0}_{m\times(n-m)}\right]$ contains the singular values of $\mathbf{A}$ in a non-increasing order. For a matrix $\mathbf{X}$, let $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{V}^*\mathbf{X}\mathbf{U}$. Then it follows from Definition 2.1 that $\mathbf{X}$ is a generalized inverse of $\mathbf{A}$ if and only if

$$\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma} = \mathbf{\Sigma}. \tag{12}$$

Denoting by r the rank of $\mathbf{A}$ and setting

$$\mathbf{\Sigma}_\square = \mathsf{diag}(\sigma_1(\mathbf{A}), \ldots, \sigma_r(\mathbf{A})), \tag{13}$$

we deduce that $\mathbf{M}$ must be of the form

$$\mathbf{M} = \begin{bmatrix} \mathbf{\Sigma}_\square^{-1} & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{bmatrix} \tag{14}$$

where $\mathbf{R} \in \mathbb{C}^{r\times(m-r)}$, $\mathbf{S} \in \mathbb{C}^{(n-r)\times r}$, and $\mathbf{T} \in \mathbb{C}^{(n-r)\times(m-r)}$ are arbitrary matrices.

For a full-rank $\mathbf{A}$, (14) simplifies to

$$\mathbf{M} = \begin{bmatrix} \mathbf{\Sigma}_\square^{-1} \\ \mathbf{S} \end{bmatrix}, \tag{15}$$

In the rest of the paper we restrict ourselves to full-rank matrices, and use the following characterization of the set of all generalized inverses of a matrix $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$:

$$\mathcal{G}(\mathbf{A}) = \{\mathbf{X} \ : \ \mathbf{X} = \mathbf{V}\mathbf{M}\mathbf{U}^* \text{ where } \mathbf{M} \text{ has the form (15)}\}. \tag{16}$$

For rank-deficient matrices, the same holds with (14) instead of (15). Using this alternative characterization to extend the main results of this paper to rank-deficient matrices is left to future work.

2.4 The Moore-Penrose Pseudoinverse (MPP)

The Moore-Penrose Pseudoinverse (MPP) has a special place among generalized inverses, thanks to its various optimality and symmetry properties.

Definition 2.4 (MPP)**.** *The Moore-Penrose pseudoinverse of the matrix* $\mathbf{A}$ *is the unique matrix* $\mathbf{A}^\dagger$ *such that*

$$\begin{aligned} \mathbf{A}\mathbf{A}^\dagger\mathbf{A} &= \mathbf{A}, & (\mathbf{A}\mathbf{A}^\dagger)^* &= \mathbf{A}\mathbf{A}^\dagger, \\ \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger &= \mathbf{A}^\dagger, & (\mathbf{A}^\dagger\mathbf{A})^* &= \mathbf{A}^\dagger\mathbf{A}. \end{aligned} \tag{17}$$

This definition is universal—it holds regardless of whether $\mathbf{A}$ is underdetermined or overdetermined, and regardless of whether it is full rank.

Under the conditions primarily considered in this paper ($m < n$, $\text{rank}(\mathbf{A}) = m$), we can express the MPP as $\mathbf{A}^\dagger = \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}$, which corresponds to the particular choice $\mathbf{S} = \mathbf{0}_{(n-m)\times m}$ in (15). The canonical dual frame $\mathbf{\Phi}$ of a frame $\mathbf{\Psi}$ is the adjoint $\mathbf{\Phi} = [\mathbf{\Phi}^\dagger]^*$ of its MPP.

There are several alternative definitions of MPP. One that pertains to our work is:

Definition 2.5. *MPP is the unique generalized inverse of* $\mathbf{A}$ *with minimal Frobenius norm.*

That this definition makes sense will be clear from the next section. As we will see in Section 4, the MPP can also be characterized as the generalized inverse minimizing many other matrix norms.

MPP has a number of interesting properties. If $\mathbf{A} \in \mathbb{C}^{m\times n}$, with $m > n$, and

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{18}$$

we can compute

$$\widehat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y}. \tag{19}$$

This vector $\widehat{\mathbf{x}}$ is what would in the noiseless case generate $\widehat{\mathbf{y}} = \mathbf{A}\mathbf{A}^\dagger\mathbf{y}$—the orthogonal projection of $\mathbf{y}$ onto the range of $\mathbf{A}$, $\mathcal{R}(\mathbf{A})$. This is also known as the least-squares solution to an inconsistent overdetermined system of linear equations, in the sense that it minimizes the sum of squared residuals over all equations. For uncorrelated, zero-mean errors of equal variance, this gives the best linear unbiased estimator (BLUE) of $\mathbf{x}$.

Note that the optimal solution to (18) in the sense of the minimum mean-squared error (MMSE) (when $\mathbf{e}$ is considered random) is not given by the MPP, but rather as the Wiener filter [Kay98],

$$\mathbf{B}_{\mathsf{MMSE}} = \mathbf{C}_x\mathbf{A}^*(\mathbf{A}\mathbf{C}_x\mathbf{A}^* + \mathbf{C}_n)^{-1}\,, \tag{20}$$

where $\mathbf{C}_x$ and $\mathbf{C}_n$ are signal and noise covariance matrices.[6] MPP for fat matrices with full row rank is a special case of this formula for $\mathbf{C}_n = \mathbf{0}$ and $\mathbf{C}_x = \mathbf{I}$.

In the underdetermined case, $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m < n$, applying the MPP yields the solution with the smallest ℓ^2 norm among all vectors $\mathbf{x}$ satisfying $\mathbf{y} = \mathbf{A}\mathbf{x}$ (among all *admissible* $\mathbf{x}$). That is,

$$\left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{x}\right\|_2 \le \|\mathbf{z}\|_2\,, \quad \forall\, \mathbf{z}\ s.t.\ \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}. \tag{21}$$

To see this, we use the orthogonality of $\mathbf{A}^\dagger\mathbf{A}$. Note that any vector $\mathbf{x}$ can be decomposed as

$$\mathbf{A}^\dagger\mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}, \tag{22}$$

and that

$$\begin{aligned}
\langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}\rangle &= \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{A}^\dagger\mathbf{A}\mathbf{x}\rangle \\
&= \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, (\mathbf{A}^\dagger\mathbf{A})^*\mathbf{x}\rangle \\
&= \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{A}^\dagger\mathbf{A}\mathbf{x}\rangle \\
&= \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{A}^\dagger\mathbf{A}\mathbf{x}, \mathbf{x}\rangle \\
&= 0,
\end{aligned} \tag{23}$$

where we applied Definition 2.4 twice. Thus $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$ is orthogonal to $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}$, and we have

$$\begin{aligned}
\|\mathbf{z}\|_2^2 &= \left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{z} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{z}\right\|_2^2 \\
&= \left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{z}\right\|_2^2 + \left\|(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{z}\right\|_2^2 \\
&\ge \left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{z}\right\|_2^2 = \left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{x}\right\|_2^2.
\end{aligned} \tag{24}$$

A natural question to ask is if there are other MPP-like linear generalized inverses for ℓ^p norms with $p \neq 2$. The answer is negative:

[6] Assuming $\mathbf{A}\mathbf{C}_x\mathbf{A}^* + \mathbf{C}_n$ is invertible.

Proposition 2.1 (Corollary 5, [NO69]). *Let $m \geq 3$ and $n > m$. For $1 < p < \infty$ define $B_{\mathbf{A}} : \mathbb{C}^m \to \mathbb{C}^n$ as*

$$B_{\mathbf{A}}(\mathbf{y}) \stackrel{\text{def}}{=} \underset{\mathbf{z}\in\mathbb{C}^n:\mathbf{A}\mathbf{z}=\mathbf{y}}{\arg\min} \|\mathbf{z}\|_p ,$$

where the minimizer is unique by the strict convexity of $\|\cdot\|_p$. Then $B_{\mathbf{A}}(\mathbf{y})$ is linear for all $\mathbf{A}$ if and only if $p = q = 2$.

3 Generalized Inverses Minimizing Matrix Norms

Despite the negative result in Proposition 2.1 saying that the MPP is in some sense an exception, an interesting way of generating different generalized inverses is by norm[7] minimization. Two central definitions of such generalized inverses will be used in this paper. The generalized inverse of $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m < n$, with minimal ν-norm is defined as ($\|\cdot\|_\nu$ is an arbitrary matrix norm or quasi-norm)

$$\mathsf{ginv}_\nu(\mathbf{A}) \stackrel{\text{def}}{=} \underset{\mathbf{X}}{\arg\min}\ \|\mathbf{X}\|_\nu \quad \text{subject to} \quad \mathbf{X} \in \mathcal{G}(\mathbf{A}).$$

The generalized inverse minimizing the μ-norm of the product $\mathbf{XA}$ is defined as

$$\mathsf{pginv}_\mu(\mathbf{A}) \stackrel{\text{def}}{=} \underset{\mathbf{X}}{\arg\min}\ \|\mathbf{XA}\|_\mu \quad \text{subject to} \quad \mathbf{X} \in \mathcal{G}(\mathbf{A}).$$

This definition, which is a particular case of the first one with $\|\cdot\|_\nu = \|\cdot\mathbf{A}\|_\mu$, will serve when considering $\mathbf{X}$ as a poor man's linear replacement for ℓ^p minimization. Strictly speaking, the above-defined pseudoinverses are sets, since the corresponding programs may have more than one solution. We will point out the cases when special care must be taken. Another important point is that both definitions involve convex programs. So, at least in principle, we can find the optimizer in the sense that any first-order scheme will lead to the global optimum.

We will treat several families of matrix norms. A matrix norm is any norm on $\mathbb{C}^{m\times n}$.

3.1 Entrywise norms

The simplest matrix norm is the entrywise ℓ^p norm. It is defined through an isomorphism between $\mathbb{C}^{m\times n}$ and $\mathbb{C}^{mn}$, that is, it is simply the ℓ^p norm of the vector of concatenated columns.

Definition 3.1. *The p-entrywise norm of $\mathbf{M} \in \mathbb{C}^{m\times n}$, where $0 \leq p \leq \infty$, is given as*

$$\|\mathbf{M}\|_p \stackrel{\text{def}}{=} \|\mathrm{vec}(\mathbf{M})\|_p . \tag{25}$$

A particular entrywise norm is the Frobenius norm associated to $p = 2$.

3.2 Induced norms—*poor man's ℓ^p minimization*

An important class is that of induced norms. To define these norms, we consider $\mathbf{M} \in \mathbb{C}^{m\times n}$ as an operator mapping vectors from $\mathbb{C}^n$ (equipped with an ℓ^p norm) to $\mathbb{C}^m$ (equipped with an ℓ^q norm).

[7] For brevity, we loosely call "norm" any quasi-norm such as ℓ^p, $p < 1$, as well as the "pseudo-norm" ℓ^0.

Definition 3.2. *The $\ell^p \to \ell^q$ induced norm of $\mathbf{M} \in \mathbb{C}^{m\times n}$, where $0 < p, q \leq \infty$ is*

$$\|\mathbf{M}\|_{\ell^p\to\ell^q} \stackrel{\text{def}}{=} \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{M}\mathbf{x}\|_q}{\|\mathbf{x}\|_p}. \tag{26}$$

It is straightforward to show that this definition is equivalent to $\|\mathbf{M}\|_{\ell^p\to\ell^q} = \sup_{\|\mathbf{x}\|_p=1} \|\mathbf{M}\mathbf{x}\|_q$. Note that while this is usually defined only for proper norms (i.e., with $1 \leq p, q \leq \infty$) the definition remains valid when $0 < p < 1$ and/or $0 < q < 1$.

3.3 Mixed norms (columnwise and rowwise)

An interesting case mentioned in the introduction is the $\ell^1 \to \ell^1$ induced norm of $\mathbf{XA}$, as it leads to a sort of optimal poor man's ℓ^1 minimization. The $\ell^1 \to \ell^1$ induced norm is a special case of the family of $\ell^1 \to \ell^q$ induced norms, which can be shown to have a simple expression as columnwise mixed norm

$$\|\mathbf{M}\|_{\ell^1\to\ell^q} = \max_{1\leq j\leq n} \|\mathbf{m}_j\|_q \stackrel{\text{def}}{=} \|\mathbf{M}\|_{|q,\infty|}. \tag{27}$$

More generally, one can consider columnwise mixed norms for any p and q:

Definition 3.3. *The columnwise mixed norm $\|\mathbf{M}\|_{|p,q|}$ is defined as*

$$\|\mathbf{M}\|_{|p,q|} \stackrel{\text{def}}{=} \left(\sum_j \|\mathbf{m}_j\|_p^q\right)^{1/q} \tag{28}$$

with the usual modification for $q = \infty$.

We deal both with column- and row-wise norms, so we introduce a mnemonic notation to easily tell them apart. Thus $\|\cdot\|_{|p,q|}$ denotes columnwise mixed norms, and $\|\cdot\|_{\underline{p,q}}$ denotes rowwise mixed norms, defined as follows.

Definition 3.4. *The rowwise mixed norm $\|\mathbf{M}\|_{\underline{p,q}}$ is defined as*

$$\|\mathbf{M}\|_{\underline{p,q}} \stackrel{\text{def}}{=} \left(\sum_i \|\mathbf{m}^j\|_p^q\right)^{1/q} \tag{29}$$

with the usual modification for $q = \infty$.

3.4 Schatten norms

Another classical norm is the spectral norm, which is the $\ell^2 \to \ell^2$ induced norm. It equals the maximum singular value of $\mathbf{M}$, so it is also a special case of Schatten norm, just as the Frobenius norm which is the ℓ^2 norm of the vector of singular values of $\mathbf{M}$. We can also define a general Schatten norm $\|\text{spec}(\mathbf{M})\|_p$ where $\text{spec}(\mathbf{M})$ is the vector of singular values.

Definition 3.5. *The Schatten norm $\|\mathbf{M}\|_{S^p}$ is defined as*

$$\|\mathbf{M}\|_{S^p} \stackrel{\text{def}}{=} \|\text{spec}(\mathbf{M})\|_p, \tag{30}$$

where $\mathrm{spec}(\mathbf{M})$ *is the vector of singular values.*

As we will see further on, these are special cases of the larger class of unitarily invariant matrix norms.

3.5 Poor man's ℓ^p minimization revisited

We conclude the overview of matrix norms by introducing certain norms based on a probabilistic signal model. Similarly to induced norms on the projection operator $\mathbf{XA}$, these norms lead to optimal ℓ^p-norm blow-up that can be achieved by a linear operator. Given that they are computed on the projection operator, they are primarily useful when considering $\mathsf{pginv}(\cdot)$, not $\mathsf{ginv}(\cdot)$.

We already pointed out in the introduction that the generalized inverse $\mathsf{pginv}_{\ell^p\to\ell^p}(\mathbf{A})$ is the one which minimizes the worst case blowup of the ℓ^p norm between $\mathbf{z}$, the minimum ℓ^p norm vector such that $\mathbf{y} = \mathbf{Az}$, and the linear estimate $\mathbf{Xy}$ where $\mathbf{X} \in \mathcal{G}(\mathbf{A})$. In this sense, $\mathsf{pginv}_{\ell^p\to\ell^p}(\mathbf{A})$ provides the best worst-case poor man's (linear) ℓ^p minimization, and solves

$$\inf_{\mathbf{X}\in\mathcal{G}(\mathbf{A})} \sup_{\mathbf{y}\neq 0} \frac{\|\mathbf{Xy}\|_p}{\inf_{\mathbf{z}:\mathbf{Az}=\mathbf{y}} \|\mathbf{z}\|_p}. \tag{31}$$

In this expression we can see explicitly the *true* ℓ^p optimal solution in the denominator of the argument of the supremum.

Instead of minimizing the worst-case blowup, we may want to minimize average-case ℓ^p blowup over a given class of input vectors. Let $\mathbf{u}$ be a random vector with probability distribution given by $\mathbb{P}_{\mathbf{u}}$. Given $\mathbf{A}$, our goal is to minimize $\mathbb{E}_{\mathbf{u}\sim\mathbb{P}_{\mathbf{u}}}[\|\mathbf{XAu}\|_p]$. We replace this minimization by a simpler proxy: we minimize $(\mathbb{E}_{\mathbf{u}\sim\mathbb{P}_{\mathbf{u}}} \|\mathbf{XAu}\|_p^p)^{\frac{1}{p}}$. It is not difficult to verify that this expectation defines a norm (or a semi-norm, depending on $\mathbb{P}_{\mathbf{u}}$).

Interestingly, for certain distributions $\mathbb{P}_{\mathbf{u}}$ this leads back to minimization of standard matrix norms:

Proposition 3.1. *Assume that* $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. *Then we have*

$$\arg\min_{\mathbf{X}\in\mathcal{G}(\mathbf{A})} \left(\mathbb{E}_{\mathbf{u}}[\|\mathbf{XAu}\|_p^p]\right)^{1/p} = \mathsf{pginv}_{\overline{2,p}}(\mathbf{A}) \tag{32}$$

Remark 3.1. *This result is intuitively satisfying. It is known [JAB11] that the* $\overline{2,1}$ *mixed norm promotes row sparsity, thus the resulting* $\mathbf{X}$ *will have rows set to zero. Therefore, even if the result may have been predicted, it is interesting to see how a generic requirement to have a small* ℓ^1 *norm of the output leads to a known group sparsity penalty on the product matrix.*

Remark 3.2. *For a general* p *the minimization* (32) *only minimizes an expected* proxy *of the output* ℓ^p *norm, but for* $p = 1$ *we get exactly the expected* ℓ^1 *norm.*

Proof.

$$\mathbb{E}_{\mathbf{u}}\left[\, \|\mathbf{XAu}\|_p^p \,\right] = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{u}}\left[\, |(\mathbf{XAu})_i|^p \,\right] \tag{33}$$

Because $\mathbf{u}$ is centered normal with covariance $\mathbf{I}_n$, the covariance matrix of $\mathbf{XAu}$ is $\mathbf{K} = (\mathbf{XA})(\mathbf{XA})^*$. Individual components are distributed according to $(\mathbf{XAu})_i \sim \mathcal{N}(0, \mathbf{K}_{ii}) =$

$\mathcal{N}(0, \left\|\mathbf{x}^i\mathbf{A}\right\|_2^2)$. A straightforward computation shows that

$$\mathbb{E}\left[\, |(\mathbf{XAu})_i|^p \,\right] = \frac{2^{p/2}\,\Gamma\left(\frac{1+p}{2}\right)}{\sqrt{\pi}} \left\|\mathbf{x}^i\mathbf{A}\right\|_2^p . \tag{34}$$

We can then continue writing

$$\mathbb{E}\left[\, \|\mathbf{XAu}\|_p^p \,\right] = \sum_{i=1}^{n} \left\|\mathbf{x}^i\mathbf{A}\right\|_2^p = \|\mathbf{XA}\|_{\underline{2,p}} , \tag{35}$$

and the claim follows. □

3.6 Fat and skinny matrices

In this paper, we concentrate on generalized inverses of fat matrices—matrices with more columns than rows. We first want to show that there is no loss of generality in making this choice. This is clear for minimizing mixed norms and Schatten norms, as for mixed norms we have that

$$\|\mathbf{M}\|_{|p,q|} = \|\mathbf{M}^*\|_{\underline{p,q}} , \tag{36}$$

and for Schatten norm we have

$$\|\mathbf{M}\|_{S_p} = \|\mathbf{M}^*\|_{S_p} . \tag{37}$$

It only remains to be shown for induced norms. We can state the following lemma:

Lemma 3.1. *Let $1 \le p, q, p^*, q^* \le \infty$ with $\frac{1}{p} + \frac{1}{p^*} = 1$ and $\frac{1}{q} + \frac{1}{q^*} = 1$. Then we have the relation $\|\mathbf{M}\|_{\ell^p \to \ell^q} = \|\mathbf{M}^*\|_{\ell^{q^*} \to \ell^{p^*}}$.*

In other words, all norms we consider on tall matrices can be converted to norms on their fat transposes, and our results apply accordingly.

As a corollary we have

$$\|\mathbf{M}\|_{\ell^p \to \ell^\infty} = \|\mathbf{M}^*\|_{\ell^1 \to \ell^{p^*}} = \|\mathbf{M}^*\|_{|p^*,\infty|} \stackrel{\text{def}}{=} \|\mathbf{M}\|_{\underline{p^*,\infty}} = \max_{1 \le i \le m} \left\|\mathbf{m}^i\right\|_{p^*} . \tag{38}$$

Next, we show that generalized inverses obtained by minimizing columnwise mixed norms always match minimizing an entrywise norm or an induced norm.

Lemma 3.2. *Consider $0 < p \le \infty$ and a full rank matrix $\mathbf{A}$.*

1. *For $0 < q < \infty$, we have the set equality $\mathsf{ginv}_{|p,q|}(\mathbf{A}) = \mathsf{ginv}_p(\mathbf{A})$.*
2. *For $q = \infty$ we have $\|\cdot\|_{|p,\infty|} = \|\cdot\|_{\ell^1 \to \ell^p}$ and the set inclusions / equalities*

$$\begin{aligned} \mathsf{ginv}_p(\mathbf{A}) \subset \mathsf{ginv}_{|p,\infty|}(\mathbf{A}) &= \mathsf{ginv}_{\ell^1 \to \ell^p}(\mathbf{A}) \\ \mathsf{pginv}_{|p,\infty|}(\mathbf{A}) &= \mathsf{pginv}_{\ell^1 \to \ell^p}(\mathbf{A}). \end{aligned}$$

Proof. For $q < \infty$, minimizing $\|\mathbf{X}\|_{|p,q|}$ under the constraint $\mathbf{X} \in \mathcal{G}(\mathbf{A})$ amounts to minimizing $\sum_j \|\mathbf{x}_j\|_p^q$ under the constraints $\mathbf{A}\mathbf{x}_j = \mathbf{e}_j$, where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$ and $\mathbf{e}_j$ the jth canonical vector. Equivalently, one can separately minimize $\|\mathbf{x}_j\|_p$ such that $\mathbf{A}\mathbf{x}_j = \mathbf{e}_j$. □

3.7 Unbiasedness of Generalized Inverses

Most of the discussion so far involved deterministic matrices and deterministic properties. In the previous subsections certain results were stated for matrices in general positions—a property which matrices from the various random ensembles verify with probability one. In this and the next section we discuss properties generalized inverses of some random matrices. We start by demonstrating a nice property of the MPP of random matrices replicated by other norm-minimizing generalized inverses, which is unbiasedness in a certain sense.

For a random Gaussian matrix $\mathbf{A}$ it holds that

$$\frac{n}{m}\mathbb{E}[\mathbf{A}^\dagger\mathbf{A}] = \mathbf{I}. \tag{39}$$

In other words, for this random matrix ensemble, applying $\mathbf{A}$ to a vector, and then the MPP to the measurements will on average retrieve the scaled version of the input vector. This aesthetically pleasing property is used to make statements about various iterative algorithms such as iterative hard thresholding. To motivate it we consider the following generic procedure: let $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{x}$ is the object we are interested in (e.g. an image), $\mathbf{A}$ a dimensionality-reducing measurement system, and $\mathbf{y}$ the resulting measurements. One admissible estimate of $\mathbf{x}$ is given by $\mathbf{X}\mathbf{A}\mathbf{x}$, where $\mathbf{X} \in \mathcal{G}(\mathbf{A})$. If the dimensionality-reducing system is random, we can compute the expectation of the reconstructed vector as

$$\mathbb{E}[\mathbf{X}\mathbf{y}] = \mathbb{E}[\mathbf{X}\mathbf{A}\mathbf{x}] = \mathbb{E}[\mathbf{X}\mathbf{A}]\mathbf{x}. \tag{40}$$

Provided that $\mathbb{E}[\mathbf{X}\mathbf{A}] = \frac{m}{n}\mathbf{I}$, we will obtain, on average, a scaled version of the object we wish to reconstruct.[8]

Clearly, this property will not hold for generalized inverses obtained by inverting a particular minor of the input matrix. As we show next, it does hold for a large class of norm-minimizing generalized inverses.

Theorem 3.1. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $m < n$ be a random matrix with iid columns such that $a_{ij} \sim (-a_{ij})$. Let further $\|\cdot\|_\nu$ be any matrix norm such that $\|\mathbf{\Pi}\cdot\|_\nu = \|\cdot\|_\nu$ and $\|\mathbf{\Sigma}\cdot\|_\nu = \|\cdot\|_\nu$, for any permutation matrix $\mathbf{\Pi}$ and modulation matrix $\mathbf{\Sigma} = \mathsf{diag}(\sigma)$, where $\sigma \in \{-1, 1\}^n$. Then, if $\mathsf{ginv}_\nu(\mathbf{A})$ and $\mathsf{pginv}_\nu(\mathbf{A})$ are singletons for all $\mathbf{A}$, we have*

$$\mathbb{E}[\mathsf{ginv}_\nu(\mathbf{A})\mathbf{A}] = \mathbb{E}[\mathsf{pginv}_\nu(\mathbf{A})\mathbf{A}] = \tfrac{m}{n}\mathbf{I}_n \tag{41}$$

More generally, consider a function $f : C \mapsto f(C) \in C \subset \mathbb{R}^{n\times m}$ that selects a particular representative for any bounded convex set C, and assume that $f(\mathbf{U}C) = \mathbf{U}f(C)$ for any unitary matrix $\mathbf{U}$ and any C. Examples of such functions f include selecting the centroid of the convex set, or selecting its element with minimum Frobenius norm. We have

$$\mathbb{E}[f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A}] = \mathbb{E}[f(\mathsf{pginv}_\nu(\mathbf{A}))\mathbf{A}] = \tfrac{m}{n}\mathbf{I}_n. \tag{42}$$

Remark 3.3. *This includes all classical norms (invariance to row permutations and sign changes) as well as any left-unitarily invariant norm (permutations and sign changes are unitary).*

To prove the theorem we use the following lemma,

[8] This linear step is usually part of a more complicated algorithm which also includes a nonlinear *denoising* step (*e.g., thresholding, non-local means*). If this denoising step is a contraction in some sense (*i.e.* it brings us closer to the object we are reconstructing), the following scheme will converge: $\mathbf{x}^{(k+1)} \stackrel{\text{def}}{=} \eta(\mathbf{x}^{(k)} + \frac{n}{m}\mathbf{X}(\mathbf{y} - \mathbf{A}\mathbf{x}^{(k)}))$.

Lemma 3.3. *Let $\mathbf{U} \in \mathbb{C}^{n\times n}$ be an invertible matrix, and $\|\cdot\|_\nu$ a norm such that $\|\mathbf{U}\cdot\|_\nu = \|\cdot\|_\nu$. Then the following claims hold,*

$$\mathsf{ginv}_\nu(\mathbf{AU}) = \mathbf{U}^{-1}\mathsf{ginv}_\nu(\mathbf{A}), \tag{43}$$

$$\mathsf{pginv}_\nu(\mathbf{AU}) = \mathbf{U}^{-1}\,\mathsf{pginv}_\nu(\mathbf{A}) \tag{44}$$

for any $\mathbf{A}$.

Proof of the lemma. We only prove the first claim; the remaining parts follow analogously using that $\mathsf{pginv}_\nu(\mathbf{A}) = \mathsf{ginv}_\mu(\mathbf{A})$ where $\|\cdot\|_\mu \stackrel{\text{def}}{=} \|\cdot\mathbf{A}\|_\nu = \|\mathbf{U}\cdot\mathbf{A}\|_\nu = \|\mathbf{U}\cdot\|_\mu$.
Feasibility: $(\mathbf{AU})(\mathbf{U}^{-1}\mathsf{ginv}_\nu(\mathbf{A})) = \mathbf{A}\mathsf{ginv}_\nu(\mathbf{A}) = \mathbf{I}_m$.
Optimality: Consider any $\mathbf{X} \in \mathcal{G}(\mathbf{AU})$. Since $(\mathbf{AU})\mathbf{X} = \mathbf{I}_m = \mathbf{A}(\mathbf{UX})$, the matrix $\mathbf{UX}$ belongs to $\mathcal{G}(\mathbf{A})$ hence

$$\|\mathbf{X}\|_\nu = \|\mathbf{UX}\|_\nu \geq \|\mathsf{ginv}_\nu(\mathbf{A})\|_\nu = \left\|\mathbf{UU}^{-1}\mathsf{ginv}_\nu(\mathbf{A})\right\|_\nu = \left\|\mathbf{U}^{-1}\mathsf{ginv}_\nu(\mathbf{A})\right\|_\nu. \tag{45}$$

□

Proof of the theorem. Since the matrix columns are iid, $\mathbf{A}$ is distributed identically to $\mathbf{A\Pi}$ for any permutation matrix $\mathbf{\Pi}$. This implies that functions of $\mathbf{A}$ and $\mathbf{A\Pi}$ have the same distribution. Thus the sets $f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A}$ and $f(\mathsf{ginv}_\nu(\mathbf{A\Pi}))\mathbf{A\Pi}$ are identically distributed. Using Lemma 3.3 with $\mathbf{U} = \mathbf{\Pi}$, we have that

$$\mathbf{M} \stackrel{\text{def}}{=} \mathbb{E}[f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A}] = \mathbb{E}[f(\mathsf{ginv}_\nu(\mathbf{A\Pi}))\mathbf{A\Pi}] \tag{46}$$

$$= \mathbb{E}[f(\mathbf{\Pi}^*\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A\Pi}] \tag{47}$$

$$= \mathbb{E}[\mathbf{\Pi}^* f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A\Pi}] = \mathbf{\Pi}^*\mathbf{M\Pi}. \tag{48}$$

This is more explicitly written $m_{ij} = m_{\pi(i)\pi(j)}$ for all i, j and π the permutation associated to the permutation matrix $\mathbf{\Pi}$. Since this holds for any permutation matrix, we can write

$$\mathbf{M} = \begin{bmatrix} c & b & \cdots & b \\ b & c & \cdots & b \\ \vdots & & \ddots & \vdots \\ b & b & \cdots & c \end{bmatrix}, \tag{49}$$

We compute the value of $c = \frac{m}{n}$ as follows:

$$nc = \mathsf{Tr}\ \mathbb{E}[f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A}] = \mathbb{E}[\mathsf{Tr}\ f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A}] = \mathbb{E}[\mathsf{Tr}\ \mathbf{A}\, f(\mathsf{ginv}_\nu(\mathbf{A}))] = \mathsf{Tr}\ \mathbf{I}_m = m. \tag{50}$$

To show that $b = 0$, we observe that since $a_{ij} \sim (-a_{ij})$, the matrices $\mathbf{A}$ and $\mathbf{A\Sigma}$ have the same distribution. As above, using Lemma 3.3 with $\mathbf{U} = \mathbf{\Sigma}$, this implies $\mathbf{M} = \mathbf{\Sigma M \Sigma}$ for any modulation matrix $\mathbf{\Sigma}$, that is to say $m_{ij} = \sigma_i m_{ij}\sigma_j$ for any i, j and $\sigma \in \{-1, 1\}^n$. It follows that $m_{ij} = 0$ for $i \neq j$. Since we already established that $c_i \equiv \frac{m}{n}$ we conclude that $\mathbb{E}[f(\mathsf{ginv}_\nu(\mathbf{A}))\mathbf{A}] = \frac{m}{n}\mathbf{I}_n$.

□

The conditions of Theorem 3.1 are satisfied by various random matrix ensembles including the common iid Gaussian ensemble.

One possible interpretation of this result is as follows: For the Moore-Penrose pseudoinverse $\mathbf{A}^\dagger$ of a fat matrix $\mathbf{A}$, we have that $\mathbf{A}^\dagger\mathbf{A}$ is an orthogonal projection. For a general pseudoinverse $\mathbf{X}$, $\mathbf{XA}$ is an oblique projection, along an angle different than $\frac{\pi}{2}$. Nevertheless, for many norm minimizing generalized inverses this angle is on average $\frac{\pi}{2}$, if the average is taken over common classes of random matrices.

4 Norms Yielding the Moore-Penrose Pseudoinverse

A particularly interesting property of the MPP is that it minimizes many of the norms in Table 1. This is related to their unitary invariance, and to geometric interpretation of the MPP [VKG14].

4.1 Unitarily invariant norms

Definition 4.1 (Unitarily invariant matrix norm)**.** *A matrix norm $\|\cdot\|$ is called unitarily invariant if and only if $\|\mathbf{UMV}\| = \|\mathbf{M}\|$ for any $\mathbf{M}$ and any unitary matrices $\mathbf{U}$ and $\mathbf{V}$.*

Unitarily invariant matrix norms are intimately related to symmetric gauge functions [Mir60], defined as vector norms invariant to sign changes and permutations of the vector entries. A theorem by Von Neumann [vN37, HJ12] states that any unitarily invariant norm $\|\cdot\|$ is a symmetric gauge function ϕ of the singular values, i.e., $\|\cdot\| = \phi(\sigma(\cdot)) \stackrel{\text{def}}{=} \|\cdot\|_\phi$. To be a symmetric gauge function, ϕ has to satisfy the following properties [Mir60]:

(i) $\phi(\mathbf{x}) \geq 0$ for $\mathbf{x} \neq 0$,

(ii) $\phi(\alpha\mathbf{x}) = |\alpha|\,\phi(\mathbf{x})$,

(iii) $\phi(\mathbf{x}+\mathbf{y}) \leq \phi(\mathbf{x}) + \phi(\mathbf{y})$,

(iv) $\phi(\mathbf{\Pi x}) = \phi(\mathbf{x})$,

(v) $\phi(\mathbf{\Sigma x}) = \phi(\mathbf{x})$,

where $\alpha \in \mathbb{R}$, $\mathbf{\Pi}$ is a permutation matrix, and $\mathbf{\Sigma}$ is a diagonal matrix with diagonal entries in $\{-1, +1\}$. Ziętak [Zie97] shows that the MPP minimizes any unitarily invariant norm.

Theorem 4.1 (Ziętak, 1997)**.** *Let $\|\cdot\|_\phi$ be a unitarily invariant norm corresponding to a symmetric gauge function ϕ. Then, for any $\mathbf{A} \in \mathbb{C}^{m\times n}$, $\left\|\mathbf{A}^\dagger\right\|_\phi = \min\left\{\|\mathbf{B}\|_\phi : \mathbf{B} \in \mathcal{G}(\mathbf{A})\right\}$. If additionally ϕ is strictly monotonic, then the set of minimizers contains a single element $\mathbf{A}^\dagger$.*

It is interesting to note that in the case of the operator norm, which is associated to the symmetric gauge function $\phi(\cdot) = \|\cdot\|_\infty$, the minimizer is not unique. Ziętak mentions a simple example for rank-deficient matrices, but multiple minimizers are present in the full-rank case too, as is illustrated by the following example.

Example 4.1. *Let the matrix $\mathbf{A}$ be*

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{51}$$

Singular values of $\mathbf{A}$ *are* $\sigma_1 = \sqrt{3}$ *and* $\sigma_2 = 1$*, and its MPP is*

$$\mathbf{A}^\dagger = \mathbf{V} \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{U}^* = \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ -1 & 2 \end{bmatrix}. \tag{52}$$

Consider now matrices of the form

$$\mathbf{A}^\ddagger = \mathbf{V} \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ 0 & 1 \\ \alpha & 0 \end{bmatrix} \mathbf{U}^*. \tag{53}$$

It is readily verified that $\mathbf{A}\mathbf{A}^\ddagger = \mathbf{I}$ *and* $\sigma(\mathbf{A}^\ddagger) = \left\{\sqrt{\alpha^2 + \frac{1}{3}},\ 1\right\}$*. Hence, whenever* $0 < |\alpha| \leq \sqrt{\frac{2}{3}}$*, we have that* $\left\|\mathbf{A}^\ddagger\right\|_{S_\infty} = \left\|\sigma(\mathbf{A}^\ddagger)\right\|_\infty = 1 = \left\|\mathbf{A}^\dagger\right\|_{S_\infty}$*, and yet* $\mathbf{A}^\ddagger \neq \mathbf{A}^\dagger$*.*

4.2 Left unitarily invariant norms

Another case of particular interest is when the norm is not fully unitarily invariant, but it is still unitarily invariant on one side. As we have restricted our attention to fat matrices, we will examine left unitarily invariant norms, because these will conveniently again lead to the MPP.

Lemma 4.1. *Let* $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{n\times m}$ *be defined as*

$$\mathbf{X} = \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{bmatrix}. \tag{54}$$

Then $\|\mathbf{X}\| \leq \|\mathbf{Y}\|$ *for any left unitarily invariant norm* $\|\cdot\|$*.*

Proof. Observe that $\mathbf{X} = \mathbf{T}\mathbf{Y}$, where

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{55}$$

with the two matrices on the right-hand side being unitary. The claim follows by applying the triangle inequality. □

With this lemma in hand, we can prove the main result for left unitarily invariant norms.

Theorem 4.2. *Let* $\|\cdot\|$ *be a left unitarily invariant norm, and let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *be full rank with* $m < n$*. Then* $\left\|\mathbf{A}^\dagger\right\| = \min\ \{\|\mathbf{X}\| : \mathbf{X} \in \mathcal{G}(\mathbf{A})\}$*. If* $\|\cdot\|$ *satisfies a strict inequality in Lemma 4.1 whenever* $\mathbf{K}_2 \neq \mathbf{0}$*, then the set of minimizers is a singleton* $\left\{\mathbf{A}^\dagger\right\}$*.*

Proof. Write $\mathbf{Y} \in \mathcal{G}(\mathbf{A})$ in the form (16). By the left unitary invariance and Lemma 4.1

$$\|\mathbf{Y}\| = \|\mathbf{M}\mathbf{U}^*\| = \left\| \begin{bmatrix} \mathbf{\Sigma}_\square^{-1}\mathbf{U}^* \\ \mathbf{S}\mathbf{U}^* \end{bmatrix} \right\| \geq \left\| \begin{bmatrix} \mathbf{\Sigma}_\square^{-1}\mathbf{U}^* \\ \mathbf{0} \end{bmatrix} \right\| = \left\|\mathbf{A}^\dagger\right\|.$$

□

4.3 Left unitarily invariant norms on the product operator

An immediate consequence of Theorem 4.2 is that the same phenomenon occurs when minimizing a left unitarily invariant norm of the product $\mathbf{XA}$. This simply comes from the observation that if $\|\cdot\|_\mu$ is left unitarily invariant, then so is $\|\cdot\|_\nu = \|\cdot\mathbf{A}\|_\mu$.

Corollary 4.1. *Let $\|\cdot\|$ be a left unitarily invariant norm, and let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be full rank with $m < n$. Then $\left\|\mathbf{A}^\dagger\right\| = \min\ \{\|\mathbf{XA}\| : \mathbf{X} \in \mathcal{G}(\mathbf{A})\}$. If $\|\cdot\|$ satisfies a strict inequality in Lemma 4.1 whenever $\mathbf{K}_2 \neq \mathbf{0}$, then the set of minimizers is a singleton $\left\{\mathbf{A}^\dagger\right\}$.*

4.4 Classical norms leading to the MPP

As a corollary, some large families of norms lead to MPP. In particular the following holds:

Corollary 4.2. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be full rank with $m < n$.*

1. **Schatten norms:** *for $1 \le p \le \infty$*

$$\mathbf{A}^\dagger \in \mathsf{ginv}_{S_p}(\mathbf{A}) \qquad \mathbf{A}^\dagger \in \mathsf{pginv}_{S_p}(\mathbf{A}).$$

   The considered sets are singletons for $p < \infty$.
   The set $\mathsf{pginv}_{S_\infty}(\mathbf{A})$ is a singleton, but $\mathsf{ginv}_{S_\infty}(\mathbf{A})$ is not necessarily *a singleton.*

2. **Columnwise mixed norms:** *for $1 \le q \le \infty$*

$$\mathbf{A}^\dagger \in \mathsf{ginv}_{|2,q|}(\mathbf{A}) \qquad \mathbf{A}^\dagger \in \mathsf{pginv}_{|2,q|}(\mathbf{A})$$

   The considered sets are singletons for $q < \infty$, but not always *for $q = \infty$.*

3. **Induced norms:** *for $1 \le p \le \infty$*

$$\mathbf{A}^\dagger \in \mathsf{ginv}_{\ell^p\to\ell^2}(\mathbf{A}) \qquad \mathbf{A}^\dagger \in \mathsf{pginv}_{\ell^p\to\ell^2}(\mathbf{A})$$

   The set $\mathsf{ginv}_{\ell^p\to\ell^2}(\mathbf{A})$ is not always *a singleton for $p \le 2$.*

Remark 4.1. *Whether $\mathsf{pginv}_{\ell^p\to\ell^2}(\mathbf{A})$, and $\mathsf{ginv}_{\ell^p\to\ell^2}(\mathbf{A})$ for $2 < p < \infty$ are singletons remains an open question.*

Remark 4.2. *Let us highlight an interesting consequence of Corollary 4.2: consider the $\|\cdot\|_{\ell^\infty\to\ell^2}$ norm, whose computation is known to be NP-complete [Lew10]. Despite this fact, Corollary 4.2 implies that we can find an optimal solution of an optimization problem involving this norm.*

Proof of Corollary 4.2 is given in Appendix A.1.

4.5 Norms that Almost Never Yield the MPP

After concentrating on matrix norms whose minimization *always* leads to the MPP, we now take a moment to look at norms that (almost) *never* lead to the MPP. The norms not covered by Corollary 4.2: columnwise mixed norms with $p \neq 2$, rowwise norms for $(p, q) \neq (2, 2)$, and

induced norms for $q \neq 2$. Our main result is the following theorem whose proof is given in Appendix A.2.

Theorem 4.3. *Consider $m < n$.*

1. *For any $m \geq 1$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $0 < p \leq \infty$, $0 < q < \infty$, $\mathsf{ginv}_{|p,q|}(\mathbf{A}) = \mathsf{ginv}_p(\mathbf{A})$. Moreover, there exists $\mathbf{A}_1 \in \mathbb{R}^{m\times n}$ such that for $0 < p, q < \infty$:*

$$\mathbf{A}_1^\dagger \in \mathsf{ginv}_{|p,q|}(\mathbf{A}_1) \iff p = 2$$

2. *For $m = 1$, and any $\mathbf{A} \in \mathbb{R}^{1\times n}$, $0 < p \leq \infty$, $0 < q < \infty$, $\mathsf{ginv}_{\underline{\overline{p,q}}}(\mathbf{A}) = \mathsf{ginv}_q(\mathbf{A})$. Hence, the matrix $\mathbf{A}_1 \in \mathbb{R}^{1\times n}$ satisfies for $0 < p \leq \infty$, $0 < q < \infty$:*

$$\mathbf{A}_1^\dagger \in \mathsf{ginv}_{\underline{\overline{p,q}}}(\mathbf{A}_1) \iff q = 2$$

   For any $m \geq 2$, there exists $\mathbf{A}_2 \in \mathbb{R}^{m\times n}$ such that for $0 < p, q < \infty$:

$$\mathbf{A}_2^\dagger \in \mathsf{ginv}_{\underline{\overline{p,q}}}(\mathbf{A}_2) \iff p = 2$$

   For any $m \geq 3$, there exists $\mathbf{A}_3 \in \mathbb{R}^{m\times n}$ such that for $0 < p, q < \infty$:

$$\mathbf{A}_3^\dagger \in \mathsf{ginv}_{\underline{\overline{p,q}}}(\mathbf{A}_3) \iff (p,q) \in \{(2,2),(1,2)\}$$

   Whether a similar construction exists for $m = 2$ remains open. Combining the above, for any $m \geq 3$, $0 < p, q < \infty$, the following properties are equivalent

   - *for any $\mathbf{A} \in \mathbb{R}^{m\times n}$ we have $\mathbf{A}^\dagger \in \mathsf{ginv}_{\underline{\overline{p,q}}}(\mathbf{A})$*
   - *$(p,q) = (2,2)$.*

   The matrix $(\mathbf{A}_3)^\dagger$ has positive entries. In fact, any $\mathbf{A}$ such that $(\mathbf{A})^\dagger$ has positive entries satisfies $(\mathbf{A})^\dagger \in \mathsf{ginv}_{\underline{\overline{1,2}}}(\mathbf{A})$.

3. *For $m = 1$, and any $\mathbf{A} \in \mathbb{R}^{1\times n}$, $0 < p, q \leq \infty$, $\mathsf{pginv}_{|p,q|}(\mathbf{A}) = \mathsf{ginv}_p(\mathbf{A})$. Hence, the matrix $\mathbf{A}_1 \in \mathbb{R}^{m\times 1}$ satisfies for $0 < p < \infty$, $0 < q \leq \infty$:*

$$\mathbf{A}_1^\dagger \in \mathsf{ginv}_{|p,q|}(\mathbf{A}_1) \iff p = 2$$

   For any $m \geq 2$, there exists $\mathbf{A}_4 \in \mathbb{R}^{m\times n}$ such that for $1 \leq p < \infty$ and $0 < q < \infty$:

$$\mathbf{A}_4^\dagger \in \mathsf{pginv}_{|p,q|}(\mathbf{A}_4) \iff p = 2$$

   Whether a similar construction is possible for $m \geq 2$ and $0 < p \leq 1$ remains open.

4. *For $m = 1$, and any $\mathbf{A} \in \mathbb{R}^{1\times n}$, $0 < p, q \leq \infty$, $\mathsf{pginv}_{\underline{\overline{p,q}}}(\mathbf{A}) = \mathsf{ginv}_q(\mathbf{A})$. Hence, the matrix $\mathbf{A}_1 \in \mathbb{R}^{m\times 1}$ satisfies for $0 < p \leq \infty$, $0 < q < \infty$:*

$$\mathbf{A}_1^\dagger \in \mathsf{pginv}_{\underline{\overline{p,q}}}(\mathbf{A}_1) \iff q = 2$$

For any $m \geq 2$, the matrix $\mathbf{A}_4$ satisfies for $1 \leq p \leq \infty$, $0 < q < \infty$:

$$\mathbf{A}_4^\dagger \in \mathsf{pginv}_{\underline{p,q}}(\mathbf{A}_4) \quad \Longleftrightarrow \quad q = 2$$

Whether a similar construction is possible for $m \geq 2$ and $0 < p \leq 1$ remains open.

For any $m \geq 3$, there exists $\mathbf{A}_5 \in \mathbb{R}^{m \times n}$ such that for $1 \leq p < \infty$ and $q = 2$:

$$\mathbf{A}_5^\dagger \in \mathsf{pginv}_{\underline{p,q}}(\mathbf{A}_3) \quad \Longleftrightarrow \quad p = 2$$

Whether a similar construction for $m = 2$ and/or $0 < p \leq 1$ is possible remains open.

Combining the above, for any $m \geq 3$, $1 \leq p < \infty$ $0 < q < \infty$, the following properties are equivalent

- *for any $\mathbf{A} \in \mathbb{R}^{m \times n}$ we have $\mathbf{A}^\dagger \in \mathsf{pginv}_{\underline{p,q}}(\mathbf{A})$*
- $(p, q) = (2, 2)$.

Remark 4.3. *It seems reasonable to expect that for "most" p, q and "most" matrices, the corresponding set of generalized inverses will not contain the MPP. However, the case of rowwise norms $\|\cdot\|_{\underline{1,2}}$ and matrices with positive entries suggests that one should be carefully about the precise statement.*

5 Matrices Having the Same Inverse for Many Norms

As we have seen, a large class of matrix norms are minimized by the Moore-Penrose pseudoinverse. We now discuss some classes of matrices whose generalized inverses minimize multiple norms.

5.1 Matrices with MPP whose non-zero entries are constant along columns

We first look at matrices for which all nonzero-entries of any column of $\mathbf{A}^\dagger$ have the same magnitude. For such matrices, the MPP actually minimizes many norms beyond those already covered by Corollary 4.2.

Theorem 5.1. *Let $\mathbf{A} \in \mathbb{C}^{m \times n}$. Suppose that every column of $\mathbf{X} = \mathbf{A}^\dagger$ has entries of the same magnitude (possibly differing among columns) over its non-zero entries; that is, $|x_{ij}| \in \{c_j, 0\}$. Then the following statements are true:*

1. *For $1 \leq p \leq \infty$, $0 < q \leq \infty$, we have*

$$\mathbf{A}^\dagger \in \mathsf{ginv}_{|p,q|}(\mathbf{A}).$$

   This set is a singleton for $1 < p < \infty$ and $0 < q < \infty$.
   NB: this includes the nonconvex case $0 < q < 1$.

2. *For $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, assuming further that $c_j = c$ is identical for all columns, we have*

$$\mathbf{A}^\dagger \in \mathsf{ginv}_{\underline{p,q}}(\mathbf{A}).$$

This set is a singleton for $1 < p < \infty$, $1 \leq q < \infty$.

Example 5.1. *Primary examples of matrices $\mathbf{A}$ satisfying the assumptions of Theorem 5.1 are tight frames $\mathbf{A}$ with entries of constant magnitude, such as partial Fourier matrices, $\mathbf{A} = \mathbf{RF}$ (resp. partial Hadamard matrices, $\mathbf{A} = \mathbf{RH}$), with $\mathbf{R}$ a restriction of the identity matrix $\mathbf{I}_n$ to some arbitrary subset of m rows and $\mathbf{F}$ (resp. $\mathbf{H}$) the Fourier (resp. a Hadamard) matrix of size n. Indeed, when $\mathbf{A}$ is a tight frame, we have $\mathbf{AA}^* \propto \mathbf{I}_m$, hence $\mathbf{A}^\dagger = \mathbf{A}^*(\mathbf{AA}^*)^{-1} \propto \mathbf{A}^*$. When in addition the entries of $\mathbf{A}$ have equal magnitude, so must the entries of $\mathbf{A}^\dagger$.*

The proof of Theorem 5.1 uses a characterization of the gradient of the considered norms.

Proof of Theorem 5.1. Consider $\mathbf{E}$ a matrix such that the matrix $\mathbf{Y} = \mathbf{X} + \mathbf{E}$ still belongs to $\mathcal{G}(\mathbf{A})$, and denote its columns as $\boldsymbol{\varepsilon}_i$. For each column we have $\mathbf{A}(\mathbf{x}_i + \boldsymbol{\varepsilon}_i) = \mathbf{e}_i = \mathbf{A}\mathbf{x}_i$, that is, $\boldsymbol{\varepsilon}_i$ must be in the nullspace of $\mathbf{A}$ for each column i. Since $\mathcal{N}(\mathbf{A}) = \mathsf{span}(\mathbf{A}^*)^\perp = \mathsf{span}(\mathbf{X})^\perp$, $\boldsymbol{\varepsilon}_i$ must be orthogonal to any column of $\mathbf{X}$, and in particular to $\mathbf{x}_i$. As a result, for each column

$$\langle \mathbf{x}_i, \boldsymbol{\varepsilon}_i \rangle_{\mathbb{C}^n} = 0. \tag{56}$$

1. To show this statement, it suffices to show that the columns $\mathbf{x}_i$ of $\mathbf{X} = \mathbf{A}^\dagger$ minimize all ℓ^p norms, $1 \leq p \leq \infty$, among the columns of $\mathbf{Y} \in \mathcal{G}(\mathbf{A})$.

   First consider $1 \leq p < \infty$ and $f(\mathbf{z}) \stackrel{\text{def}}{=} \|\mathbf{z}\|_p$ for $\mathbf{z} \in \mathbb{C}^n$. By convexity of $\mathbb{R}^n \times \mathbb{R}^n \ni (\mathbf{x}_R, \mathbf{x}_I) \mapsto f(\mathbf{x}_R + j\mathbf{x}_I)$, we have that for any $\mathbf{u} \in \frac{\partial f}{\partial \mathbf{z}}(\mathbf{x}_i)$ (as defined in Lemma A.2)

   $$\|\mathbf{x}_i + \boldsymbol{\varepsilon}\|_p \geq \|\mathbf{x}_i\|_p + \langle \Re(\mathbf{u}), \boldsymbol{\varepsilon}_R \rangle_{\mathbb{R}^n} + \langle \Im(\mathbf{u}), \boldsymbol{\varepsilon}_I \rangle_{\mathbb{R}^n} = \|\mathbf{x}_i\|_p + \Re\left(\langle \mathbf{u}, \boldsymbol{\varepsilon} \rangle_{\mathbb{C}^n}\right),$$

   where we recall that $\langle \mathbf{u}, \boldsymbol{\varepsilon} \rangle_{\mathbb{C}^n} = \sum_i u_i \epsilon_i^*$ involves a complex conjugation of $\boldsymbol{\varepsilon}$. By Lemma A.2, since the column $\mathbf{x}_i$ of $\mathbf{X}$ has all its nonzero entries of the same magnitude there is $c = c(p, \mathbf{x}_i) > 0$ such that $\mathbf{u} \stackrel{\text{def}}{=} c\mathbf{x}_i \in \frac{\partial f}{\partial \mathbf{z}}(\mathbf{x}_i)$. In particular, we choose the element of the subdifferential with zeros for entries corresponding to zeros in $\mathbf{x}_i$. Using (56) then gives:

   $$\|\mathbf{x}_i + \boldsymbol{\varepsilon}_i\|_p \geq \|\mathbf{x}_i\|_p + c\,\Re\left(\langle \mathbf{x}_i, \boldsymbol{\varepsilon}_i \rangle_{\mathbb{C}^n}\right) = \|\mathbf{x}_i\|_p. \tag{57}$$

   Since this holds true for any $1 \leq p < \infty$, we also get $\|\mathbf{x}_i + \boldsymbol{\varepsilon}_i\|_\infty \geq \|\mathbf{x}_i\|_\infty$ by considering the limit when $p \to \infty$. For $1 < p < \infty$ and $0 < q < \infty$, the strict convexity of f and the strict monotonicity of the ℓ^q (quasi)norm imply that the inequality is strict whenever $\mathbf{E} \neq \mathbf{0}$, hence the uniqueness result.

2. First consider $1 \leq p < \infty$ and $1 \leq q < \infty$, and $f(\mathbf{Z}) \stackrel{\text{def}}{=} \|\mathbf{Z}\|_{\underline{p,q}}$ for $\mathbf{Z} \in \mathbb{C}^{n \times m}$. By convexity of $(\mathbf{X}_R, \mathbf{X}_I) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \mapsto f(\mathbf{X}_R + j\mathbf{X}_I)$ we have for any $\mathbf{U} \in \frac{\partial f}{\partial \mathbf{Z}}(\mathbf{X})$ (as defined in Lemma A.2)

   $$\|\mathbf{X} + \mathbf{E}\|_{\underline{p,q}} \geq \|\mathbf{X}\|_{\underline{p,q}} + \Re\left(\langle \mathbf{U}, \mathbf{E} \rangle_{\mathbb{C}^{n \times m}}\right).$$

   By Lemma A.2 since $\mathbf{X}$ has all its nonzero entries of the same magnitude there is a constant $c = c(p, q, \mathbf{X}) > 0$ such that $\mathbf{U} \stackrel{\text{def}}{=} c \cdot \mathbf{X} \in \frac{\partial f}{\partial \mathbf{Z}}(\mathbf{X})$ (we again choose the element of the subdifferential with zeros for entries corresponding to zeros in $\mathbf{X}$). Using (56) we have $\langle \mathbf{X}, \mathbf{E} \rangle_{\mathbb{C}^{n \times m}} = 0$ and it follows that:

   $$\|\mathbf{X} + \mathbf{E}\|_{\underline{p,q}} = \|\mathbf{X}\|_{\underline{p,q}} + c \cdot \Re\left(\langle \mathbf{X}, \mathbf{E} \rangle_{\mathbb{C}^{n \times m}}\right) = \|\mathbf{X}\|_{\underline{p,q}}. \tag{58}$$

As above, the inequality $\|\mathbf{X}+\mathbf{E}\|_{\overline{p,q}} \geq \|\mathbf{X}\|_{\overline{p,q}}$ is extended to $p=\infty$ and/or $q=\infty$ by considering the limit. For $1<p<\infty$, $1 \leq q < \infty$, the function f is strictly convex implying that the inequality is strict whenever $\mathbf{E} \neq \mathbf{0}$. This establishes the uniqueness result.

□

5.2 Matrices with a highly sparse generalized inverse

Next we consider matrices for which a single generalized inverse, which is *not* the MPP, simultaneously minimizes several norms. It is known that if $\mathbf{x}$ is sufficiently sparse, then it can be uniquely recovered from $\mathbf{y}=\mathbf{A}\mathbf{x}$ by ℓ^p minimization with $0 \leq p \leq 1$. Denote $k_p(\mathbf{A})$ the largest integer k such that this holds true for any k-sparse vector $\mathbf{x}$:

$$k_p \stackrel{\text{def}}{=} \max \left\{ k \ : \ \underset{\mathbf{A}\widehat{\mathbf{x}}=\mathbf{A}\mathbf{x}}{\arg\min} \|\widehat{\mathbf{x}}\|_p = \{\mathbf{x}\}, \ \forall \mathbf{x} \text{ such that } \|\mathbf{x}\|_0 \leq k \right\}. \tag{59}$$

Using this definition we have the following theorem:

Theorem 5.2. *Consider $\mathbf{A} \in \mathbb{C}^{m\times n}$, and assume there exists a generalized inverse $\mathbf{X} \in \mathcal{G}(\mathbf{A})$ such that every of its columns $\mathbf{x}_i$ is $k_{p_0}(\mathbf{A})$-sparse, for some $0 < p_0 \leq 1$. Then, for all $0 < q < \infty$, and all $0 \leq p \leq p_0$, we have*

$$\mathsf{ginv}_{|p,q|}(\mathbf{A}) = \{\mathbf{X}\}.$$

Proof. It is known [GN07] that for $0 \leq p \leq p_0 \leq 1$ and any $\mathbf{A}$ we have $k_p(\mathbf{A}) \geq k_{p_0}(\mathbf{A})$. Hence, the column $\mathbf{x}_i$ of $\mathbf{X}$ is the unique minimum ℓ^p norm solution to $\mathbf{A}\mathbf{x}=\mathbf{e}_i$. When $q<\infty$ this implies $\mathsf{ginv}_{|p,q|}(\mathbf{A}) = \{\mathbf{X}\}$. When $q=\infty$ this simply implies $\mathsf{ginv}_{|p,q|}(\mathbf{A}) \ni \mathbf{X}$. □

Example 5.2. *Let $\mathbf{A}$ be the Dirac-Fourier (resp. the Dirac-Hadamard) dictionary, $\mathbf{A}=[\mathbf{I},\ \mathbf{F}]$ (resp. $\mathbf{A}=[\mathbf{I},\ \mathbf{H}]$). It is known that $k_1(\mathbf{A}) \geq (1+\sqrt{m})/2$ (see,* e.g.*, [GN07]). Moreoever*

$$\mathbf{X} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \in \mathcal{G}(\mathbf{A})$$

has k-sparse columns with $k=1$. As a result, $\mathsf{ginv}_{|p,q|}(\mathbf{A}) = \{\mathbf{X}\}$ for any $0 \leq p \leq 1$ and $0<q<\infty$. On this specific case, the equality can also be checked for $q=\infty$. Thus, $\mathsf{ginv}_{|p,q|}(\mathbf{A})$ is distinct from the MPP of $\mathbf{A}$, $\mathbf{A}^\dagger = \mathbf{A}^*/2$.

6 Computation of Norm-Minimizing Generalized Inverses

For completeness, we briefly discuss the computation of the generalized inverses associated with various matrix norms. For simplicity, we only discuss *real-valued* matrices.

General-purpose convex optimization packages such as CVX [GB14, GB08] transform the problem into a semidefinite or second-order cone program, and then invoke the corresponding interior point solvers. This makes them quite slow, especially when the program cannot be reduced to a vector form.

For the kind of problems that we aim to solve, it is more appropriate to use methods such as the projected gradient method, or the alternating direction method of multipliers (ADMM) [PB14], also known as the Douglas-Rachford splitting method [EB92]. We focus on ADMM as it nicely handles non-smooth objectives.

6.1 Norms that Reduce to the Vector Case

In certain cases, it is possible to reduce the computation of a norm-minimizing generalized inverse to a collection of independent vector problems. This is the case with the columnwise mixed norms (and consequently entrywise norms).

Consider the minimization for $\mathsf{ginv}_{|p,q|}(\mathbf{A})$, with $q < \infty$. As $x \mapsto x^{\frac{1}{q}}$ is monotonically increasing over $\mathbb{R}^+$, it follows that

$$\underset{\mathbf{X}\in\mathcal{G}(\mathbf{A})}{\arg\min}\left(\sum_j \|\mathbf{x}_j\|_p^q\right)^{\frac{1}{q}} = \underset{\mathbf{AX}=\mathbf{I}}{\arg\min}\sum_j \|\mathbf{x}_j\|_p^q\,. \tag{60}$$

On the right-hand side there is no interaction between the columns of $\mathbf{X}$, because the constraint $\mathbf{AX} = \mathbf{I}$ can be separated into m independent constraints $\mathbf{A}\mathbf{x}_j = \mathbf{e}_j,\ j \in \{1, 2, \ldots, m\}$. We can therefore perform the optimization separately for every $\mathbf{x}_j$,

$$\hat{\mathbf{x}}_j = \underset{\mathbf{A}\mathbf{x}_j=\mathbf{e}_j}{\arg\min}\ \|\mathbf{x}_j\|_p^q = \underset{\mathbf{A}\mathbf{x}_j=\mathbf{e}_j}{\arg\min}\ \|\mathbf{x}_j\|_p\,. \tag{61}$$

The above procedure also gives us *a* minimizer for $q = \infty$.

This means that, in order to compute any $\mathsf{ginv}_{|p,q|}(\mathbf{A})$, we can use our favorite algorithm for finding the ℓ^p-minimal solution of an underdetermined system of linear equations. The most interesting cases are $p \in \{1, 2, \infty\}$. For $p = 2$, we of course get the MPP, and for the other two cases, we have efficient algorithms at our disposal [BJMO11, DSSSC08].

6.2 Alternating Direction Method of Multipliers (ADMM)

The ADMM is a method to solve minimization problems of the form

$$\text{minimize}\ \ f(x) + g(x), \tag{62}$$

where in our case x is a matrix $\mathbf{X}$, and g is a monotonically increasing function of a matrix norm $\|\mathbf{X}\|$. An attractive property of the method is that neither f nor g need be smooth.

The ADMM algorithm. Generically, the ADMM updates are given as follows [PB14]:

$$\begin{aligned} x^{k+1} &\stackrel{\text{def}}{=} \mathsf{prox}_{\lambda f}(z^k - u^k) \\ z^{k+1} &\stackrel{\text{def}}{=} \mathsf{prox}_{\lambda g}(x^{k+1} + u^k) \\ u^{k+1} &\stackrel{\text{def}}{=} u^k + x^{k+1} - z^{k+1}. \end{aligned} \tag{63}$$

The algorithm relies on the iterative computation of proximal operators,

$$\mathsf{prox}_{\varphi}(y) = \arg\min_z \tfrac{1}{2}\,\|y - x\|_2^2 + \varphi(x),$$

which can be seen as a generalization of projections onto convex sets. Indeed, when φ is an indicator function of a convex set, the proximal mapping is a projection.

Linearized ADMM. Another attractive aspect of ADMM is that it warrants an immediate extension from ginv to pginv via the so called *linearized ADMM*, which addresses optimization problems of the form

$$\text{minimize} \;\; f(x) + g\big(A(x)\big), \tag{64}$$

with A being a linear operator. The corresponding update rules, which involve its adjoint operator $A^\star$, are [PB14]:

$$\begin{aligned} x^{k+1} &\stackrel{\text{def}}{=} \mathsf{prox}_{\lambda f}\big(x^k - (\mu/\lambda)A^\star\big(A(x^k) - z^k + u^k\big)\big) \\ z^{k+1} &\stackrel{\text{def}}{=} \mathsf{prox}_{\lambda g}\big(A(x^{k+1}) + u^k\big) \\ u^{k+1} &\stackrel{\text{def}}{=} u^k + A(x^{k+1}) - z^{k+1}, \end{aligned} \tag{65}$$

where λ and μ satisfy $0 < \mu \leq \lambda / \|A\|^2_{\ell^2 \to \ell^2}$. Thus we may use almost the same update rules to optimize for $\|\mathbf{X}\|$ and for $\|\mathbf{XA}\|$, at a disadvantage of having to tune an additional parameter μ.

As far as convergence goes, it can be shown that under mild assumptions, and with a proper choice of λ, ADMM converges to the optimal value. The convergence rate is in general sublinear; nevertheless, as Boyd notes in [BPC11], "Simple examples show that ADMM can be very slow to converge to high accuracy. However, it is often the case that ADMM converges to modest accuracy—sufficient for many applications—within a few tens of iterations."

6.3 ADMM for Norm-Minimizing Generalized Inverses

In order to use ADMM, we have to transform the constrained optimization program

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \|\mathbf{X}\| \\ \text{subject to} \quad & \mathbf{X} \in \mathcal{G}(\mathbf{A}) \end{aligned} \tag{66}$$

into an unconstrained program of the form (62). This is achieved by using indicator functions, defined as

$$\mathbb{I}_{\mathcal{S}}(x) \stackrel{\text{def}}{=} \begin{cases} 0, & x \in \mathcal{S} \\ +\infty, & x \notin \mathcal{S}. \end{cases} \tag{67}$$

The indicator function rules out any x that does not belong to the argument set $\mathcal{S}$; in our case, $\mathcal{S}$ is the affine space $\mathcal{G}(\mathbf{A})$. Using these notations, we can rewrite (66) as an unconstrained program for computing $\mathsf{ginv}_{\|\cdot\|}(\mathbf{A})$:

$$\underset{\mathbf{X}}{\text{minimize}} \quad \mathbb{I}_{\mathcal{G}(\mathbf{A})}(\mathbf{X}) + h(\|\mathbf{X}\|), \tag{68}$$

where h is a monotonically increasing function on $\mathbb{R}^+$ (typically a power). Using similar reasoning, we can write the unconstrained program for computing $\mathsf{pginv}_{\|\cdot\|}(\mathbf{A})$ as

$$\underset{\mathbf{X}}{\text{minimize}} \quad \mathbb{I}_{\mathcal{G}(\mathbf{A})}(\mathbf{X}) + h(\|\mathbf{XA}\|). \tag{69}$$

Next, we need to compute the proximal mappings for $h(\|\cdot\|)$ and for the indicator function.

6.3.1 Proximal operator of the indicator function $\mathbb{I}_{\mathcal{G}(\mathbf{A})}(\mathbf{X})$

The proximal operator of an indicator function of a convex set is simply the Euclidean projection onto that set,

$$\mathsf{prox}_{\mathbb{I}_{\mathcal{S}}}(v) = \mathsf{proj}_{\mathcal{S}}(v) = \arg\min_{x \in \mathcal{S}} \ \|x - v\|_2^2 .$$

In our case, $\mathcal{S} = \mathcal{G}(\mathbf{A})$, which is an affine subspace

$$\mathcal{G}(\mathbf{A}) = \left\{ \mathbf{A}^\dagger + \mathbf{N}\mathbf{Z} \ : \ \mathbf{Z} \in \mathbb{R}^{(n-m)\times(n-m)} \right\},$$

where we assumed that $\mathbf{A}$ has full rank, and the columns of $\mathbf{N} \in \mathbb{R}^{n\times(n-m)}$ form an orthonormal basis for the nullspace of $\mathbf{A}$. To project $\mathbf{V}$ orthogonally on $\mathcal{G}(\mathbf{A})$, we can translate it and project it on the parallel linear subspace, and then translate the result back into the affine subspace:

$$\mathsf{prox}_{\mathbb{I}_{\mathcal{G}(\mathbf{A})}}(\mathbf{V}) = \mathsf{proj}_{\mathcal{G}(\mathbf{A})}(\mathbf{V}) = \mathbf{A}^\dagger + \mathbf{N}\mathbf{N}^\top(\mathbf{V} - \mathbf{A}^\dagger) = \mathbf{A}^\dagger + \mathbf{N}\mathbf{N}^\top\mathbf{V}. \tag{70}$$

6.3.2 Proximal operators of some matrix norms

Proximal operators associated with matrix norms are typically more involved than projections onto affine spaces. In what follows, we discuss the proximal operators for some mixed and induced norms.

An important ingredient is a useful expression for the proximal operator of any norm in terms of a projection onto a dual norm ball. We have that for any scalar $\lambda > 0$,

$$\mathsf{prox}_{\lambda\|\cdot\|}(\mathbf{V}) = \mathbf{V} - \lambda \mathsf{proj}_{\left\{\mathbf{X} : \|\mathbf{X}\|_\star \leq 1\right\}}(\mathbf{V}), \tag{71}$$

with $\|\cdot\|_\star$ the dual norm to $\|\cdot\|$. Thus computing the proximal operator of a norm amounts to projecting onto a norm ball of the dual norm. This means that we can compute the proximal operator efficiently if we can project efficiently, and vice-versa.

Mixed norms (columnwise and rowwise). Even though we can compute $\mathsf{ginv}_{|p,q|}(\mathbf{A})$ by solving a series of vector problems, it is useful to rely on a common framework for the computations in terms of proximal operators; this will help later when dealing with $\mathsf{pginv}_{\cdot}()$.

Instead of computing the proximal mapping for $\|\cdot\|_{|p,q|}$, we compute it for $\|\cdot\|_{|p,q|}^q$ as it yields the same generalized inverse. Similarly as in Section 6.1, we have that

$$\begin{aligned}
\mathsf{prox}_{\lambda\|\cdot\|_{|p,q|}^q}(\mathbf{V}) &= \arg\min_{\mathbf{X}} \sum_j \lambda \|\mathbf{x}_j\|_p^q + \tfrac{1}{2}\|\mathbf{V} - \mathbf{X}\|_F^2 \\
&= \arg\min_{\mathbf{X}} \sum_j \left(\lambda \|\mathbf{x}_j\|_p^q + \tfrac{1}{2}\|\mathbf{v}_j - \mathbf{x}_j\|_2^2 \right) \\
&= \left(\mathsf{prox}_{\lambda\|\cdot\|_p^q}(\mathbf{v}_j) \right)_j
\end{aligned} \tag{72}$$

where, unlike in Section 6.1, special care must be taken for $q = \infty$ (see Lemma 6.1 below).

Conveniently, the exact same logic holds for rowwise mixed norms: their proximal mappings can be constructed from vector proximal mappings by splitting into a collection of vector problems

as follows:

$$\begin{aligned}\mathsf{prox}_{\lambda\|\cdot\|^q_{\underline{p,q}}}(\mathbf{V}) &= \arg\min_{\mathbf{X}} \sum_i \lambda \left\|\mathbf{x}^i\right\|_p^q + \tfrac{1}{2}\left\|\mathbf{V}-\mathbf{X}\right\|_F^2 \\ &= \arg\min_{\mathbf{X}} \sum_i \left(\left\|\mathbf{x}^i\right\|_p^q + \tfrac{1}{2}\left\|\mathbf{v}^i-\mathbf{x}^i\right\|_2^2\right) \\ &= \left(\mathsf{prox}_{\lambda\|\cdot\|_p^q}((\mathbf{v}^i)^\top)\right)_i^\top\end{aligned} \tag{73}$$

Interestingly, even though the computation of the generalized inverses corresponding to rowwise mixed norms does not decouple over rows or columns, we can decouple the computation of the corresponding proximal mappings as long as $q < \infty$. This is possible because the minimization for the latter is unconstrained.

Now the task is to find efficient algorithms to compute the proximal mappings for powers of vector ℓ^p-norms, $\|\cdot\|_p^q$. We list some known results for the most interesting (and the simplest) case of $q = 1$, and $p \in \{1, 2, \infty\}$.

1. When $p = 1$ we have the so-called soft thresholding operator,

$$\mathsf{prox}_{\lambda\|\cdot\|_1}(\mathbf{v}) = \mathsf{sign}(\mathbf{v}) \odot (|\mathbf{v}| - \lambda\mathbf{1})_+.$$

   where $|\mathbf{a}|$, $\mathbf{a} \odot \mathbf{b}$, $(\mathbf{a})_+$ denote entrywise absolute value, multiplication, positive part, and $\mathbf{1}$ has all entries equal to one.

2. For $p = 2$ we have that

$$\mathsf{prox}_{\lambda\|\cdot\|_2}(\mathbf{v}) = \max\left\{1 - \lambda/\left\|\mathbf{v}\right\|_2,\ 0\right\}\mathbf{v}.$$

   The case of $p = 2$ is interesting for rowwise mixed norms; for the columnwise norms, we simply recover the MPP.

3. Finally, for $p = \infty$, it is convenient to exploit the relationship with the projection operator (71). The dual norm to ℓ^∞-norm is the ℓ^1-norm. We can project on its unit norm-ball as follows:

$$\mathsf{proj}_{\left\{\mathbf{z}:\|\mathbf{z}\|_1\le 1\right\}}(\mathbf{v}) = \mathsf{sign}(\mathbf{v}) \odot (|\mathbf{v}| - \lambda\mathbf{1})_+,$$

   where

$$\lambda = \begin{cases} 0 & \|\mathbf{v}\|_1 \le 1 \\ \text{solution of } \sum_i \max(|x_i| - \lambda, 0) = 1 & \text{otherwise.}\end{cases}$$

   This can be computed in linear time [LY09] and the proximal mapping is then given by (71).

In all cases, to obtain the corresponding matrix proximal mapping, we simply concatenate the columns (or rows) obtained by vector proximal mappings.

Norms that do not (Easily) Reduce to the Vector Case The analysis of the proximal mappings in the previous section fails when $q = \infty$. On the other hand, computing the poor man's ℓ^1-minimization generalized inverse $\mathsf{ginv}_{\ell^1\to\ell^1}(\mathbf{A})$ using ADMM requires us to find the proximal operator for $\|\cdot\|_{\ell^1\to\ell^1} = \|\cdot\|_{|1,\infty|}$. Note that we could compute $\mathsf{ginv}_{|1,\infty|}(\mathbf{A})$ directly by decoupling, but we cannot do it for the corresponding $\mathsf{pginv}()$ nor for the rowwise norms. We can again derive the corresponding proximal mapping using (71). The following lemma will prove useful:

Lemma 6.1. *The dual norm of* $\|\cdot\|_{\ell^1\to\ell^q} = \|\cdot\|_{|q,\infty|}$ *is*

$$\|\cdot\|^*_{\ell^1\to\ell^q} = \|\cdot\|_{|q^*,1|}\,,$$

and the dual norm of $\|\cdot\|_{\ell^p\to\ell^\infty} = \|\cdot\|_{\underline{p^*,\infty}}$ *is*

$$\|\cdot\|^*_{\ell^p\to\ell^\infty} = \|\cdot\|_{\underline{p,1}}\,,$$

where $1/p + 1/p^* = 1/q + 1/q^* = 1$.

Proof. The first result is a direct consequence of the characterization (27) of $\ell^1 \to \ell^q$ norms in terms of columnwise mixed norms $|q,\infty|$, and of [Sra12, Lemma 3]. The second result follows from the characterization (38) of $\ell^p \to \ell^\infty$ norms in terms of $\underline{p^*,\infty}$ and, again, of [Sra12, Lemma 3]. □

Combining (71) with Lemma 6.1 shows that for $\|\cdot\|_{\ell^1\to\ell^q}$, where $q \in \{1,2,\infty\}$, computing the proximal operator means projecting onto the ball of the $\|\cdot\|_{|q^*,1|}$ norm. The good news is that these projections can be computed efficiently. We have already seen the cases when $q^* \in \{1,2\}$ in the previous section (corresponding to $p \in \{2,\infty\}$).

For $q^* = \infty$ we can use the algorithm of Quattoni *et al.* [QCCD09] that computes the projection on the $\|\cdot\|_{|\infty,1|}$ (or equivalently $\|\cdot\|_{\underline{\infty,1}}$) norm ball in time $O(mn\log(mn))$. Even for a general q, the projection can be computed efficiently, but the algorithm becomes more involved [LY10, WLY13].[9] In summary, we can efficiently compute the proximal mappings for induced norms $\|\cdot\|_{\ell^1\to\ell^q}$, and equivalently for induced norms $\|\cdot\|_{\ell^p\to\ell^\infty}$, because these read as proximal mappings for certain mixed norms.

6.4 Generic ADMM for matrix norm minimization

Given a norm $\|\cdot\|$, we can now summarize the ADMM update rules for computing $\mathsf{ginv}_{\|\cdot\|}(\mathbf{A})$ and $\mathsf{pginv}_{\|\cdot\|}(\mathbf{A})$, assuming that $\mathsf{prox}_{\lambda\|\cdot\|^q}(\cdot)$ can be computed efficiently for some $0 < q < \infty$.

6.4.1 ADMM for the computation of $\mathsf{ginv}(\cdot)$

To compute $\mathsf{ginv}(\cdot)$, we simply solve (68) with $h(t) = \lambda t^q$ by running the iterates

$$\begin{aligned}
\mathbf{X}^{k+1} &\stackrel{\text{def}}{=} \mathbf{A}^\dagger + \mathbf{N}\mathbf{N}^\top(\mathbf{X}^k - \mathbf{U}^k)\\
\mathbf{Z}^{k+1} &\stackrel{\text{def}}{=} \mathsf{prox}_{\lambda\|\cdot\|^q}(\mathbf{X}^{k+1} + \mathbf{U}^k)\\
\mathbf{U}^{k+1} &\stackrel{\text{def}}{=} \mathbf{U}^k + \mathbf{X}^{k+1} - \mathbf{Z}^{k+1}.
\end{aligned} \tag{74}$$

While there are a number of references that study the choice of the regularization parameter λ for particular f and g, this choice still remains somewhat of a black art. Discussion of this topic falls out of the scope of the present contribution. In our implementations[10] we used the values of λ for which the algorithm was empirically verified to converge.

[9] Our notation differs from the notation used in [LY10, WLY13, QCCD09]; the roles of p and q are reversed.
[10] Available online at `https://github.com/doksa/altginv`.

6.4.2 Linearized ADMM for the computation of $\mathsf{pginv}(\cdot)$

Even for entrywise and columnwise mixed norms, things get more complicated when instead of $\mathsf{ginv}_{|p,q|}(\mathbf{A})$ we want to compute $\mathsf{pginv}_{|p,q|}(\mathbf{A})$. This is because the objective $\|\mathbf{XA}\|_{|p,q|}$ now mixes the columns of $\mathbf{X}$ so that they cannot be untangled. A similar issue arises when trying to compute $\mathsf{pginv}_{\overline{p,q}}(\mathbf{A})$. This issue is elegantly addressed by the linearized ADMM, and by using the proximal mappings described in the previous section.

The linearized ADMM allows us to easily adapt the updates (74) for norms on the matrix $\mathbf{XA}$, without computing the new proximal operator. To compute $\mathsf{pginv}(\cdot)$, we express (69) with $h(t) = \lambda t^q$. This has the form

$$\text{minimize } f(\mathbf{X}) \ + \ g(\mathbf{XA}). \tag{75}$$

where $g(\mathbf{V}) = h(\|\mathbf{V}\|)$. It is easy to verify that the adjoint of $A : \mathbf{X} \in \mathbb{R}^{n\times m} \to \mathbf{XA} \in \mathbb{R}^{n\times n}$ is $A^\star : \mathbf{V} \in \mathbb{R}^{n\times n} \to \mathbf{VA}^\top \in \mathbb{R}^{n\times m}$, and that $\|A\|_{\ell^2\to\ell^2} = \|\mathbf{A}\|_{\ell^2\to\ell^2}$ so that the updates for $\mathsf{pginv}_{\|\cdot\|}(\mathbf{A})$ are given as

$$\begin{aligned}
\mathbf{X}^{k+1} &\stackrel{\text{def}}{=} \mathbf{A}^\dagger + \mathbf{N}\mathbf{N}^\top \left[\mathbf{X}^k - (\mu/\lambda)(\mathbf{X}^k\mathbf{A} - \mathbf{Z}^k + \mathbf{U}^k)\mathbf{A}^\top\right] \\
\mathbf{Z}^{k+1} &\stackrel{\text{def}}{=} \mathsf{prox}_{\lambda\|\cdot\|^q}(\mathbf{X}^{k+1}\mathbf{A} + \mathbf{U}^k) \\
\mathbf{U}^{k+1} &\stackrel{\text{def}}{=} \mathbf{U}^k + \mathbf{X}^{k+1}\mathbf{A} - \mathbf{Z}^{k+1},
\end{aligned} \tag{76}$$

where $0 < \mu \leq \lambda / \|\mathbf{A}\|^2_{\ell^2\to\ell^2}$.

7 Conclusion

We presented a collection of new results on generalized matrix inverses which minimize various matrix norms. Our study is motivated by the fact that the Moore-Penrose pseudoinverse minimizes e.g. the Frobenius norm, by the intuition that minimizing the entrywise ℓ_1 norm of the inverse matrix should give us sparse pseudoinverses, and by the fact that poor man's ℓ^p minimization—a linear procedure which minimizes the worst-case ℓ^p norm blowup—is achieved by a generalized inverse which minimizes the induced $\ell^1 \to \ell^1$ norm of the associated projection matrix.

Most of the presented findings address the relation in which various norms and matrices stand with respect to the MPP. In this regard, a number of findings make our work appear Sisyphean since for various norms and matrices we merely reestablish the optimality of the MPP. We could summarize this in a maxim "When in doubt, use the MPP", which most practitioners will not find surprising.

On the other hand, we identify classes of matrix norms for which the above statement does not hold, and whose minimization leads to matrices with very different properties, potentially useful in applications. Perhaps the most interesting such generalized inverse—the sparse pseudoinverse—is studied in Part II of this two-paper series.

Future work related to the results presented in this Part I involves extensions of our results to rank deficient matrices, further study of $\mathsf{pginv}(\cdot)$ operators, and filling several "holes" in the results as we could not answer all the posed questions for all combinations of norms and matrices (see Table 1), in particular for induced norms.

8 Acknowledgments

The authors would like to thank Mihailo Kolundžija, Miki Elad, Jakob Lemvig, and Martin Vetterli for the discussions and input in preparing this manuscript, and Laurent Jacques for pointing out related work. This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906).

Appendices

A Proofs of Formal Statements

A.1 Proof of Corollary 4.2

We begin by showing that the MPP is *a* minimizer of the considered norms. The result for $\|\cdot\|_{S_p}$ with $1 \leq p \leq \infty$, for the induced norm $\|\cdot\|_{\ell^2 \to \ell^2}$ and columnwise mixed norm $\|\cdot\|_{|2,2|}$ follow from their unitarily invariance and Theorem 4.1. In contrast, the norms $\|\cdot \mathbf{A}\|_{S_p}$ generally fail to be unitarily invariant. Similarly, for $p \neq 2$ and $q \neq 2$, induced norms and columnwise mixed norms are *not* unitarily invariant, hence the results do not directly follow from Theorem 4.1. Instead, the reader can easily check that all considered norms are left unitarily invariant, hence the MPP is *a* minimizer by Theorem 4.2 and Corollary 4.1.

Uniqueness cases for Schatten norms and columnwise mixed norms. Since Schatten norms with $1 \leq p < \infty$ are fully unitarily invariant and associated to a strictly monotonic symmetric gauge function, the uniqueness result of Theorem 4.1 applies. To establish uniqueness with the norms $\|\cdot \mathbf{A}\|_{S_p}$, $1 \leq p < \infty$ we exploit the following useful Lemma (see, e.g., [Zie97, Lemma 7] and references therein for a proof).

Lemma A.1. *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{m \times n}$ be given as*

$$\mathbf{X} = \begin{bmatrix} \mathbf{K}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{K}_1 & \mathbf{K}_2 \\ \mathbf{K}_3 & \mathbf{K}_4 \end{bmatrix} \tag{77}$$

where the block $\mathbf{K}_1$ is of size r. Then we have that $\sigma_j(\mathbf{X}) \leq \sigma_j(\mathbf{Y})$ for $1 \leq j \leq r$, and for any unitarily invariant norm $\|\cdot\|_\phi$ associated to a symmetric gauge ϕ, we have $\|\mathbf{X}\|_\phi \leq \|\mathbf{Y}\|_\phi$. When ϕ is strictly monotonic, $\|\mathbf{X}\|_\phi = \|\mathbf{Y}\|_\phi$ if and only if $\mathbf{K}_2, \mathbf{K}_3$ and $\mathbf{K}_4$ are zero blocks.

Considering $\mathbf{X} \in \mathcal{G}(\mathbf{A})$ and its representation as given in (16), using the unitary invariance of the Schatten norm, we have

$$\|\mathbf{X}\mathbf{A}\|_{S_p} = \|\mathbf{M}\mathbf{\Sigma}\|_{S_p} = \left\| \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{S}\mathbf{\Sigma}_\square & \mathbf{0} \end{bmatrix} \right\|_{S_p} > \left\| \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right\|_{S_p} = \left\|\mathbf{A}^\dagger \mathbf{A}\right\|_{S_p} \tag{78}$$

as soon as $\mathbf{S}\mathbf{\Sigma}_\square \neq \mathbf{0}$, that is to say whenever $\mathbf{S} \neq \mathbf{0}$. For both types of columnwise mixed norms with $1 \leq q < \infty$, the strictness of the inequality in Lemma 4.1 when $\mathbf{K}_2 \neq \mathbf{0}$ is easy to check, and we can apply the uniqueness result of Corollary 4.1.

Uniqueness for $\mathsf{pginv}_{S_\infty}(\mathbf{A})$. To prove uniqueness for $\mathsf{pginv}_{S_\infty}(\mathbf{A})$, consider again equation (78) with $p = \infty$. Denote $\mathbf{Z} \stackrel{\text{def}}{=} \mathbf{S}\mathbf{\Sigma}_\square \neq \mathbf{0}$, and let $\mathbf{z}^*$ be its first nonzero row. We bound the largest singular value σ_1 of the matrix $\mathbf{Y}$ defined as follows

$$\mathbf{Y} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{z}^* & 0 \end{bmatrix}.$$

By the variational characterization of eigenvalues of a Hermitian matrix, we have

$$\sigma_1^2 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{Y}\mathbf{Y}^*\mathbf{x} = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^* \begin{bmatrix} \mathbf{I}_m & \mathbf{z} \\ \mathbf{z}^* & \|\mathbf{z}\|_2^2 \end{bmatrix} \mathbf{x}. \tag{79}$$

Partitioning $\mathbf{x}$ as $\mathbf{x}^* = (\widetilde{\mathbf{x}}^*,\ x_n^*)$ with $x_n \in \mathbb{C}$ we write the maximization as

$$\max_{\|\widetilde{\mathbf{x}}\|_2^2+|x_n|^2=1} \|\widetilde{\mathbf{x}}\|_2^2 + 2\Re(x_n^*\mathbf{z}^*\widetilde{\mathbf{x}}) + |x_n|^2 \|\mathbf{z}\|_2^2 . \tag{80}$$

Since $\mathbf{z} \neq 0$ by assumption, it must have at least one non-zero entry. Let i be the index of a non-zero entry, and restrict $\widetilde{\mathbf{x}}$ to the form $\widetilde{\mathbf{x}} = \alpha\mathbf{e}_i$, $\alpha \in \mathbb{C}$. We have that

$$\max_{\|\widetilde{\mathbf{x}}\|_2^2+|x_n|^2=1} \|\widetilde{\mathbf{x}}\|_2^2+2\Re(x_n^*\mathbf{z}^*\widetilde{\mathbf{x}})+|x_n|^2 \|\mathbf{z}\|_2^2 \geq \max_{|\alpha|\leq 1} |\alpha|^2+2\sqrt{1-|\alpha|^2}\cdot|z_i|\cdot|\alpha|+(1-|\alpha|^2)|z_i|^2, \tag{81}$$

where we used the fact that maximization over a smaller set can only diminish the optimum, as well as $\|\mathbf{z}\|_2^2 \geq |z_i|^2$. Straightforward calculus shows that the maximum of the last expression is $1+|z_i|^2$, which is strictly larger than 1. Since $\mathbf{Y}\mathbf{Y}^*$ is a principal minor of

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{S}\mathbf{\Sigma}_\square & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{S}\mathbf{\Sigma}_\square & \mathbf{0} \end{bmatrix}^*$$

it is easy to see that

$$\left\| \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{S}\mathbf{\Sigma}_\square & \mathbf{0} \end{bmatrix} \right\|_{S_\infty} \geq \sqrt{\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{Y}\mathbf{Y}^*\mathbf{x}} = \sigma_1 \geq \sqrt{1+|z_i|^2} \overset{\text{strictly}}{>} 1 = \left\|\mathbf{A}^\dagger\mathbf{A}\right\|_{S_\infty} .$$

Cases of possible lack of uniqueness. The construction of $\mathbf{A}^\ddagger \neq \mathbf{A}^\dagger$ in Example 4.1 provides a matrix $\mathbf{A}$ for which $\mathsf{ginv}_{S_\infty}(\mathbf{A}) \supset \{\mathbf{A}^\dagger, \mathbf{A}^\ddagger\}$ is not reduced to the MPP.

A counterexample for $\mathsf{ginv}_{|2,\infty|}(\cdot)$ is as follows: consider $0 < \eta < 1$ and the following matrix,

$$\mathbf{A} = \begin{bmatrix} \eta^{-1} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{82}$$

Its MPP is simply $\mathbf{A}^\dagger = \begin{bmatrix} \eta & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^\top$, and we have that $\left\|\mathbf{A}^\dagger\right\|_{|2,\infty|} = 1$. One class of generalized inverses is given as

$$\mathbf{A}^\ddagger(\alpha) = \begin{bmatrix} \eta & 0 \\ 0 & 1 \\ \alpha & 0 \end{bmatrix}. \tag{83}$$

Clearly for all α with $|\alpha| \leq \sqrt{1-\eta^2}$, $\left\|\mathbf{A}^\ddagger(\alpha)\right\|_{|2,\infty|} = \left\|\mathbf{A}^\dagger\right\|_{|2,\infty|} = 1$, hence $\mathsf{ginv}_{|2,\infty|}(\cdot)$ is not a singleton. To construct a counterexample for $\mathsf{pginv}_{|2,\infty|}(\cdot)$, consider a full rank matrix $\mathbf{A}$ with the SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. Then all the generalized inverses $\mathbf{A}^\ddagger$ are given as (16), so that

$$\left\|\mathbf{A}^\ddagger\mathbf{A}\right\|_{|2,\infty|} = \left\| \begin{bmatrix} \mathbf{V}_1^* \\ \mathbf{S}\mathbf{\Sigma}\mathbf{V}_1^* \end{bmatrix} \right\|_{|2,\infty|}, \tag{84}$$

where we applied the left unitary invariance and we partitioned $\mathbf{V} = [\mathbf{V}_1\ \mathbf{V}_2]$. Clearly, setting $\mathbf{S} = \mathbf{0}$ gives the MPP and optimizes the norm. To construct a counterexample, note that generically $\mathbf{V}_1^*$ will have columns with different 2-norms. Let $\mathbf{v}$ be its column with the largest

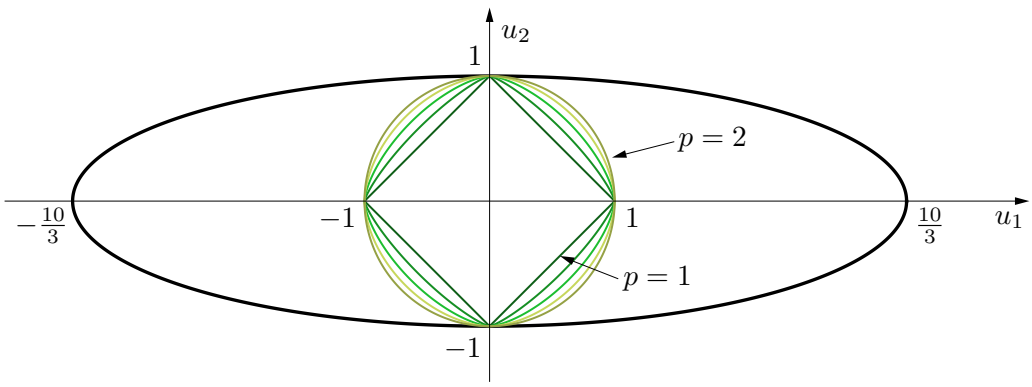

Figure 2: Geometry of the optimization problem (86). Unit norm balls $|u_1|^p + |u_2|^p = 1$ are shown for $p \in \{1, 1.25, 1.5, 1.75, 2\}$.

2-norm; we choose a matrix $\mathbf{P}$ so that $\mathbf{P\Sigma v} = \mathbf{0}$ while $\mathbf{P\Sigma V}_1^* \neq \mathbf{0}$. Then there exists ε_0 so that for any $\varepsilon \leq \varepsilon_0$,

$$\left\| \begin{bmatrix} \mathbf{V}_1^* \\ \varepsilon \mathbf{P\Sigma V}_1^* \end{bmatrix} \right\|_{|2,\infty|} = \left\| \begin{bmatrix} \mathbf{V}_1^* \\ \mathbf{0} \end{bmatrix} \right\|_{|2,\infty|} = \left\| \mathbf{A}^\dagger \mathbf{A} \right\|_{|2,\infty|}. \tag{85}$$

We can reuse counterexample (82) to show the lack of uniqueness for the induced norms (note that we already have two counterexamples: for $p = 1$ as this gives back the columnwise norm, and for $p = 2$ as this gives back the spectral norm which is the Schatten infinity norm). Let us look at $\left\| \mathbf{A}^\ddagger(\alpha) \right\|_{\ell^p \to \ell^2} = \max_{\mathbf{u} \in \mathbb{R}^2 : \|\mathbf{u}\|_p = 1} \left\| \mathbf{A}^\ddagger(\alpha)\mathbf{u} \right\|_2$. We can write this as (after squaring the 2-norm)

$$\max_{u_1^p + u_2^p = 1} (\eta^2 + \alpha^2) u_1^2 + u_2^2. \tag{86}$$

The optimization problem (86) is depicted geometrically in Figure 2: consider the ellipse with equation

$$(\eta^2 + \alpha^2) u_1^2 + u_2^2 = R^2, \tag{87}$$

We are searching for the largest R so that there exists a point on this ellipse with unit ℓ^p norm. In other words, we are squeezing the ellipse until it touches the ℓ^p norm ball. The semi-axes of the ellipse are $R/\sqrt{\eta^2 + \alpha^2}$ and R, so that when η and α are both small, the ellipses are elongated along the horizontal axis and the intersection between the squeezed ellipse and the ℓ^p ball will be close to the vertical axis. In fact, for $p \leq 2$, one can check[11] that there exists $0 < \beta < 1$ such that, whenever $\eta^2 + \alpha^2 \leq \beta^2$, the squeezed ellipse touches the ℓ^p ball only at the points $[0, \pm 1]^\top$ (as seen on Figure 2). That is, the maximum is achieved for $R = 1$. The value of the cost function at this maximum is

$$(\eta^2 + \alpha^2) 0^2 + (\pm 1)^2 = 1 \tag{88}$$

Therefore, choosing $\eta < \beta$, the value of $\left\| \mathbf{A}^\ddagger(\alpha) \right\|_{\ell^p \to \ell^2}$ is constant for all $\alpha^2 \leq \beta^2 - \eta^2$, showing that there are many generalized inverses yielding the same $\ell^p \to \ell^2$ norm as the MPP, which we know achieves the optimum.

[11] This is no longer the case for $p > 2$ for the ℓ^p ball is "too smooth" at the point $[0, 1]$: around this point on the ball we have $1 - u_2 = c|u_1|^p(1 + o(u_1)) = o(u_1^2)$.

A.2 Proof of Theorem 4.3

We will use a helper lemma about gradients and subdifferentials of matrix norms:

Lemma A.2. *For a real-valued function of a complex matrix $\mathbf{Z} = (z_{k\ell}) \in \mathbb{C}^{m\times n}$, $f(\mathbf{Z})$, denote*

$$\nabla_{\mathbf{Z}} f(\mathbf{Z}) \stackrel{\text{def}}{=} \left[\frac{\partial}{\partial x_{k\ell}} f(\mathbf{X}+j\mathbf{Y}) + j\frac{\partial}{\partial y_{k\ell}} f(\mathbf{X}+j\mathbf{Y})\right]_{k\ell}$$

with j the imaginary unit, $\mathbf{Z} = \mathbf{X} + j\mathbf{Y}$, $\mathbf{X} = (x_{k\ell}) \in \mathbb{R}^{m\times n}$, $\mathbf{Y} = (y_{k\ell}) \in \mathbb{R}^{m\times n}$. Denote $|\cdot|^r$ the entrywise r-th power of the (complex) magnitude of a vector or matrix, $\odot$ the entrywise multiplication, $\mathsf{sign}(\cdot)$ the entrywise sign with $\mathsf{sign}(z) \stackrel{\text{def}}{=} z/|z|$ for nonzero $z \in \mathbb{C}$, and $\mathsf{sign}(0) \stackrel{\text{def}}{=} \{u \in \mathbb{C}, |u| \le 1\}$. The chain rule yields:

- *for $f(z) = |z|$ the complex magnitude, we have for any nonzero scalar $z \in \mathbb{C}$*

  $$\nabla_z |z| = \frac{z}{|z|} = \mathsf{sign}(z).$$

  At $z = 0$, the subdifferential of f is $\partial_z f(0) = \mathsf{sign}(0)$.

- *for $f(z) = |z|^p$ with $0 < p < \infty$, we have for any nonzero scalar $z \in \mathbb{C}$*

  $$\nabla_z |z|^p = p\,|z|^{p-1}\,\nabla_z |z| = p\,|z|^{p-1}\,\mathsf{sign}(z).$$

  At $z = 0$, this also holds for $p > 1$, yielding $\nabla_z f(0) = 0$. See above for $p = 1$.

- *for $f(\mathbf{z}) = \|\mathbf{z}\|_p = (\sum_{k=1}^m |z_k|^p)^{\frac{1}{p}}$ with $0 < p < \infty$, we have for any vector with nonzero entries $\mathbf{z} \in \mathbb{C}^m$*

  $$\nabla_{\mathbf{z}} \|\mathbf{z}\|_p = \tfrac{1}{p}(\|\mathbf{z}\|_p^p)^{\frac{1}{p}-1}\left[\nabla_{z_k} |z_k|^p\right]_k = \|\mathbf{z}\|_p^{1-p}\left(|\mathbf{z}|^{p-1} \odot \mathsf{sign}(\mathbf{z})\right).$$

  When $\mathbf{z}$ has some (but not all) zero entries: this also holds for $p > 1$, and the entries of $\nabla_{\mathbf{z}} f(\mathbf{z})$ corresponding to zero entries of $\mathbf{z}$ are zero; for $p = 1$, the above expression yields the subdifferential of f at $\mathbf{z}$, $\partial_{\mathbf{z}} f(\mathbf{z}) = \mathsf{sign}(\mathbf{z})$.
  At $\mathbf{z} = 0$, the subdifferential of f is $\partial_{\mathbf{z}} f(\mathbf{0}) = \mathcal{B}_{p^*} = \left\{\mathbf{u} \in \mathbb{C}^m, \|\mathbf{u}\|_{p^*} \le 1\right\}$ with $\frac{1}{p^*} + \frac{1}{p} = 1$.

- *for $f(\mathbf{Z}) = \|\mathbf{Z}\|_{|p,q|} = \left(\sum_{\ell=1}^n \|\mathbf{z}_\ell\|_p^q\right)^{\frac{1}{q}}$, $0 < p, q < \infty$, we have for any matrix with nonzero entries $\mathbf{Z} \in \mathbb{C}^{m\times n}$:*

  $$\begin{aligned}
  \nabla_{\mathbf{Z}} \|\mathbf{Z}\|_{|p,q|} &= \tfrac{1}{q}\left(\|\mathbf{Z}\|_{|p,q|}^q\right)^{\frac{1}{q}-1}\left[\nabla_{\mathbf{z}_\ell} \|\mathbf{z}_\ell\|_p^q\right]_\ell = \tfrac{1}{q}\|\mathbf{Z}\|_{|p,q|}^{1-\frac{1}{q}}\left[q\,\|\mathbf{z}_\ell\|_p^{q-1}\,\nabla_{\mathbf{z}_\ell}\|\mathbf{z}_\ell\|_p\right]_\ell \\
  &= \|\mathbf{Z}\|_{|p,q|}^{1-\frac{1}{q}}\left[\|\mathbf{z}_\ell\|_p^{q-1}\,\|\mathbf{z}_\ell\|_p^{1-p}\,(|\mathbf{z}_\ell|^{p-1} \odot \mathsf{sign}(\mathbf{z}_\ell))\right]_\ell \\
  &= \|\mathbf{Z}\|_{|p,q|}^{1-\frac{1}{q}}\left(|\mathbf{Z}|^{p-1} \odot \mathsf{sign}(\mathbf{Z})\right)\mathbf{D}_{col}^{q-p}
  \end{aligned}$$

  with $\mathbf{D}_{col} \stackrel{\text{def}}{=} \mathsf{diag}(\|\mathbf{z}_\ell\|_p)$.

When $\mathbf{Z}$ has nonzero columns but some zero entries: this also holds for $p > 1$, and the entries of $\nabla_{\mathbf{Z}} \|\mathbf{Z}\|_{|p,q|}$ corresponding to zero entries of $\mathbf{Z}$ are zero; for $p = 1$ the above expression yields the subdifferential of f at $\mathbf{Z}$, $\partial_{\mathbf{Z}} f(\mathbf{Z}) = \|\mathbf{Z}\|_{|1,q|}^{1-\frac{1}{q}} \mathsf{sign}(\mathbf{Z})\mathbf{D}_{col}^{q-1}$.

When $\mathbf{Z}$ has some (but not all) zero columns, we get

- *For $p > 1$ and $q > 1$ the above extends, with zeros in the gradient entries (resp. columns) corresponding to zero entries (resp. columns) in $\mathbf{Z}$.*
- *For $p > 1$ and $q = 1$, $f(\mathbf{Z}) = \|\mathbf{Z}\|_{|p,1|} = \sum_{\ell=1}^{n} \|\mathbf{z}_\ell\|_p$, and we get a subdifferential*

$$\partial_{\mathbf{Z}} \|\mathbf{Z}\|_{|p,1|} = \left\{ \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \;\middle|\; \mathbf{v}_i = \nabla \|\mathbf{v}_i\|_p \;\; \textit{if } \mathbf{z}_i \neq \mathbf{0},\; \mathbf{v}_i \in \partial \|\mathbf{z}_i\|_p \;\; \textit{otherwise} \right\}$$

- *For $p = 1$ and $q \geq 1$ we again get the subdifferential $\partial_{\mathbf{Z}} \|\mathbf{Z}\|_{|1,q|} = \|\mathbf{Z}\|_{|p,q|}^{1-\frac{1}{q}} \mathsf{sign}(\mathbf{Z})\mathbf{D}_{col}^{q-p}$ as above. In particular, when $p = q = 1$, we have that $\partial_{\mathbf{Z}} \|\mathbf{Z}\|_{|1,1|} = \mathsf{sign}(\mathbf{Z})$.*

At $\mathbf{Z} = 0$ the subdifferential is again the unit ball of the dual norm (this follows from the definition of a subdifferential and the definition of a dual norm).

- *similarly for $f(\mathbf{Z}) = \|\mathbf{Z}\|_{\underline{p,q}} = \left(\sum_k \|\mathbf{z}^k\|_p^q\right)^{\frac{1}{q}}$, $0 < p, q < \infty$, we have for any matrix with nonzero entries $\mathbf{Z} \in \mathbb{C}^{m \times n}$:*

$$\nabla_{\mathbf{Z}} \|\mathbf{Z}\|_{\underline{p,q}} \;=\; \|\mathbf{Z}\|_{\underline{p,q}}^{1-\frac{1}{q}} \mathbf{D}_{row}^{q-p} \left(|\mathbf{Z}|^{p-1} \odot \mathsf{sign}(\mathbf{Z}) \right)$$

with $\mathbf{D}_{row} \stackrel{\text{def}}{=} \mathsf{diag}(\|\mathbf{z}^k\|_p)$.

The same extensions to $\mathbf{Z}$ with some (but not all) zero entries (resp. zero rows) hold. In particular for $p = 1$ and $\mathbf{Z}$ with nonzero rows, $\partial_{\mathbf{Z}} f(\mathbf{Z}) = \|\mathbf{Z}\|_{\underline{1,q}}^{1-\frac{1}{q}} \mathbf{D}_{row}^{q-1} \mathsf{sign}(\mathbf{Z})$.

The same results hold for the same functions and norms of real-valued *vectors and matrices.*

Proof of Theorem 4.3. For simplicity we use $g(\mathbf{X})$ to denote the considered norm (respectively $\|\mathbf{X}\|_{|p,q|}$, $\|\mathbf{XA}\|_{|p,q|}$, $\|\mathbf{X}\|_{\underline{p,q}}$ or $\|\mathbf{XA}\|_{\underline{p,q}}$). Our goal is to build a matrix $\mathbf{A}$ such that: for some $\mathbf{X} \in \mathcal{G}(\mathbf{A})$ we have $g(\mathbf{A}^\dagger) > g(\mathbf{X})$.

For any $\mathbf{A}$ with full column rank (a property that will be satisfied by the matrices we will construct), any $\mathbf{X} \in \mathcal{G}(\mathbf{A})$ can be written as $\mathbf{A}^\dagger + \mathbf{NZ}$, where $\mathbf{N} \in \mathbb{R}^{n \times (n-m)}$ is a basis for the nullspace of $\mathbf{A}$, and $\mathbf{Z} \in \mathbb{R}^{(n-m) \times m}$. Using this representation and denoting $f(\mathbf{Z}) \stackrel{\text{def}}{=} g(\mathbf{A}^\dagger + \mathbf{NZ})$, it is enough to show that the *subdifferential* of f at $\mathbf{Z} = \mathbf{0}$ does not contain $\mathbf{0}$, i.e., that

$$\mathbf{0}_{(n-m)\times m} \notin \partial f(\mathbf{0}) = \left\{ \mathbf{F} \in \mathbb{R}^{(n-m)\times m} : \; \forall \mathbf{Z},\; f(\mathbf{Z}) - f(\mathbf{0}) \geq \langle \mathbf{F}, \mathbf{Z} \rangle_F \right\}.$$

which is in turn equivalent to $\mathbf{0}_{(n-m)\times m} \notin \mathbf{N}^\top \partial g(\mathbf{A}^\dagger)$. Put differently, there is no $n \times m$ matrix $\mathbf{G} \in \partial g(\mathbf{A}^\dagger)$ such that its columns are all orthogonal to the null space of $\mathbf{A}$, $\ker(\mathbf{A}) = \mathcal{R}(\mathbf{N})$, i.e., for any $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_m] \in \partial g(\mathbf{A}^\dagger)$, there is a column index $j \in \{1, \ldots, m\}$ such that $\mathbf{g}_j \notin \mathcal{R}(\mathbf{N})^\perp = \mathcal{R}(\mathbf{A}^\top) = \mathcal{R}(\mathbf{A}^\dagger)$. At this point we have shown that, given a full rank matrix $\mathbf{A}$,

there exists $\mathbf{X} \in \mathcal{G}(\mathbf{A})$ such that $g(\mathbf{A}^\dagger) > g(\mathbf{X})$ if, and only if,

$$\mathbf{0}_{n\times m} \notin (\mathbf{I}_n - \mathbf{A}^\dagger\mathbf{A})\partial g(\mathbf{A}^\dagger). \tag{89}$$

Our new goal is thus to exhibit such a matrix $\mathbf{A}$ for all the considered norms.

1. **Case of $\mathsf{ginv}(\cdot)$ for columnwise norms.** Since $\mathsf{ginv}_{|p,q|}(\cdot) = \mathsf{ginv}_{|p,p|}(\cdot) = \mathsf{ginv}_p(\cdot)$ for $q < \infty$, it is sufficient to consider $q = p$, i.e. $g(\mathbf{X}) = \|\mathbf{X}\|_p = \left(\sum_{j=1}^m \|\mathbf{x}\|_p^p\right)^{1/p}$. By Lemma A.2, we need to consider

$$h^{\mathsf{ginv}}_{|p,p|}(\mathbf{A}) = (\mathbf{I}_n - \mathbf{A}^\dagger\mathbf{A})\left(\left|\mathbf{A}^\dagger\right|^{p-1} \odot \mathsf{sign}(\mathbf{A}^\dagger)\right)$$

and exhibit some full rank $\mathbf{A}$ such that $h^{\mathsf{ginv}}_{|p,p|}(\mathbf{A}) \ni \mathbf{0}$ if, and only if, $p = 2$. Since $\mathbf{A}^\dagger\mathbf{A}$ is the orthogonal projection onto the span of $\mathbf{A}^\dagger$, it is sufficient to find a full rank $\mathbf{B} \in \mathbb{R}^{n\times m}$ with strictly positive entries so that $(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\,|\mathbf{B}|^{p-1} = \mathbf{0}$ if, and only if, $p = 2$, with $\mathbf{P}_\mathbf{B} = \mathbf{B}\mathbf{B}^\dagger$ the orthogonal projection onto the span of $\mathbf{B}$ (the matrix $\mathbf{B}$ plays the role of $\mathbf{A}^\dagger$, and $\mathbf{B}^\dagger$ the role of $\mathbf{A}$).

   Let $\mathbf{C} \in \mathbb{R}^{n\times m}$ be a matrix with 0/1 entries and $\mathbf{B} \stackrel{\text{def}}{=} \mathbf{1}_n\mathbf{1}_m^\top + \mathbf{C}$. All entries of $\mathbf{B}$ are ones except those associated to nonzero entries in $\mathbf{C}$, which are twos. Hence

$$|\mathbf{B}|^{p-1} = \mathbf{1}_n\mathbf{1}_m^\top + (2^{p-1}-1)\mathbf{C} = \mathbf{1}_n\mathbf{1}_m^\top + (2^{p-1}-1)(\mathbf{B} - \mathbf{1}_n\mathbf{1}_m^\top) = (2^{p-1}-1)\mathbf{B} - (2^{p-1}-2)\mathbf{1}_n\mathbf{1}_m^\top \tag{90}$$

   and

$$(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\,|\mathbf{B}|^{p-1} = -(2^{p-1}-2)(\mathbf{I}_n - \mathbf{P})\mathbf{1}_n\mathbf{1}_m^\top.$$

   Provided that $\mathbf{1}_n$ does not belong to the span of $\mathbf{B}$, $(\mathbf{I}_n - \mathbf{P})\mathbf{1}_n\mathbf{1}_m^\top$ is nonzero, and the above expression is zero if, and only if, $2^{p-1} = 2$, which is equivalent to $p = 2$.

   To conclude, we just need to show we can choose a binary $\mathbf{C}$ so that: (i) $\mathbf{B}$ is full rank; and (ii) its span does not contain the vector $\mathbf{1}_n$. We let the reader check that for any $m \geq 1$ this can be achieved with

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0}_{(n-m)\times m} \end{bmatrix}.$$

   This shows the desired property with $\mathbf{A}_1 \stackrel{\text{def}}{=} \mathbf{B}^\dagger$.

2. **Case of $\mathsf{ginv}(\cdot)$ for rowwise norms.**

   For $m = 1$ with rowwise norms, we have for any $0 < p, q \leq \infty$ and $\mathbf{X} \in \mathbb{R}^{n\times 1}$: $\|\mathbf{X}\|_{\underline{p,q}} = \|\mathbf{X}\|_q$. Hence for any $0 < p, q \leq \infty$ and any $\mathbf{A} \in \mathbb{R}^{1\times n}$, $\mathsf{ginv}_{\underline{p,q}}(\mathbf{A}) = \mathsf{ginv}_q(\mathbf{A})$. This yields the desired property with $\mathbf{A}_1$ obtained above.

   For $m \geq 2$, by Lemma A.2, what we seek is $\mathbf{A}$ so that

$$h^{\mathsf{ginv}}_{\underline{p,q}}(\mathbf{A}) = (\mathbf{I}_n - \mathbf{A}^\dagger\mathbf{A})\mathsf{diag}\left(\left\|(\mathbf{A}^\dagger)^i\right\|_p^{q-p}\right)_{i=1}^n \left(\left|\mathbf{A}^\dagger\right|^{p-1} \odot \mathsf{sign}(\mathbf{A}^\dagger)\right) \ni \mathbf{0}$$

   if, and only if, (p, q) have particular values. As previously, it is sufficient to find a full rank $\mathbf{B} \in \mathbb{R}^{n\times m}$ with no zero entry so that, with $\mathbf{D}_{\text{row}} = \mathsf{diag}(\left\|\mathbf{b}^i\right\|_p)$,

$$(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{D}_{\text{row}}^{q-p}(|\mathbf{B}|^{p-1} \odot \mathsf{sign}(\mathbf{B})) = \mathbf{0}$$

   only for these values of (p, q).

We consider $\mathbf{B} \stackrel{\text{def}}{=} \mathbf{1}_n \mathbf{1}_m^\top + \mathbf{C}'$ with

$$\mathbf{C}' \stackrel{\text{def}}{=} \left[\begin{array}{c} \mathbf{I}_m \\ \hline -2, 1, \mathbf{0}_{m-2}^\top \\ \hline \mathbf{0}_{(n-1-m)\times(m-1)}; \mathbf{1}_{(n-1-m)\times 1} \end{array}\right]$$

so that $b_{(m+1)1} = -1$. In each line of $|\mathbf{B}|$, one entry is a 2 and the remaining $m-1$ entries are ones, hence for $0 < p, q < \infty$ we have $\mathbf{D}_{\text{row}}^{q-p} = C_{p,q}\mathbf{I}_n$ and

$$(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{D}_{\text{row}}^{q-p}\left[|\mathbf{B}|^{p-1} \odot \mathsf{sign}(\mathbf{B})\right] = C_{p,q}(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\left[|\mathbf{B}|^{p-1} \odot \mathsf{sign}(\mathbf{B})\right].$$

The first column of $\left[|\mathbf{B}|^{p-1} \odot \mathsf{sign}(\mathbf{B})\right]$ is $\mathbf{y} \stackrel{\text{def}}{=} \left[2^{p-1}, \mathbf{1}_{m-1}, -1, \mathbf{1}_{n-1-m}\right]^\top$. To conclude we show that this lives in the span of $\mathbf{B}$ if, and only if, $p = 2$.

Assume that $\mathbf{y} = \mathbf{B}\mathbf{z}$. Specializing to rows 2 to $m+1$, we obtain $[\mathbf{1}_{m-1}^\top, -1]^\top = \mathbf{B}'\mathbf{z}$ with $\mathbf{B}'$ the restriction of $\mathbf{B}$ to the considered m rows. We let the reader check that $\mathbf{B}'$ is invertible and that this implies $\mathbf{z} = \mathbf{e}_1$. Specializing now to the first row, we have $2^{p-1} = y_1 = (\mathbf{B}\mathbf{e}_1)_1 = 2$ that is to say $p = 2$. The converse is immediate. This shows the desired property with $\mathbf{A}_2 \stackrel{\text{def}}{=} \mathbf{B}^\dagger$.

To further characterize the role of q when $p = 2$ we now consider $m \geq 3$ and construct a full rank $\mathbf{B}$ from $\mathbf{C}$ where $\mathbf{C}$ is binary with the same number of ones in all rows except the first one, so that $\mathbf{D}_{\text{row}}$ has all diagonal entries equal except the first one and $\mathbf{D}_{\text{row}}^{q-p} = c(\mathbf{I} + \lambda\mathbf{e}_1\mathbf{e}_1^\top)$ where $c > 0$, and $\lambda \neq 0$ as soon as $q \neq p$. With $\mathbf{b}^1$ the first row of $\mathbf{B}$ (a row vector),

$$\begin{aligned}
\mathbf{D}_{\text{row}}^{q-p}|\mathbf{B}|^{p-1} &= c|\mathbf{B}|^{p-1} + c\lambda\mathbf{e}_1\left|\mathbf{b}^1\right|^{p-1} \\
&\stackrel{(90)}{=} c(2^{p-1}-1)\mathbf{B} - c(2^{p-1}-2)\mathbf{1}_n\mathbf{1}_m^\top + c\lambda\mathbf{e}_1\left|\mathbf{b}^1\right|^{p-1} \\
(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{D}_{\text{row}}^{q-p}|\mathbf{B}|^{p-1} &= c(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\left(\lambda\mathbf{e}_1\left|\mathbf{b}^1\right|^{p-1} - (2^{p-1}-2)\mathbf{1}_n\mathbf{1}_m^\top\right) \\
&= c\left(\lambda\mathbf{u}\left|\mathbf{b}^1\right|^{p-1} - (2^{p-1}-2)\mathbf{v}\mathbf{1}_m^\top\right)
\end{aligned}$$

with $\mathbf{u} \stackrel{\text{def}}{=} (\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{e}_1$ and $\mathbf{v} \stackrel{\text{def}}{=} (\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{1}_n$.

For $m \geq 3$, consider the specific choice

$$\mathbf{C} = \left[\begin{array}{c} \mathbf{c}^1 \\ \hline \mathbf{I}_m \\ \hline \mathbf{0}_{(n-1-m)\times(m-1)},\ \mathbf{1}_{n-1-m} \end{array}\right]$$

where the last block is empty if $n = m+1$, and the first row is $\mathbf{c}^1 = [1, 1, \mathbf{0}_{m-2}^\top]$.

For any vector $\mathbf{z}$ and scalars α, β such that $\mathbf{B}\mathbf{z} = \alpha\mathbf{1}_n + \beta\mathbf{e}_1$, one can show that $\mathbf{z} = \beta\mathbf{1}_m$ and $\alpha = (m+1)\beta$. In particular:

(i) The matrix $\mathbf{B}$ is full rank as the equality $\mathbf{B}\mathbf{z} = \mathbf{0}_n = \alpha\mathbf{1}_n + \beta\mathbf{e}_1$ with $\alpha = \beta = 0$ implies $\mathbf{z} = \mathbf{0}_m$.

(ii) Similarly $\mathbf{B}\mathbf{z} = \alpha\mathbf{e}_1 + \beta\mathbf{1}_n$ with $\beta = 0$ implies $\mathbf{z} = \mathbf{0}$, hence $\mathbf{e}_1 \notin \mathcal{R}(\mathbf{B})$ and $\mathbf{u} \neq \mathbf{0}$;

(iii) By the same argument, $\mathbf{1}_n \notin \mathcal{R}(\mathbf{B})$, and $\mathbf{v} \neq \mathbf{0}$;

(iv) For $m \geq 3$, $(m+1)\mathbf{1}_n + \mathbf{e}_1 = \mathbf{B}\mathbf{1}_m$ *does* belong to the span of $\mathbf{B}$. Up to a scaling it is the only linear combination of $\mathbf{1}_n$ and $\mathbf{e}_1$ that lives there; if there were another one, then by linearity the whole 2D subspace would be in the span of $\mathbf{B}$ including $\mathbf{1}_n$ and $\mathbf{e}_1$ for which we have shown that this is not the case;

(v) This implies that $\mathbf{u}$ and $\mathbf{v}$ are colinear with $\mathbf{u} = -(m+1)\mathbf{v}$ as $\alpha\mathbf{v} + \beta\mathbf{u} = \mathbf{0}$ if, and only if, $\alpha\mathbf{1}_n + \beta\mathbf{e}_1 \in \mathcal{R}(\mathbf{B})$, which is equivalent to $\alpha = (m+1)\beta$.

As a result

$$(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{D}_{\text{row}}^{q-p}\,|\mathbf{B}|^{p-1} = c \cdot \mathbf{u}\left(\lambda\,|\mathbf{b}|^{p-1} + \tfrac{2^{p-1}-2}{m+1}\mathbf{1}_m^\top\right).$$

This is zero if and only if

$$\lambda\,|\mathbf{b}|^{p-1} = \tfrac{2^{p-1}-2}{m+1}\mathbf{1}_m^\top. \tag{91}$$

As $\left|\mathbf{b}^1\right|^{p-1} = [2^{p-1}, 2^{p-1}, \mathbf{1}_{m-2}^\top]$, (91) is equivalent to

$$\lambda = -(m+1)^{-1}(2^{p-1}-2)2^{-(p-1)} \tag{92}$$

$$\lambda = -(m+1)^{-1}(2^{p-1}-2). \tag{93}$$

- For $p \notin \{1,2\}$ and any q, the right-hand side of these two equalities have incompatible values, so the only way for (92)-(93) to hold is to have $p \in \{1,2\}$.
- For $p = 2$ the right-hand side of these equalities is zero, hence they are satisfied only if $\lambda = 0$. Since $\mathbf{D}_{\text{row}}^p = \mathbf{D}_{\text{row}}^2 = \mathsf{diag}(m+6, m+3, \ldots, m+3) = (m+3)\mathsf{diag}(\frac{m+6}{m+3}, 1, \ldots, 1)$, we have $\lambda = \left(\frac{m+6}{m+3}\right)^{q/p-1} - 1$ and the condition $\lambda = 0$ can only hold if $q = p$. This means that the only way for (91) to hold when $p = 2$ is to have $(p,q) = (2,2)$.
- For $p = 1$ and any q, the right-hand sides are compatible and yield the constraint $\lambda + 1 = \frac{m+2}{m+1}$. For this value of p we have $\mathbf{D}_{\text{row}} = \mathsf{diag}(m+2, m+1, \ldots, m+1)$, $c = (m+1)^{q-1}$, and $\lambda = \left(\frac{m+2}{m+1}\right)^{q-1} - 1$. Combining we get the condition

$$\frac{m+2}{m+1} = \left(\frac{m+2}{m+1}\right)^{q-1} \tag{94}$$

  which can only hold for $q = 2$.

This shows the desired property with $\mathbf{A}_3 \stackrel{\text{def}}{=} \mathbf{B}^\dagger$.

For $(p,q) = (1,2)$, the construction of $\mathbf{A}_3$ does not give us a counterexample. In fact, for this choice of p and q, whenever $\mathbf{B}$ has all entries positive, we have $\mathbf{D}_{\text{row}}^{q-p} = \mathbf{D}_{\text{row}} = \mathsf{diag}(\mathbf{B}\mathbf{1}_m)$ and

$$\begin{aligned}(\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{D}_{\text{row}}^{q-p}\left[|\mathbf{B}|^{p-1} \odot \mathsf{sign}(\mathbf{B})\right] &= (\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathsf{diag}\left(\mathbf{B}\mathbf{1}_m\right)\mathbf{1}_n\mathbf{1}_m^\top \\ &= (\mathbf{I}_n - \mathbf{P}_\mathbf{B})\mathbf{B}\mathbf{1}_m\mathbf{1}_m^\top = \mathbf{0}_{n\times m}\end{aligned}$$

where the last equality follows because $\mathbf{B}\mathbf{1}_m$ is always in the range of $\mathbf{B}$. However the counter-example built with $\mathbf{A}_2$ is valid for $(p,q) = (1,2)$ hence $(p,q) \neq (2,2)$ implies the existence of a full-rank $\mathbf{A}$ such that $\mathbf{A}^\dagger \notin \mathsf{ginv}_{\underline{p,q}}(\mathbf{A})$.

3. **Case of $\mathsf{pginv}(\cdot)$ for columnwise norms**

For columnwise norms, by Lemma A.2, we seek $\mathbf{A}$ so that

$$h^{\mathsf{pginv}}_{|p,q|}(\mathbf{A}) = (\mathbf{I}_n - \mathbf{A}^\dagger\mathbf{A})\left(\left|\mathbf{A}^\dagger\mathbf{A}\right|^{p-1} \odot \mathsf{sign}(\mathbf{A}^\dagger\mathbf{A})\right)\mathsf{diag}\left(\left\|(\mathbf{A}^\dagger\mathbf{A})_j\right\|_p^{q-p}\right)_{j=1}^n \mathbf{A}^\top \ni \mathbf{0}$$

only for specific values of (p,q), where for $p=1$ the notation $|\mathbf{M}|^{p-1} \odot \mathsf{sign}(\mathbf{M})$ should be replaced by $\mathsf{sign}(\mathbf{M})$ Again, some manipulations indicate that it is sufficient to find a full rank $\mathbf{B} \in \mathbb{R}^{n\times m}$ so that $\mathbf{P_B}$ has no zero column and we have, with $\mathbf{D} = \mathsf{diag}(\left\|(\mathbf{P_B})^i\right\|_p)$

$$(\mathbf{I}_n - \mathbf{P_B})(|\mathbf{P_B}|^{p-1} \odot \mathsf{sign}(\mathbf{P_B}))\mathbf{D}^{q-p}\mathbf{P}_B \ni \mathbf{0} \tag{95}$$

only for these values of (p,q). In other words, it is sufficient to find an orthogonal projection matrix $\mathbf{P}$ of rank m with no zero column so that with $\mathbf{D} = \mathsf{diag}(\left\|(\mathbf{P})^i\right\|_p)$:

$$(\mathbf{I}_n - \mathbf{P})(|\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}))\mathbf{D}^{q-p}\mathbf{P} \ni \mathbf{0}$$

only for controlled values of (p,q).

For rowwise norms, by Lemma A.2, we need to exhibit $\mathbf{A}$ so that we control for which (p,q) we have

$$h^{\mathsf{pginv}}_{\underline{p,q}}(\mathbf{A}) = (\mathbf{I}_n - \mathbf{A}^\dagger\mathbf{A})\mathsf{diag}\left(\left\|(\mathbf{A}^\dagger\mathbf{A})^i\right\|_p^{q-p}\right)_{j=1}^n \left(\left|\mathbf{A}^\dagger\mathbf{A}\right|^{p-1} \odot \mathsf{sign}(\mathbf{A}^\dagger\mathbf{A})\right)\mathbf{A}^\top \ni \mathbf{0}.$$

With the same reasoning as above it is sufficient to find an orthogonal projection matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ of rank m with no zero column so that we control the values of (p,q) for which

$$(\mathbf{I}_n - \mathbf{P})\mathbf{D}^{q-p}(|\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}))\mathbf{P} \ni \mathbf{0} \tag{96}$$

where $\mathbf{D} = \mathsf{diag}(\|(\mathbf{P})_j\|_p) = \mathsf{diag}(\left\|(\mathbf{P})^i\right\|_p)$ by the symmetry of $\mathbf{P}$.

We now proceed to the desired constructions for various dimensions m, n.

- For $m=1$ we have for any $0 < p,q \le \infty$ and $\mathbf{X} \in \mathbb{R}^{n\times 1}$: $\|\mathbf{XA}\|_{|p,q|} = \|\mathbf{X}\|_p \|\mathbf{A}\|_q$ and $\|\mathbf{XA}\|_{\underline{p,q}} = \|\mathbf{X}\|_q \|\mathbf{A}\|_p$. Hence for any $0 < p,q \le \infty$ and any $\mathbf{A} \in \mathbb{R}^{1\times n}$, $\mathsf{pginv}_{|p,q|}(\mathbf{A}) = \mathsf{ginv}_p(\mathbf{A})$ and $\mathsf{pginv}_{\underline{p,q}}(\mathbf{A}) = \mathsf{ginv}_q(\mathbf{A})$. This allows reusing the matrix $\mathbf{A}_1$.
- For $m \ge 2$ and columnwise norms, we construct $\mathbf{P}$ as follows: choose $0 < \theta < \pi/2$, $\theta \ne \pi/4$ and set $c \stackrel{\text{def}}{=} \cos\theta > 0$, $s \stackrel{\text{def}}{=} \sin\theta > 0$, $\mathbf{u} \stackrel{\text{def}}{=} (c,s)$; build a block diagonal projection matrix of rank m,

  $$\mathbf{P} \stackrel{\text{def}}{=} \mathsf{blockdiag}(\mathbf{uu}^\top, \mathbf{P}')$$

  where $\mathbf{P}'$ is an arbitrary $(n-2)\times(n-2)$ projection matrix of rank $m-1$ with nonzero columns. Denoting $\mathbf{v} \stackrel{\text{def}}{=} (-s,c)$ and $\mathbf{w} \stackrel{\text{def}}{=} (c^{p-1}, s^{p-1})$, we have $\mathbf{I}_2 - \mathbf{uu}^\top = \mathbf{vv}^\top$ and

  $$\begin{aligned}
  \mathbf{I}_n - \mathbf{P} &= \mathsf{blockdiag}(\mathbf{vv}^\top, \mathbf{I}_{n-m} - \mathbf{P}') \\
  \mathbf{D} &= \mathsf{blockdiag}(\|\mathbf{u}\|_p \mathsf{diag}(\mathbf{u}), \mathbf{D}') \\
  |\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}) &= \begin{cases} \mathsf{blockdiag}(\mathbf{ww}^\top, \mathbf{W}), & \text{for } 1 < p < \infty \\ \mathsf{sign}(\mathbf{P}) = \left\{ \begin{bmatrix} \mathbf{ww}^\top & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{W} \end{bmatrix}, \|\mathbf{S}_i\|_{|\infty,\infty|} \le 1 \right\} & \text{for } p = 1, \end{cases}
  \end{aligned}$$

where $\mathbf{D}'$ is diagonal and $\mathbf{W}$ is some matrix. For $p = 1$ we abused the notation and instead of specifying the whole subdifferential as a set, we kept the decomposition into parts and assigned the set to the only ambiguous term. For $p > 1$ it follows that

$$\begin{aligned}(\mathbf{I}_n - \mathbf{P})(|\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}))\mathbf{D}^{q-p}\mathbf{P} &= \|\mathbf{u}\|_p^{q-p}\ \mathsf{blockdiag}(\mathbf{v}\mathbf{v}^\top \mathbf{w}\mathbf{w}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{u}\mathbf{u}^\top, \mathbf{W}') \\ &= \|\mathbf{u}\|_p^{q-p} \cdot \langle \mathbf{v}, \mathbf{w}\rangle \cdot (\mathbf{w}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{u}) \\ &\quad \cdot \mathsf{blockdiag}(\mathbf{v}\mathbf{u}^\top, \mathbf{W}'').\end{aligned}$$

for some matrices $\mathbf{W}'$, $\mathbf{W}''$. For $p = 1$ we similarly get

$$(\mathbf{I}_n - \mathbf{P})\mathsf{sign}(\mathbf{P})\mathbf{D}^{q-p}\mathbf{P} = \|\mathbf{u}\|_p^{q-p} \cdot \langle \mathbf{v}, \mathbf{w}\rangle \cdot \mathbf{w}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{u} \left\{ \begin{bmatrix} \mathbf{v}\mathbf{u}^\top & \mathbf{W}_1 \\ \mathbf{W}_2 & \mathbf{W}_3 \end{bmatrix}, \mathbf{W}_i \in \mathcal{W}_i \right\}$$

for some matrix sets $\mathcal{W}_i$.
As $\mathbf{v}\mathbf{u}^\top \neq \mathbf{0}$, $\mathbf{w}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{u} = c^q + s^q > 0$, and $\langle \mathbf{v}, \mathbf{w}\rangle = -sc^{p-1} + cs^{p-1}$ this is zero if, and only if, $cs^{p-1} = sc^{p-1}$, i.e. $(s/c)^{p-2} = 1$. As $s/c = \tan\theta \neq 1$, this is equivalent to $p = 2$.

4. **Case of $\mathsf{pginv}(\cdot)$ for rowwise norms**

- For $m \geq 2$, we proceed similarly as for the columnwise norms:

$$\begin{aligned}(\mathbf{I}_n - \mathbf{P})\mathbf{D}^{q-p}(|\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}))\mathbf{P} &= \|\mathbf{u}\|_p^{q-p}\ \mathsf{blockdiag}(\mathbf{v}\mathbf{v}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{w}\mathbf{w}^\top \mathbf{u}\mathbf{u}^\top, \mathbf{W}) \\ &= \|\mathbf{u}\|_p^{q-p} \cdot \mathbf{v}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{w} \cdot \ \langle \mathbf{w}, \mathbf{u}\rangle \\ &\quad \cdot \mathsf{blockdiag}(\mathbf{v}\mathbf{u}^\top, \mathbf{W}').\end{aligned}$$

with the appropriate adaptation for $p = 1$. As $\langle \mathbf{w}, \mathbf{u}\rangle = c^p + s^p > 0$ and $\mathbf{v}^\top \mathsf{diag}(|\mathbf{u}|^{q-p})\mathbf{w} = -sc^{q-1} + cs^{q-1}$, the same reasoning shows this is zero if, and only if, $q = 2$.

- For $m = 2$, $n = 3$, $q = 2$, consider the unit norm vector $\mathbf{v} \stackrel{\text{def}}{=} [2, 1, 1]^\top/\sqrt{6}$ and define $\mathbf{P} \stackrel{\text{def}}{=} \mathbf{I}_3 - \mathbf{v}\mathbf{v}^\top$, which is a rank-$(m = 2)$ projection matrix. We have for any $0 < p < \infty$

$$\begin{aligned}\mathbf{P} = \tfrac{1}{6}\begin{bmatrix} 2 & -2 & -2 \\ -2 & 5 & -1 \\ -2 & -1 & 5 \end{bmatrix} &= \tfrac{1}{3}\begin{bmatrix} 1 & -1 & -1 \\ -1 & 5/2 & -1/2 \\ -1 & -1/2 & 5/2 \end{bmatrix} \\ |\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}) &\propto \begin{bmatrix} 1 & -1 & -1 \\ -1 & (5/2)^{p-1} & -(1/2)^{p-1} \\ -1 & -(1/2)^{p-1} & (5/2)^{p-1} \end{bmatrix} \stackrel{\text{def}}{=} \mathbf{Q}_p\end{aligned}$$

Moreover denoting $c_p \stackrel{\text{def}}{=} 1 + (5/2)^p + (1/2)^p$ we have $\mathbf{D}_{\text{row}}^{q-p} \propto (\mathbf{I}_3 + \lambda \mathbf{e}_1\mathbf{e}_1^\top)$ with $\lambda = (3/c_p)^{q/p-1} - 1$, hence

$$(\mathbf{I}_3 - \mathbf{P})\mathbf{D}^{q-p}(|\mathbf{P}|^{p-1} \odot \mathsf{sign}(\mathbf{P}))\mathbf{P} \propto \mathbf{v}\mathbf{v}^\top \left(\mathbf{I}_3 + \lambda \mathbf{e}_1\mathbf{e}_1^\top\right) \mathbf{Q}_p\mathbf{Q}_2 \tag{97}$$

As $\mathbf{v} \neq \mathbf{0}_3$, the right-hand side of (97) vanishes if and only if

$$\mathbf{0}_3 = \mathbf{Q}_2\mathbf{Q}_p(\mathbf{I}_3 + \lambda \mathbf{e}_1\mathbf{e}_1^\top)\mathbf{v} = \mathbf{Q}_2\mathbf{Q}_p\mathbf{v} + \lambda v_1 \mathbf{Q}_2\mathbf{Q}_p\mathbf{e}_1 \propto \mathbf{Q}_2\mathbf{Q}_p \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} + 6\lambda \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \tag{98}$$

We let the reader check that

$$\mathbf{Q}_2\mathbf{Q}_p \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = 2(2 - (5/2)^{p-1} + (1/2)^{p-1}) \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

hence (98) is equivalent to $\lambda = -((1/2)^{p-1} - (5/2)^{p-1} + 2)/3$. We derived before that $\lambda = (3/c_p)^{q/p-1} - 1$ hence the right-hand side of (97) vanishes if and only if:

$$\left[\frac{3}{1 + \left(\frac{5}{2}\right)^p + \left(\frac{1}{2}\right)^p}\right]^{\frac{2-p}{p}} = \frac{1}{3}\left[-\left(\frac{1}{2}\right)^{p-1} + \left(\frac{5}{2}\right)^{p-1} + 1\right]. \tag{99}$$

Setting $g(p) \stackrel{\text{def}}{=} \frac{1}{3}\left[1 + \left(\frac{5}{2}\right)^p + \left(\frac{1}{2}\right)^p\right]$ for $p \in \mathbb{R}$ this can be written for $0 < p < \infty$ as

$$g(p)^{\frac{p-2}{p}} = g(p-1) - \frac{1}{3}\left(\frac{1}{2}\right)^{p-2}.$$

- For $p = 2$ the equation is satisfied and the right-hand side of (97) indeed vanishes.
- For $p > 2$, note that $g(p) < \frac{1}{2}\left(\frac{5}{2}\right)^p$ since it is equivalent to

  $$\tfrac{1}{6}\left(\tfrac{5}{2}\right)^p > \tfrac{1}{3} + \tfrac{1}{3}\left(\tfrac{1}{2}\right)^p$$

  where the left-hand side is strictly increasing, and the right-hand side strictly decreasing and smaller than the left-hand side for $p = 2$. Therefore, for $p > 2$,

  $$g(p)^{\frac{p-2}{p}} < \left[\tfrac{1}{2}\left(\tfrac{5}{2}\right)^p\right]^{\frac{p-2}{p}} = \tfrac{2^{2/p}}{5}\left(\tfrac{5}{2}\right)^{p-1}.$$

  On the other hand,

  $$\left[\tfrac{1}{2}\left(\tfrac{5}{2}\right)^p\right]^{\frac{p-2}{p}} < g(p-1) - \tfrac{1}{3}\left(\tfrac{1}{2}\right)^{p-2}.$$

  To see that this is true, we rewrite it as

  $$\tfrac{3\cdot 2^{2/p}-5}{15}\left(\tfrac{5}{2}\right)^{p-1} < \tfrac{1}{3} - \tfrac{1}{3}\left(\tfrac{1}{2}\right)^{p-1}$$

  which holds since for $p = 2$ the two sides coincide, the right-hand side is strictly increasing, and the left-hand side strictly decreasing (the latter can be verified directly by studying the sign of the derivative).[12]
  In summary, for $p > 2$ we have

  $$g(p)^{\frac{p-2}{p}} < \tfrac{2^{2/p}}{5}\left(\tfrac{5}{2}\right)^{p-1} < g(p-1) - \left(\tfrac{1}{2}\right)^{p-2}$$

  and the equation (99) cannot be consistent.

[12] The derivative of the left-hand side is $\left\{\left(\frac{2}{5}\right)^{1-p}\left[4^{1/p}\left(3p^2\log\left(\frac{5}{2}\right) - 2\log(8)\right) - 5p^2\log\left(\frac{5}{2}\right)\right]\right\}/(15p^2)$. When $p > \log(4)/\log(5/3) \approx 2.71$, then $4^{1/p}\cdot 3 < 5$ and the expression is negative. For $2 < p < 3$, $4^{1/p}\left(3p^2\log\left(\frac{5}{2}\right) - 2\log(8)\right) - 5p^2\log\left(\frac{5}{2}\right) < 2\left(3p^2\log\left(\frac{5}{2}\right) - 2\log(8)\right) - 5p^2\log\left(\frac{5}{2}\right) = p^2\log\left(\frac{5}{2}\right) - 4\log(8) < 9\log\left(\frac{5}{2}\right) - 4\log(8) \approx -0.07$.

– For $0 < p < 2$, the following holds:

$$g(p)^{(p-2)/p} > 2^{p-2}.$$

To see this, note that it is equivalent to (remember that $p-2<0$ so raising both sides to the power $1/(p-2)$ reverses the inequality)

$$1 + \left(\tfrac{5}{2}\right)^p + \left(\tfrac{1}{2}\right)^p < 3 \cdot 2^p,$$

or

$$\left(\tfrac{1}{2}\right)^p + \left(\tfrac{5}{4}\right)^p + \left(\tfrac{1}{4}\right)^p < 3,$$

which holds since $p \mapsto \left(\frac{1}{2}\right)^p + \left(\frac{5}{4}\right)^p + \left(\frac{1}{4}\right)^p$ is convex on $\mathbb{R}$, equal to 3 for $p=0$, and equal to $\frac{15}{8} < 3$ for $p=2$. On the other hand, we have that

$$2^{p-2} > g(p-1) - \tfrac{1}{3}\left(\tfrac{1}{2}\right)^{p-2}$$

for $0 < p < 2$ and that the equation holds for $p = 0$. This follows since the two sides coincide for $p = 2$ and since the function

$$a(p) \stackrel{\text{def}}{=} 2^{p-2} - \left[g(p-1) - \tfrac{1}{3}\left(\tfrac{1}{2}\right)^{p-2}\right]$$

is decreasing on $[0,\ 2]$. Indeed

$$a'(p) = \tfrac{1}{15} 2^{-p-2} \underbrace{\left[4^p \cdot 15\log(2) - 5^p \cdot 8\log\left(\tfrac{5}{2}\right) - 40\log(2)\right]}_{\stackrel{\text{def}}{=} b(p)},$$

and

$$b'(p) = 15 \cdot 4^p \log(2)\log(4) - 8 \cdot 5^p \log\left(\tfrac{5}{2}\right)\log(5).$$

It can be verified that $b(p)$ has a single critical point on $[0,2]$ at $p_1 \stackrel{\text{def}}{=} \log\left(\frac{15\log(2)\log(4)}{8\log\left(\frac{5}{2}\right)\log(5)}\right) / \log\left(\frac{5}{4}\right)$ which is a maximum, and that $b(p_1) \approx -22.72$. Hence, $a'(p) < 0$ on $(0,2)$.

In conclusion, for (98) to vanish when $q = 2$, it must hold that $p = 2$.

- For any $3 \le m < n$, one can find $\mathbf{P}_1 \in \mathbb{R}^{(n-3)\times(n-3)}$, a projection matrix of rank $m-2 \ge 1$ with no zero column, and build $\mathbf{P}' = \mathsf{blockdiag}(\mathbf{P}, \mathbf{P}_1)$ with the rank-2 matrix $\mathbf{P} \in \mathbb{R}^{3\times 3}$ we have just built. The same reasoning as before leads to a construction so that for $1 \le p < \infty$, $q = 2$, $\mathbf{A}_5^\dagger \in \mathsf{pginv}_{\underline{p,q}}(\mathbf{A}_5) \iff p = 2$.

□

References

[BD53] R. Bott and R. J. Duffin, "On the algebra of networks," 1953.

[BDP+14] A. Bourrier, M. E. Davies, T. Peleg, P. Perez, and R. Gribonval, "Fundamental Performance Limits for Ideal Decoders in High-Dimensional Linear Inverse Problems," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7928–7946, 2014.

[BIG03] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, ser. CMS Books in Mathematics. New York: Springer, Jun. 2003.

[BJMO11] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with Sparsity-Inducing Penalties," *arXiv*, Aug. 2011.

[BLPY09] J. Blum, M. Lammers, A. M. Powell, and Ö. Yılmaz, "Sobolev Duals in Frame Theory and Sigma-Delta Quantization," *J. Fourier Anal. Appl.*, vol. 16, no. 3, pp. 365–381, 2009.

[BPC11] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers.* Now Publishers Inc, 2011.

[CHKK] P. G. Casazza, A. Heinecke, F. Krahmer, and G. Kutyniok, "Optimally Sparse Frames," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7279–7287.

[DKV13] I. Dokmanić, M. Kolundžija, and M. Vetterli, "Beyond Moore-Penrose: Sparse pseudoinverse," in *IEEE ICASSP.* IEEE, 2013, pp. 6526–6530.

[DLL94] I. Daubechies, H. J. Landau, and Z. Landau, "Gabor Time-Frequency Lattices and the Wexler-Raz Identity," *J. Fourier Anal. Appl.*, vol. 1, no. 4, pp. 437–478, 1994.

[DMA97] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive Greedy Approximations," *Constr. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.

[DSSSC08] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the l 1 -ball for learning in high dimensions.* New York, New York, USA: ACM, Jul. 2008.

[EB92] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators," *Math. Program.*, vol. 55, no. 1-3, pp. 293–318, 1992.

[GB08] M. Grant and S. Boyd, "Graph Implementations for Nonsmooth Convex Programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[GB14] ——, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[GLP+12] C. S. Güntürk, M. Lammers, A. M. Powell, R. Saab, and Ö. Yılmaz, "Sobolev Duals for Random Frames and $\Sigma\Delta$ Quantization of Compressed Sensing Measurements," *Foundations of Computational Mathematics*, vol. 13, no. 1, pp. 1–36, 2012.

[GN07] R. Gribonval and M. Nielsen, "Highly Sparse Representations From Dictionaries Are Unique and Independent of the Sparseness Measure," *Appl. Comput. Harmon. Anal.*, vol. 22, no. 3, pp. 335–355, May 2007.

[HJ12] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, Oct. 2012.

[JAB11] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured Variable Selection with Sparsity-Inducing Norms," *The Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, Feb. 2011.

[Kay98] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*, 1998.

[KC08] J. Kovačević and A. Chebira, *An Introduction to Frames.* Now Publishers Inc, 2008.

[KKL12] F. Krahmer, G. Kutyniok, and J. Lemvig, "Sparsity and Spectral Properties of Dual Frames," *Linear Algebra Appl.*, vol. 439, no. 4, pp. 1–17, Dec. 2012.

[Lew10] A. D. Lewis, "A Top Nine List: Most Popular Induced Matrix Norms," Queen's University, Kingston, Ontario, Tech. Rep., 2010.

[LH10] J. Lopez and D. Han, "Optimal Dual Frames for Erasures ," *Linear Algebra Appl.*, vol. 432, no. 1, pp. 471–482, Jan. 2010.

[LHH] J. Leng, D. Han, and T. Huang, "Optimal Dual Frames for Communication Coding With Probabilistic Erasures," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5380–5389.

[LLM13] S. Li, Y. Liu, and T. Mi, "Sparse Dual Frames and Dual Gabor Functions of Minimal Time and Frequency Supports," *J. Fourier Anal. Appl.*, vol. 19, no. 1, pp. 48–76, 2013.

[LPY08] M. Lammers, A. M. Powell, and Ö. Yılmaz, "Alternative Dual Frames for Digital-to-Analog Conversion in Sigma–Delta Quantization," *Adv Comput Math*, vol. 32, no. 1, pp. 73–102, Jul. 2008.

[LY09] J. Liu and J. Ye, *Efficient Euclidean projections in linear time.* New York, New York, USA: ACM, Jun. 2009.

[LY10] ——, "Efficient ℓ^1/ℓ^q Norm Regularization," Sep. 2010.

[Mir60] L. Mirsky, "Symmetric Gauge Functions and Unitarily Invariant Norms," *Q. J. Math. Q. J. Math. Q. J. Math. Q. J. Math.*, vol. 11, no. 1, pp. 50–59, 1960.

[MN70] N. Minamide and K. Nakamura, "A Restricted Pseudoinverse and Its Application to Constrained Minima," *SIAM J. Appl. Math.*, 1970.

[Moo20] E. H. Moore, "On the Reciprocal of the General Algebraic Matrix," *Bulletin of the American Mathematical Society*, vol. 26, pp. 394–395, 1920.

[Nat95] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 25, no. 2, pp. 227–234, 1995.

[NO69] T. G. Newman and P. L. Odell, "On the Concept of a $p-q$ Generalized Inverse of a Matrix," *SIAM J. Appl. Math.*, vol. 17, no. 3, pp. 520–525, 1969.

[PB14] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2014.

[Pen08] R. Penrose, "A Generalized Inverse for Matrices," *Math. Proc. Camb. Phil. Soc.*, vol. 51, no. 03, p. 406, Oct. 2008.

[PHSB14] N. Perraudin, N. Holighaus, P. L. Søndergaard, and P. Balazs, "Designing Gabor Windows Using Convex Optimization," *arXiv*, Jan. 2014.

[QCCD09] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, "An Efficient Projection for $\ell_{1,\infty}$ Regularization," in *ICML '09.* New York, New York, USA: ACM, Jun. 2009, pp. 857–864.

[SG10] A. Suárez and L. González, "Applied Mathematics and Computation," *Appl. Math. Comput.*, vol. 216, no. 2, pp. 514–522, Mar. 2010.

[Sra12] S. Sra, "Fast Projections Onto Mixed-Norm Balls with Applications," *Data Min. Knowl. Disc.*, vol. 25, no. 2, pp. 358–377, Sep. 2012.

[VKG14] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing.* Cambridge, UK: Cambridge University Press, 2014.

[vN37] J. von Neumann, "Some Matrix Inequalities and Metrization of Matrix-Space ," *Tomsk University Review*, no. 1, pp. 286–300, 1937.

[WLY13] J. Wang, J. Liu, and J. Ye, "Efficient Mixed-Norm Regularization: Algorithms and Safe Screening Methods," *arXiv*, Jul. 2013.

[Zie97] K. Zietak, "Strict Spectral Approximation of a Matrix and Some Related Problems," *Appl. Math.*, vol. 24, no. 3, pp. 267–280, 1997.